

**ECONOMIC RESEARCH REPORTS**

VERIFICATION AND STABILITY:  
A GAME THEORETIC ANALYSIS

by

Steven J. Brams

and

D. Marc Kilgour

R.R. #86-13

April 1986

**C. V. STARR CENTER  
FOR APPLIED ECONOMICS**



**NEW YORK UNIVERSITY  
FACULTY OF ARTS AND SCIENCE  
DEPARTMENT OF ECONOMICS  
WASHINGTON SQUARE  
NEW YORK, N.Y. 10003**

VERIFICATION AND STABILITY: A GAME THEORETIC ANALYSIS

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

D. Marc Kilgour\*\*  
Department of Mathematics  
Wilfrid Laurier University  
Waterloo, Ontario  
Canada N2L 3C5

\* Steven J. Brams gratefully acknowledges the financial support of the Ford Foundation under Grant No. 845-0354 and the National Science Foundation under Grant No. SES84-08505.

\*\* D. Marc Kilgour gratefully acknowledges the financial support of the Natural Sciences and Engineering Research Council of Canada under Grant No. A8974.

ABSTRACT

## VERIFICATION AND STABILITY: A GAME-THEORETIC ANALYSIS

A two-person, variable-sum "verification game" is analyzed between an inspector and an inspectee. The inspectee most wants its claimed compliance with an arms-control treaty accepted, whereas the inspector most desires actual compliance. The inspector, however, can only imperfectly detect the actions of the inspectee.

An optimal inducement strategy of the inspector, whereby it announces a mixed strategy to induce an optimal pure-strategy best response from the inspectee, is found. Given that the inspector's detection probabilities (of both compliance and noncompliance) are above a certain threshold, inducement by the inspector improves its payoff and does not significantly reduce the inspectee's, relative to the payoffs at the dominant Nash equilibrium. Below this threshold, inducement cannot occur; the only Nash equilibrium is noncooperative and Pareto-inferior to the other equilibria above the threshold.

The appropriateness of the inducement and equilibrium solution concepts is considered in the context of superpower conflict. It is suggested that weapons such as cruise missiles that are difficult to detect and verify may impede finding solutions that are both cooperative and stable in the verification game.

## VERIFICATION AND STABILITY: A GAME-THEORETIC ANALYSIS

Steven J. Brams and D. Marc Kilgour

### 1. Introduction

Arms control between the superpowers seems at a dead end unless the provisions of any treaty they negotiate can be verified. In a recent assessment, one expert argued that "verification has become the most important standard against which arms control agreements -- both past and prospective -- are measured" (Potter, 1985, p. 1), and then went on to point out that

five years ago ... verification was a contentious issue in the arms control debate, but it was not the decisive one. Today, this situation has changed, and verification stands as the litmus test by which arms control proposals are assessed (Potter, 1985, p. 4).

Another expert expressed a similar viewpoint about the primacy of verification:

The ability of the United States to verify Soviet compliance with the provisions of arms control agreements ... has been taken by many as a fundamental requirement for establishing that a treaty is in the interests of the United States (Richelson, 1985, p. 202).

The formal analysis of verification, based on strategic models of rational choice, has a curious history. Twenty years ago, such analysts as Maschler (1966, 1967) and Rapoport (1966) developed rather elaborate game-theoretic models to analyze the strategic implications of different inspection procedures. Then, except for some scattered and mostly unpublished reports, game theory and other formal tools of rational strategic analysis were not again applied to arms-control issues until

recently. We shall review some of this work in the next section but first we shall outline the scope of this paper and its main results.

First, as a basic model of verification we shall propose a variable-sum game, between an inspector and an inspectee, different from games that have previously been analyzed in the literature. For reasons to be given, we believe that this game better captures the crucial features of the strategic conflict between the superpowers over the issue of verification.

Endowing the inspector with an ability to detect the choices of the inspectee, but only imperfectly, we shall first analyze the payoffs of the players from the inspector's inducing the inspectee to respond to its announced strategy. Inducement may be thought of as a device for one player's leading the other to do something, or deterring it from doing something, in such a way as ultimately to benefit itself.

Next we shall systematically examine all Nash (1951) equilibria in this game. The dependence of equilibria and inducement outcomes on the inspector's detection probabilities -- both to ascertain compliance and uncover noncompliance -- will be shown.

The two solution concepts of inducement and equilibrium will then be compared, and the applicability of each to practical verification issues discussed. Our general goal is to provide a theoretical framework for viewing such issues, indicating what kinds of strategies are appropriate for achieving outcomes that help to dispel distrust and stabilize cooperation.

Trust and stability are closely linked: adversaries can afford to trust each other if an outcome is stable -- that is, a unilateral deviation will hurt the deviator -- whether the players' choices are simultaneous (Nash equilibrium) or sequential (inducement outcomes). Thus, if the

inspectee is robbed of the incentive to cheat, and the inspector to hide or manipulate the information it collects, trust will be fostered.

Moreover, this is trust based not on mere blind faith but rather on calculations of advantage and disadvantage. In the context of arms control, we shall derive conditions under which it is more costly to violate than not violate the provisions of a treaty when the players choose strategies that are in some sense stable.

Thereby we are able to connect verification to stability and ultimately to trust. In fact, the connection is two-sided: verification can give one good reason both to trust an opponent and to be trustworthy to him.

This is not to say that present technologies for verifying adherence to arms-control agreements are adequate for all weapons systems. Immense practical problems exist in detecting the deployment of certain strategic weapons, such as cruise missiles. An important by-product of our game-theoretic analysis -- in addition to clarifying the qualitative notions of stability and inducement -- is the demonstration that there are quantitative thresholds, as measured by the level of detector reliability, sufficient to ensure different kinds of stable and cooperative outcomes.

## 2. The Verification Game

A number of different games have been proposed as models of the strategic situation facing a possible violator of an arms-control agreement, whom we shall call Inspectee, and a detector who signals possible violations, whom we shall call Inspector. Before reviewing these, we shall posit our own game, which we call the (Basic) Verification Game (VG), and then draw some comparisons between this game and others

that have been proposed as models of verification.<sup>1</sup>

In VG we assume Inspectee may either comply with (C) or violate (i.e., not comply with  $\bar{C}$ ) an arms-control agreement, but that Inspectee always claims that he chose C. Inspector may either accept (A) or challenge (i.e., not accept  $\bar{A}$ ) Inspectee's stated compliance. (Later we shall use mixed strategies to model the possibility that Inspectee may choose any level of compliance or noncompliance, and Inspector any degree of acceptance or nonacceptance, but this complication is not necessary to describe the basic game.)

In the basic version of VG, each player has two strategies that lead to four possible outcomes at the intersection of each pair of strategy choices, as shown in Figure 1. We assume that the payoffs to the players

---

Figure 1 about here

---

are normalized, so that the best outcomes for Inspectee (ROW) and Inspector (COL) have utilities  $r_4 = 1$  and  $c_4 = 1$ , and the worst have utilities  $r_1 = 0$  and  $c_1 = 0$ , respectively. The players' next-best payoffs are  $r_3$  and  $c_3$ , and next-worst  $r_2$  and  $c_2$ . In summary, the payoffs satisfy

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

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

VG is defined by the payoffs at the four outcomes shown in Figure 1. These payoffs are assumed to be cardinal utilities and are indicated by ordered pairs  $(r_i, c_j)$ . For example,  $(r_3, 1)$  at the upper-left outcome of Figure 1, associated with the strategies C and A, gives ROW its next-best payoff and COL its best payoff.

The payoff rankings in VG can be characterized by the primary (1) and

FIGURE 1  
THE BASIC VERIFICATION GAME (VG)

		<u>Inspector (COL)</u>		
		Accept (A)	Challenge/ Don't accept ( $\bar{A}$ )	
<u>Inspectee (ROW)</u>	Comply (C)	$(r_3, 1)$	$(0, c_3)$	(s)
	Violate/ Don't comply ( $\bar{C}$ )	$(1, 0)$	$(r_2, c_2)$	(1 - s)
		(u)	(1 - u)	

Key:  $(r_i, c_j) = (\text{payoff to ROW, payoff to COL})$

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

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

secondary (2) goals of each player:

ROW (Inspectee)

1. Prefers COL accept its claimed compliance (two best outcomes associated with A, two worst with  $\bar{A}$ ).
2. Prefers to violate agreement (whether COL chooses A or  $\bar{A}$ , ROW prefers  $\bar{C}$ ).

COL (Inspector)

1. Prefers ROW comply (two best outcomes associated with C, two worst with  $\bar{C}$ ).
2. Prefers policy of tit-for-tat (if ROW chooses C, COL prefers A; if ROW chooses  $\bar{C}$ , COL prefers  $\bar{A}$ ).

These goals determine a lexicographic order (Fishburn, 1974), whereby the primary goal distinguishes the two best from the two worst outcomes for each player; given this distinction, the secondary goal orders the two best, on the one hand, and the two worst, on the other, for each player.

Thus for ROW, its primary goal says that its two best outcomes are in the A column of Figure 1, and its secondary goal says that between these two outcomes it prefers the one associated with  $\bar{C}$ . Therefore, its payoff from strategy pair  $\bar{C}A$  is  $r_4 = 1$ , and its payoff from strategy pair  $CA$  is  $r_3$ ; similarly, between its two worst payoffs in the  $\bar{A}$  column (as implied by its primary goal) its secondary goal says that  $\bar{C}\bar{A}$  yields  $r_2$  and  $C\bar{A}$  yields  $r_1 = 0$ .

The primary goals of each player would certainly appear plausible: ROW would want its claimed compliance accepted, and COL would want actual compliance. Also, ROW's secondary goal, if morally dubious, is probably realistic in many situations. For if violating an agreement were not profitable, then there would be

no reason for ROW not to comply, making for a trivial game with a mutually best (1, 1) outcome.

Similarly, COL's secondary goal seems eminently defensible. First, given ROW chooses C, COL's preference for A is reasonable, for why should COL not cooperate by choosing A when ROW cooperates by choosing C? The second part of this tit-for-tat goal seems equally reasonable -- that is, for COL to be most hurt when it accepts a violation.

If the reasonableness of the goals of the players is not at issue, the question of which goal is primary and which is secondary may be more controversial. Conceivably, ROW may prefer violating an agreement over having it accepted by COL, and COL may prefer a policy of tit-for-tat over ROW's compliance. A reversal in the priority of primary and secondary goals by one player would give rise to two new games, and a reversal by both would generate a third.

Different variable-sum games have been analyzed by other theorists. For example, Rapoport (1966, pp. 169-185) applied a number of different solution concepts to a game between an inspector and an evader, in which the primary and secondary goals of the inspector duplicated ours in VG. However, the evader had a primary goal of deceiving the inspector -- not simply having its claimed compliance accepted; its secondary goal was the same as our inspector's primary goal. Maschler (1966, 1967) proposed more elaborate inspection games, involving chance and an inspector who could decide whether or not to investigate a suspicious event. In his model, Maschler assumed that the inspector could announce and commit itself to a mixed strategy, against which the potential violator would maximize its expected payoff.

Brams and Davis (1985) and Brams (1985, ch. 4) used this idea of

inducement, which was originally proposed by Stackelberg (1934) to study price leadership, and applied it to a "truth game" to model superpower verification. In this game, which reverses the priority of the primary and secondary goals of the inspector but not the evader in Rapoport's (1966) game, they investigated both "inducement" and "guarantee" strategies. More specifically, they analyzed the benefits that not only the inspector but also the inspectee could realize by inducing the other player to respond to it, comparing these with the payoffs that the players could guarantee for themselves whatever the opponent's choice (analogous to minimax/maximin strategies in two-person constant-sum games). More recently, Brams and Kilgour (1986) studied the Nash-equilibrium strategies of the players in Rapoport's inspector-evader game and the truth game.

Fichtner (1986) compared various game-theoretic approaches the inspection and verification of arms-control agreements; as well, he applied similar solution concepts to auditing, consumer protection, and nuclear safeguards. Avenhaus's (1986) treatment of the latter issue, in particular, makes significant use of game theory; among other things, he finds optimal strategies for an inspector, with goals the same as ours in VG, to induce compliance by an inspectee, whose primary but not secondary goal is the same as ours in VG.

There is probably no "best" game to model all aspects of verification. We have already justified VG in terms of the primary and secondary goals of the players, but it is worth noting why the VG ranking seems especially plausible in the case of the superpowers. Consider each of the four outcomes in turn, and assume that one superpower (ROW) contemplates cheating on an arms-control treaty and the other (COL), which monitors its compliance, can challenge possible violation:

$\bar{CA} -- (1,0)$ : A successful violation of a treaty, giving one side a substantial edge in the arms race, would certainly seem the best outcome for ROW and the worst for COL.

$CA -- (r_3,1)$ : This is the best outcome for COL, for it validates the treaty, but it is definitely inferior for ROW because an unchallenged violation could give ROW the edge mentioned above.

$\bar{CA} -- (0,c_3)$ : COL gets compliance but, without at first recognizing it, creates some distress for itself by erroneously challenging ROW; for ROW, on the other hand, a false charge of cheating is its worst outcome, undermining the benefits of the treaty for no gain.

$\bar{CA} -- (r_2,c_2)$ : A rather unsatisfactory outcome for both sides, because COL's justified challenge of violations underscores the treaty's fragile status, perhaps leading to its abrogation.

One might contend that  $\bar{CA}$  would be higher than next-worst in COL's preferences, for challenging noncompliance is certainly in COL's interest. We believe, however, that COL would prefer compliance at  $\bar{CA}$  to noncompliance at  $\bar{CA}$ , even though compliance at the former outcome includes the embarrassment of a false charge.

The reason for this preference is that the damage done by an unjustified accusation can more easily be undone than that of an actual violation, even detected and challenged. After forty years of conflict, both superpowers almost certainly would prefer to have their adversary adhere to SALT and other arms-control treaties, even at the cost of occasionally making unsubstantiated charges, than to catch their adversary in a lie and challenge

real violations.

Whether  $c_2$  and  $c_3$  are interchanged in VG, however, makes no difference for the rational strategy choice of ROW, which is to choose  $\bar{C}$ . For this strategy is dominant: whether COL chooses A or  $\bar{A}$ ,  $\bar{C}$  is better than C for ROW. In a game of complete information, COL would know that ROW has an unconditionally best strategy choice, and presuming ROW would choose it, COL could do no better than choose  $\bar{A}$ , leading to  $(r_2, c_2)$ .

This is the unique Nash equilibrium in VG: once at this outcome, neither player would have an incentive to depart from it unilaterally because it would do worse if it did. Unfortunately for both players, however, this outcome is Pareto-inferior: it is worse for both players than  $(r_3, 1)$ . Yet the latter outcome is not in equilibrium because ROW has an incentive to depart from it to  $(1, 0)$ , its best outcome.

The Pareto-inferior rational solution to this game can be circumvented if, assuming ROW chooses its strategy first, COL has perfect information about ROW's choice and ROW knows this. In this case, it is easy to show that COL would have a dominant strategy of tit-for-tat -- choose A if ROW chooses C, and  $\bar{A}$  if ROW chooses  $\bar{C}$  -- and Row, anticipating this dominant-strategy choice, would choose C, resulting in  $(r_3, 1)$ , a Nash equilibrium in the resulting  $2 \times 4$  game.

But, of course, the superpowers in general will have only imperfect information about each other's choices when each plays the role of Inspector in VG. How, then, can they use their imperfect detection equipment to choose optimally? In the next section, we shall propose two different notions of "optimal" and analyze how rational strategies in each case depend on the quality of the detection equipment in a more realistic version of VG.

### 3. Inducement in the Verification Game with Detection

We assume that COL is equipped with a detector and has the option of consulting it before choosing A or  $\bar{A}$ . COL's detector is characterized by parameters  $x$  and  $y$ , which are conditional probabilities that describe its reliability and are assumed to be known by both players:

$$x = \Pr\{\text{detector signals violation} \mid \text{ROW chose } \bar{C}\}$$

$$y = \Pr\{\text{detector signals no violation} \mid \text{ROW chose } C\}$$

Note that  $0 \leq x, y \leq 1$ . The detector is perfect when  $x = y = 1$  and would seem worthless when  $x$  and/or  $y$  are near or at 0.<sup>2</sup>

We assume that ROW chooses its strategy before COL. ROW has two pure, or single, strategies,  $C$  and  $\bar{C}$ , so its mixed strategy, which is simply a probability distribution over its set of pure strategies, can be represented by a single probability  $s$ :

$$\Pr\{\text{ROW chooses } C\} = s.$$

Because COL has a detector, its choices are more complicated. Besides its two pure strategies,  $A$  and  $\bar{A}$ , which we treat as unconditional choices (i.e., made without consulting the detector), it has a third pure strategy of consulting its detector ( $D$ ). If COL chooses  $D$ , we assume it will follow a policy of tit-for-tat by picking  $\bar{A}$  if the detector signals a violation and  $A$  otherwise. Altogether, we represent COL's mixed strategy by two probabilities,  $t$  and  $u$ , where

$$\Pr\{\text{COL chooses } D\} = t$$

$$\Pr\{\text{COL chooses } A\} = u$$

Of course,  $\Pr\{\text{COL chooses } \bar{A}\} = 1 - t - u$ , just as  $\Pr\{\text{ROW chooses } \bar{C}\} = 1 - s$ .

Because the mixed strategies are probabilities,

$$0 \leq s \leq 1; \quad 0 \leq u \leq 1; \quad 0 \leq t \leq 1; \quad u + t \leq 1;$$

the latter sum is the probability that COL chooses either  $A$  or  $D$  -- that is,

does not choose  $\bar{A}$ .<sup>3</sup>

We have indicated the probabilities of ROW's choices C and  $\bar{C}$  ( $s$  and  $1 - s$ ), and the unconditional probabilities of COL's choices A and  $\bar{A}$  ( $u$  and  $1 - u$ ), in parentheses to the right and below these strategies in Figure 1. Combining these probabilities with the probability,  $t$ , of COL's consulting its detector (D), and then choosing either A or  $\bar{A}$ , we obtain probabilities that each of the four possible outcomes of the game will occur:

$$CA: s(u + ty)$$

$$C\bar{A}: s(1 - u - ty)$$

$$\bar{C}A: (1 - s)(t + u - tx)$$

$$\bar{C}\bar{A}: (1 - s)(1 - t - u + tx)$$

For example, CA is chosen when ROW chooses C (with probability  $s$ ) and COL chooses either A unconditionally (with probability  $u$ ) or consults its detector (with probability  $t$ ) and, detecting no violation (with probability  $y$ ), chooses A.

These outcome probabilities may now be combined with the payoffs that the players obtain at the outcomes (see Figure 1) to give the players' expected payoffs (E's) in VG. To simplify notation, we shall henceforth distinguish the E's by subscripts R (for ROW) and C (for COL):

$$E_R(s; t, u) = s(u + ty)r_3 + (1 - s)(t + u - tx) + (1 - s)(1 - t - u + tx)r_2$$

$$E_C(t, u; s) = s(u + ty) + s(1 - u - ty)c_3 + (1 - s)(1 - t - u + tx)c_2$$

In the Appendix, we show that the maximin strategies for ROW and COL -- that is, the strategies that maximize their minimum expected payoffs, whatever the opponent does -- are  $s = 0$  and  $u = t = 0$ , respectively. These strategies, which are for ROW always to choose  $\bar{C}$  and for COL always to choose  $\bar{A}$  unconditionally (and never consult its detector), give the players

maximin values of  $r_2$  and  $c_2$ , respectively. Recall that these are the payoffs to the players from choosing their Nash-equilibrium strategies in the Figure 1 game (without the possibility of detection by COL).

Later we shall investigate the Nash equilibria that arise when COL has an imperfect detector that it can use to try to discern ROW's (prior) strategy choice. They can be different from the unique Nash equilibrium in the Figure 1 game without detection.

First, however, we shall investigate how COL, by announcing and committing itself to a mixed strategy  $(u, t)$ , can induce ROW to respond to this strategy in such a way that COL maximally benefits. As we shall show, COL's optimal inducement strategy depends on its detector's having detection probabilities,  $x$  and  $y$ , above a certain minimum. We shall assume, for this exposition, that both  $x < 1$  and  $y < 1$ ; this assumption is dropped in the Appendix.

Because ROW has only two pure strategies, COL can induce either one or the other by arranging that it be ROW's unique best response. It cannot induce a mixture since no mixed strategy could constitute ROW's best response unless all do, which is to say that no strategy, pure or mixed, is better than any other.

We show in the Appendix that  $s = 0$  (always choose  $\bar{C}$ ) is a best response for ROW to any strategy of COL if

$$x(1 - r_2) + yr_3 \leq 1.$$

But then COL cannot receive more than  $c_2$ , whereas if  $s = 1$  COL cannot receive less than  $c_3 > c_2$  (see Figure 1).

Obviously COL would prefer to induce  $s = 1$ , and he can do so provided that

$$x(1 - r_2) + yr_3 > 1.$$

We show in the Appendix that, under this condition, COL induces  $s = 1$ , and benefits maximally, when it chooses

$$u = \left[ \frac{x(1-r_2) - 1 + yr_3}{x(1-r_2) - r_3 + yr_3} \right]^- ; \quad t = 1 - u = \left[ \frac{1 - r_3}{x(1-r_2) - r_3 + yr_3} \right]^+ , \quad (2)$$

where the "-" and "+" superscripts indicate values slightly less and greater, respectively, than those given in the brackets. This strategy makes  $s = 1$  ROW's unique best response, and among all strategies that do so this one maximizes COL's expected payoff.

What are the benefits to COL of its optimal inducement strategy? It turns out that this strategy yields COL

$$E_C^* = E_C(u, t; 1) = \left[ \frac{x(1-r_2) - (1-y)(1-c_3+c_3r_3)}{x(1-r_2) - (1-y)r_3} \right]^- ,$$

and  $c_3 < E_C^* < 1$ . Thus, COL receives strictly more than its next-best payoff of  $c_3$ ; it receives somewhat less than its best payoff of 1 because of detector unreliability.

For ROW the benefits are not so great -- in terms of the comparative rankings of the players -- but neither are they terrible:

$$E_R^* = E_R(1; u, t) = \left[ r_3 \left( \frac{x(1-r_2) - 1 + y}{x(1-r_2) - r_3 + yr_3} \right) \right]^- ,$$

and  $r_2 < E_R^* < r_3$ . In other words, ROW receives strictly more than its maximin value of  $r_2$ , but strictly less than its payoff at the "cooperative" outcome  $(r_3, 1)$ , when  $s = u = 1$  and  $t = 0$ .

Indeed,  $(r_3, 1)$  give a better payoff to both players than what COL can induce, so inducement is Pareto-inferior to the pure-strategy outcome  $(r_3, 1)$ . Moreover, even these inferior inducement payoffs are unattainable unless the conditional detection probabilities,  $x$  and  $y$ , are sufficiently high that inequality (1) is satisfied. Given that (1) is satisfied, COL

can induce either  $s = 0$  or  $s = 1$ ; by choosing  $u$  and  $t$  according to (2), it will induce  $(E_R^*, E_C^*)$ , which is certainly better for both players than  $(r_2, c_2)$  (obtained by optimally inducing  $s = 0$ ), even if it is Pareto-inferior to  $(r_3, 1)$ .

It might be thought that ROW could turn the tables on COL and induce COL to respond to its own mixed strategy. However, ROW would have no reason to announce that it might ever choose  $\bar{C}$  -- violate the treaty -- because this could only steer COL toward the choice of  $\bar{A}$ , which leads to ROW's two worst outcomes.

In the next section we assume that COL is not able to seize the initiative and announce its optimal inducement strategy to evoke a best response from ROW. Instead we ask whether COL, again with only an imperfect detector, and ROW can simultaneously (or at least in ignorance of each other) choose strategies such that neither would have an incentive to depart from its choice unilaterally.

#### 4. Nash-Equilibrium Strategies

Assume that ROW does not respond to COL's prior choice -- as under optimal inducement by COL -- perhaps because COL is unable to make a credible commitment to a particular mixed strategy. Rather, suppose that both players, knowing that COL can imperfectly detect ROW's prior choice, act in light of this knowledge (i.e., of the conditional probabilities,  $x$  and  $y$ ) and the payoffs of VG shown in Figure 1.

Under the assumption that  $x > 0$  and  $y > 0$  (i.e., the detector has nonzero probabilities of being correct), we derive in the Appendix all Nash equilibria in VG with detection. These equilibria can be classified into five distinct groupings:

I. Cooperative Equilibrium, with payoffs of  $(r_3, 1)$ . Since this equilibrium requires  $y = 1$  (perfect detection of compliance by COL) it is unrealistic and we shall not consider it further.

II. Noncooperative Equilibria, with payoffs of  $(r_2, c_2)$ . These equilibria can always (i.e., for any values of  $x$  and  $y$ ) be achieved by COL's choosing  $\bar{A}$  for certain and ROW's choosing  $\bar{C}$  for certain. If  $x=1$ , COL may instead consult his detector some of the time without changing the equilibrium.

In any event, these equilibria give the players only their maximin values and are dominated by all other equilibria whenever other equilibria exist (more on this point below).

III. Constant-Detection Equilibrium, with payoffs

$$(yr_3, xc_2 + s[c_3 + y(1 - c_2) - xc_2]),$$

where  $x(1 - r_2) + yr_3 = 1$ . COL always consults its detector at this equilibrium and so follows a policy of tit-for-tat. This equilibrium is not very important because it occurs very rarely -- only when the values of  $x$  and  $y$  precisely satisfy a linear equation.

IV. Never-Accept (Unconditionally) Equilibrium, with payoffs

$$\left( \frac{yr_2r_3}{yr_3 - (1-x)(1-r_2)}, \frac{yc_2(1-c_3) + (1-x)c_2c_3}{y(1-c_3) + (1-x)c_2} \right),$$

where  $x(1 - r_2) + yr_3 > 1$ . COL either consults its detector or chooses  $\bar{A}$  but never chooses A; because this equilibrium is dominated by V below (i.e., V is better for both players than IV), it will not be considered further.

V. Never-Challenge (Unconditionally) Equilibrium, with payoffs

$$\left( \frac{x(1-r_2)r_3 - (1-y)r_3}{x(1-r_2) - (1-y)r_3}, \frac{xc_2}{(1-y)(1-c_3) + xc_2} \right),$$

where  $x(1 - r_2) + yr_3 > 1$ . COL either consults its detector or chooses A

but never chooses  $\bar{A}$ .

To summarize, we have dismissed I as unrealistic because perfect detection is unattainable, III as unimportant because it almost never occurs, and IV because it is dominated by V. We also show in the Appendix that II is dominated by V (when V occurs); only the  $\bar{CA}$  form of II will be considered here for the  $x=1$  variant requires perfect detection.

This leaves II and V. If

$$x(1 - r_2) + yr_3 < 1, \quad (3)$$

II is the only equilibrium. The unshaded region below and to the left of the line shown in Figure 2 satisfies inequality (3).

Figure 2 about here

Thus, if the detection probabilities,  $x$  and  $y$ , are so low as to satisfy (3), only strategies associated with the Noncooperative Equilibrium are stable in VG. Formally, these strategies are

$$s = 0; u = 0, t = 0,$$

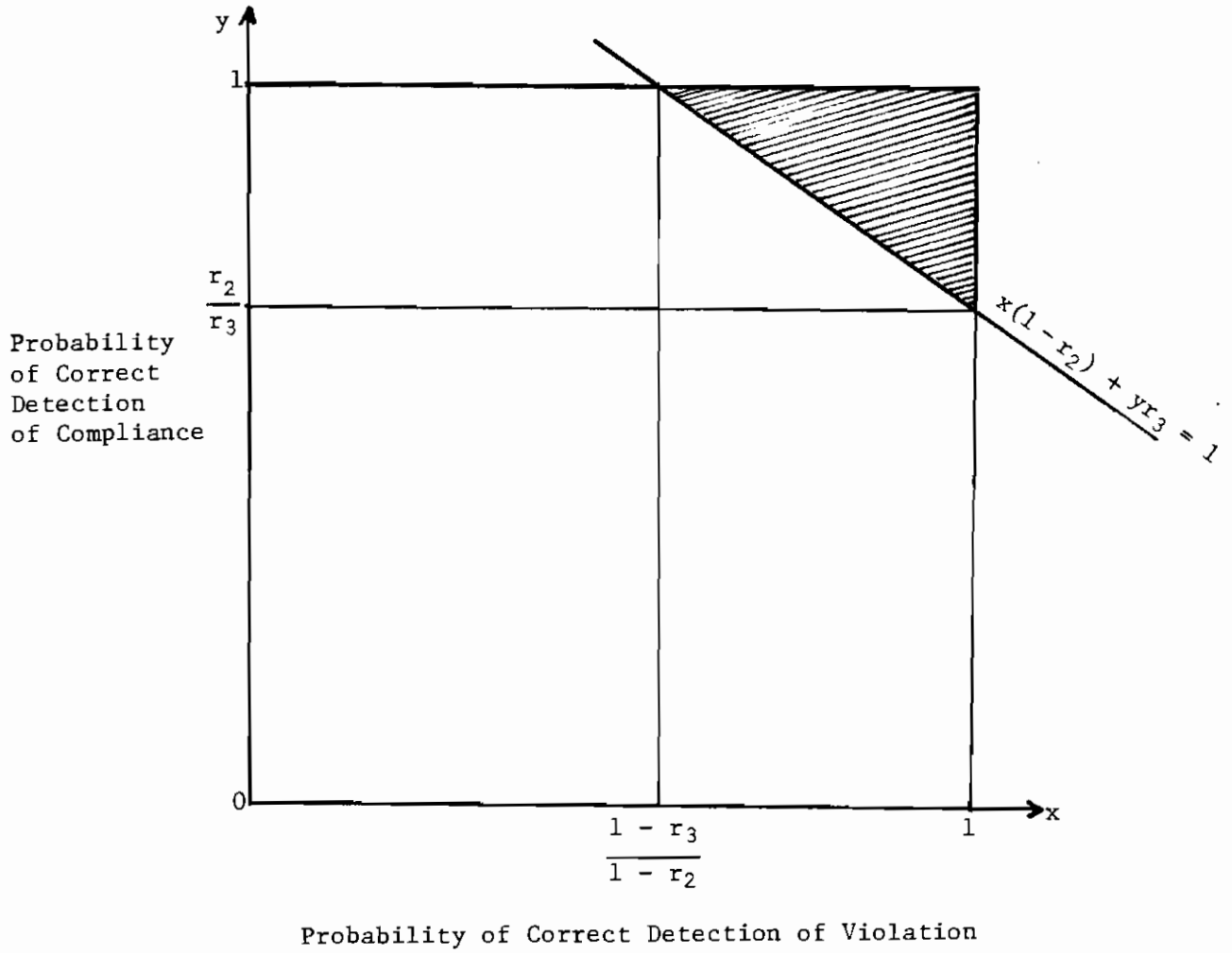
where  $x < 1$ .

The situation is a little better for the players if equality holds in (3), for Equilibrium III, which dominates II, then exists. But this occurs only on the line  $x(1 - r_2) + yr_3 = 1$ , shown in Figure 2.

The picture is still brighter if inequality (3) is reversed, which is in fact inequality (1) in section 2. If  $x$  and  $y$  are sufficiently high that  $x(1 - r_2) + yr_3 > 1$ , then ROW and COL can obtain the payoffs given by Equilibrium V. Of course, Equilibria II and IV are also available in this region, which lies above and to the right of the line  $x(1 - r_2) + yr_3 = 1$ , shown in Figure 2. Presumably, however, the players will choose the dominant equilibrium in this region -- that which is better for both players.

FIGURE 2

REGION (SHADED) IN WHICH EQUILIBRIUM V DOMINATES ALL OTHER NASH EQUILIBRIA AND INDUCEMENT IS POSSIBLE



Consider the region in which Equilibrium V exists and is dominant. This region, which is defined by inequality (1), is also the region in which COL can induce ROW to choose  $s = 1$  (i.e., always comply), as shown in section 3.

A comparison of COL's payoff at Equilibrium V with what it can obtain through optimal inducement is instructive. As shown in the Appendix, COL does better under optimal inducement, obtaining  $E_C^*$ , than it does at Equilibrium V; ROW does marginally worse.

In fact, COL's optimal inducement strategy given by (2) is almost identical to its strategy at Equilibrium V (see Appendix for details). In either case, COL sometimes consults its detector, sometimes chooses A, but never chooses  $\bar{A}$  without first consulting its detector. ROW's Equilibrium V strategy is mixed (see Appendix), but of course its best response to COL's optimal inducement is pure (C).

As the detection probabilities,  $x$  and  $y$ , approach 1 -- allowing for perfect detection -- the payoffs to the players approach  $(r_3, 1)$  both under optimal inducement and at Equilibrium V. In other words, the players can come closer and closer to the payoffs of the Cooperative Equilibrium (I) as detection improves, which is a hopeful sign in VG. Moreover, COL need never resort to choosing  $\bar{A}$  unconditionally -- before consulting its detector -- as long as inequality (1) is satisfied.

Only below the line shown in Figure 2, where inequality (3) is satisfied and detection is relatively poor, do the players have no equilibrium strategies other than always to choose  $\bar{C}$  and  $\bar{A}$ . Above this line, it is advantageous for COL sometimes to consult its detector and sometimes to accept unconditionally. This mixed strategy evokes greater compliance from ROW, which renders the outcome at Equilibrium V stable and yields higher payoffs to both players than are available at any other Nash equilibrium.

However, whenever Equilibrium V is available, it is always in COL's interest to attempt optimal inducement, for this approach would substantially improve COL's expected payoff without any significant effect on ROW's. We illustrate these ideas with a numerical example: if  $r_2 = .2$ ,  $r_3 = .7$ ,  $c_2 = .4$ ,  $c_3 = .7$ , and  $x = y = .8$ , then (1) holds, and the players have equilibria II, IV, and V available, paying  $(.2, .4)$ ,  $(.28, .48)$ , and  $(.62, .84)$ , respectively. At equilibrium V, ROW complies 84% of the time; COL consults its detector 60% of the time, and accepts unconditionally otherwise. For optimal inducement, COL credibly commits itself to consulting its detector slightly more than 60% of the time and accepting unconditionally otherwise. ROW's best response is to increase compliance to 100%, and the payoffs become  $(.62^-, .96^-)$ ; hence, COL's gain of almost .12 is achieved at essentially no cost to ROW.

## 5. Conclusions

There are both auspicious and inauspicious implications of our analysis of the Verification Game with detection. Inauspiciously, the detection probabilities must be above threshold values before the detector can be helpful at all, or inducement by Inspector is worth trying. Indeed, below these values the detector is worse than useless and should never be consulted. Instead, the players' strategies of always violating and always challenging are the only strategies in equilibrium.

Auspiciously, though the players' noncooperative strategies remain in equilibrium above the threshold values, this equilibrium is dominated by more cooperative equilibria that yield higher payoffs for both players. These equilibria involve mixed strategies, whereby Inspector sometimes relies on its detector, sometimes accepts (unconditionally) Inspectee's claimed compliance, but never

challenges unconditionally. Inspectee's equilibrium strategy is also mixed, but between compliance and noncompliance, with the probability of compliance rising as Inspector's detector improves.

By announcing a mixed strategy to which Inspectee responds, Inspector can induce a pure-strategy response by Inspectee that raises Inspector's payoff above its greatest Nash-equilibrium value (in the shaded region in Figure 2). Inspectee is slightly hurt in this region when it is induced to comply.

Whatever the differences between the inducement and equilibrium payoffs, the key to "solving" the Verification Game lies not so much in the solution concept selected but in being the favorable region (shaded in Figure 2), which triggers both the possibility of inducement and Nash equilibria that are Pareto-superior to the Noncooperative Equilibrium. Inducement probably makes most sense if there is a third party or neutral inspector, which can credibly announce acceptance conditions to which the parties to a conflict will respond. On the other hand, in a two-person game between adversaries such as the superpowers -- wherein each adversary plays the role of both Inspector and Inspectee -- the Nash-equilibrium strategies may be more sensible because of their symmetry with respect to the players.

It is difficult to say whether the verification equipment of the superpowers is sufficiently good to place them in the favorable region. Nevertheless, the nuclear arms-control agreements reached so far, though limited in scope, argue against unbridled pessimism.

Verification of the most significant agreements that the United States and the Soviet Union have achieved, beginning with the Limited Nuclear Test Ban Treaty prohibiting nuclear weapons tests in the atmosphere, outer space, and under water in 1963, has been based on

"national technical means" (NTM). The use of NTM means that each super-power relies primarily on its own surveillance of the other side's activities, using reconnaissance satellites and the like. The importance of NTM since the early 1960s should not be underestimated; in fact, without NTM, there would be no Verification Game.<sup>4</sup>

Perhaps we cannot be so sanguine about new and future weapons systems, including those, like cruise missiles, that are highly mobile or those, like anti-satellite weapons, that are hard to detect for technical reasons. Yet, because it is in both sides' interest -- whether in the role of Inspector or Inspectee -- to be in the favorable region, our analysis suggests that each side's apparent obsession to conceal as much as possible is probably ill-founded.

It is worth stressing that stability, except of the noncooperative kind, is impossible unless detection thresholds are reached and surpassed. In the absence of such stability achieved by equilibrium or inducement, verification will surely fail, for the rational strategies of players -- to violate and challenge -- will undermine it. Thus, an overarching goal of verification, we believe, must be to develop procedures for exceeding the thresholds at which cooperative behavior is stabilized. The players have a mutual interest in being both more compliant and more willing to develop and allow the use of better detection methods.

## APPENDIX

Maximin Strategies and Values

We begin by calculating the maximin strategies and values for the players in the Verification Game with detection, defined in section 3, where we showed that the expected payoffs of ROW and COL are

$$E_R(s; t, u) = s(u + ty)r_3 + (1 - s)(t + u - tx) + (1 - s)(1 - t - u + tx)r_2 \quad (A1)$$

$$E_C(t, u; s) = s(u + ty) + s(1 - u - ty)c_3 + (1 - s)(1 - t - u + tx)c_2 \quad (A2)$$

Observe that  $E_R$  can be rewritten as

$$E_R(s; u, t) = (1 - s)r_2 + u[sr_3 + (1 - s)(1 - r_2)] + t[syr_3 + (1 - x)(1 - r_2)] \quad (A3)$$

Since the coefficients of  $u$  and  $t$  in (A3) are non-negative, it follows that

$$\min_{u, t} E_R(s; u, t) = E_R(s; 0, 0) = (1 - s)r_2,$$

which is maximized by  $s = 0$ . Thus, ROW's maximin strategy is  $s = 0$ , and its maximin value is  $r_2$ . These maximin results are identical to those for VG without detection.

To find COL's maximin strategy, note from (A2) that

$$\frac{\partial E_C}{\partial s} = (c_3 - c_2) + u[1 - c_3 + c_2] + t[y(1 - c_3) + (1 - x)c_2] > 0. \quad (A4)$$

From (A4) it follows that

$$\min_s E_C(u, t; s) = E_C(u, t; 0) = (1 - u - t + tx)c_2,$$

so COL's maximin value is  $c_2$ , and its maximin strategy is  $u = 0, t = 0$  (or, if  $x = 1, u = 0, t$  arbitrary). Again, these maximin results are the same as for the no-detection case.

### Inducement by COL

Inducement occurs when COL announces in advance its (mixed) strategy, inviting ROW to make its best response. In some detection games (Brams and Davis, 1985; Brams, 1985, ch. 4), COL can do better by means of this strategem than by the players' simultaneously selecting maximin strategies.

Since ROW has only two pure strategies, COL can induce only one or the other (C or  $\bar{C}$ ). Suppose that COL has chosen some  $u$  and  $t$ , and consider what ROW's best response is. From (A1) it follows that

$$\frac{\partial E_R}{\partial s} = -r_2 + u[r_2 + r_3 - 1] + t[yr_3 - (1 - x)(1 - r_2)]. \quad (A5)$$

Assume first that  $x(1 - r_2) + yr_3 \leq 1$ . Then

$$yr_3 - (1 - x)(1 - r_2) = x(1 - r_2) + yr_3 - 1 + r_2 \leq r_2,$$

so by (A5),

$$\frac{\partial E_R}{\partial s} \leq -r_2 + ur_2 + tr_2 = -r_2 + r_2(u + t) \leq 0$$

since  $u + t \leq 1$ .

We have shown that  $s = 0$  is ROW's best response to any strategy of COL if  $x(1 - r_2) + yr_3 \leq 1$ . Thus, COL cannot induce  $s = 1$  in this case. This puts COL in an unfavorable position, for it is easy to verify from (A2) that COL cannot receive more than  $c_2$  if  $s = 0$ , and cannot receive less than  $c_3 > c_2$  if  $s = 1$ .

Now assume that (see Figure 2)

$$x(1 - r_2) + yr_3 > 1. \quad (A6)$$

Then  $yr_3 - (1 - x)(1 - r_2) > r_2$ , and it is easy to show from (A5) that

$\frac{\partial E_R}{\partial s} > 0$ , so COL induces  $s = 1$  iff (if and only if)

$$t > \frac{r_2 + u(1 - r_3 - r_2)}{yr_3 - (1 - x)(1 - r_2)}. \quad (A7)$$

But  $t \leq 1 - u$  is also required, so  $u$  must satisfy

$$\frac{r_2 + u(1 - r_3 - r_2)}{yr_3 - (1-x)(1-r_2)} < 1 - u,$$

which can be proven to be equivalent to

$$u < \frac{x(1 - r_2) - 1 + yr_3}{x(1 - r_2) - r_3 + yr_3}. \quad (\text{A8})$$

[From (A6), the right side of (A8) is well-defined; the numerator is positive and less than the denominator.] Hence, we have shown that, if (A6) holds, COL can induce  $s = 1$  by choosing any  $u \geq 0$  satisfying (A8), and then any  $t \leq 1 - u$  satisfying (A7).

We now ask which, among the choices of  $u$  and  $t$  that induce  $s = 1$ , does COL most prefer? This choice will make COL's inducement optimal. From (A3) it follows that

$$E_C(u, t; 1) = c_3 + (1 - c_3)(u + ty),$$

so COL maximizes its expected payoff by choosing  $(u + ty)$  as large as possible.

If  $y < 1$ , COL must choose

$$u = \left[ \frac{x(1 - r_2) - 1 + yr_3}{x(1 - r_2) - r_3 + yr_3} \right]^-; \quad t = 1 - u = \left[ \frac{1 - r_3}{x(1 - r_2) - r_3 + yr_3} \right]^+. \quad (\text{A9})$$

If  $y = 1$ , any  $u$  satisfying (A8) and  $t = 1 - u$  is sufficient.

We have shown that, if (A6) fails, COL cannot induce any strategy of ROW other than  $s = 0$ ; COL then obtains at most  $c_2$ . If (A5) holds, COL can induce either  $s = 0$  or  $s = 1$ .

We next show that optimal inducement of  $s = 1$  is indeed beneficial to COL. Substitution of (A9) and  $s = 1$  into (A2) yields

$$E_C^* = E_C(u, t; 1) = \left[ \frac{x(1 - r_2) - (1 - y)(1 - c_3 + c_3r_3)}{x(1 - r_2) - (1 - y)r_3} \right]^- \quad (\text{A10})$$

[The right side of (A10) is well-defined since the denominator exceeds the numerator, and the numerator exceeds  $x(1 - r_2) - (1 - y) > x(1 - r_2) + yr_3 - 1 > 0$ , by (A6).]

Next we show that optimal inducement yields COL strictly more than  $c_3$

by evaluating

$$\begin{aligned} & x(1 - r_2) - (1 - y)(1 - c_3 + c_3 r_3) - x(1 - r_2)c_3 - (1 - y)r_3 c_3 \\ & = x(1 - r_2)(1 - c_3) - (1 - y)(1 - c_3) \\ & = (1 - c_3)[x(1 - r_2) - (1 - y)] > 0. \end{aligned}$$

It is now clear that  $c_3 < E_C^* < 1$ .

Finally, we determine the effects of COL's optimal inducement on ROW by substituting (A9), and  $s = 1$ , into (A1):

$$E_R^* = E_R(1; u, t) = r_3 \left[ \frac{x(1 - r_2) - 1 + y}{x(1 - r_2) - r_3 + yr_3} \right]^{-1} \quad (\text{A11})$$

[Again it is easy to use (A6) to show that the fraction on the right side of (A11) is well-defined.] To show that COL's optimal inducement yields ROW more than  $r_2$ , we calculate

$$\begin{aligned} & x(1 - r_2)r_3 - (1 - y)r_3 - x(1 - r_2)r_2 + (1 - y)r_2 r_3 \\ & = (1 - r_2)[x(r_3 - r_2) - (1 - y)r_3] > 0 \end{aligned}$$

because, by (A6),

$$yr_3 - (1 - x) - xr_2 > 0,$$

so that

$$yr_3 - r_3(1 - x) - xr_2 > 0.$$

It now follows that  $r_2 < E_R^* < r_3$ .

### Nash Equilibria

We now determine all Nash equilibria of VG with detection. To avoid trivialities, assume that the detector satisfies  $x > 0$  and  $y > 0$ .

We first identify all those equilibria with  $s = 1$ . From (A2),

$$E_C(u, t; 1) = c_3 + (u + ty)(1 - c_3),$$

so that at any equilibrium with  $s = 1$ , COL must maximize  $(u + ty)$  because  $1 - c_3 > 0$ . Thus, for an equilibrium with  $s = 1$ , either  $u = 1$ ,  $t = 0$ , or if  $y = 1$ ,  $u$  is arbitrary and  $t = 1 - u$ .

To begin with, suppose  $y < 1$ . If  $u = 1$ ,  $t = 0$ , (A1) shows that

$$E_R(s; 1, 0) = 1 - s(1 - r_3).$$

But since  $1 - r_3 > 0$ , the choice of  $s = 1$  does not maximize ROW's expected payoff; therefore, there are no equilibria with  $s = 1$  if  $y < 1$ .

We next search for equilibria with  $s = 1$  under the assumption that  $y = 1$ . If  $u = 1 - t$ , then (A1) shows that

$$E_R(s; 1 - t, t) = 1 - tx(1 - r_2) + s[tx(1 - r_2) - (1 - r_3)].$$

It follows that  $s = 1$  maximizes ROW's expected payoff iff

$$t \geq \frac{1 - r_3}{x(1 - r_2)},$$

which, because  $t \leq 1$  is required, can be arranged iff  $x \geq \frac{1 - r_3}{1 - r_2}$ .

It can be verified that these necessary conditions are also sufficient.

We can now characterize all Nash equilibria with  $s = 1$  in VG:

---

I. Cooperative Equilibrium. There is an equilibrium with  $s = 1$  iff

$x \geq \frac{1 - r_3}{1 - r_2}$  and  $y = 1$ . In this case, these equilibria are precisely those combinations of strategies satisfying

$$s = 1; u = 1 - t, t \geq \frac{1 - r_3}{x(1 - r_2)}.$$

Payoffs at any equilibrium of type I are  $(r_3, 1)$ .  $\square$

---

Analogously, we can identify all equilibria with  $s = 0$ . First, from (A2),

$$E_C(u, t; 0) = [1 - u - t(1 - x)]c_2,$$

so COL maximizes its expected payoff by picking  $u = 0$ ,  $t = 0$ , or, if  $x = 1$ ,  $u = 0$ ,  $t$  arbitrary. If  $x < 1$ , then by (A1),

$$E_R(s; 0, 0) = (1 - s)r_2,$$

so  $s = 0$  is indeed a best response for ROW. If  $x = 1$ , then (A1) shows that

$$E_R(s; 0, t) = r_2 + s[tyr_3 - r_2],$$

so  $s = 0$  maximizes ROW's expected payoff iff

$$t \leq \frac{r_2}{yr_3}.$$

(This inequality holds for any choice of  $t$  if  $y \leq r_2/r_3$ ; if  $y > r_2/r_3$ , some values of  $t$  are excluded.) Because the above necessary conditions can also be shown to be sufficient, we have

II. Noncooperative Equilibria. There are essentially two equilibria of this type. The strategy combination

$$s = 0; u = 0, t = 0$$

is type IIa. There are no other equilibria with  $s = 0$  unless  $x = 1$ , in which case the equilibria with  $s = 0$  are precisely those strategy combinations satisfying

$$s = 0; u = 0, t \leq \frac{r_2}{yr_3},$$

which is type IIb. At any type II equilibrium, payoffs are  $(r_2, c_2)$ .  $\square$

Having identified all equilibria with  $s = 0$  or  $s = 1$ , we turn to the case  $0 < s < 1$ . Differentiation of (A1) yields

$$\frac{\partial E_R}{\partial s} = -r_2 + u[r_2 + r_3 - 1] + t[yr_3 - (1 - x)(1 - r_2)]. \quad (\text{A12})$$

Observe that  $\frac{\partial E_R}{\partial s} = 0$  at any equilibrium with  $0 < s < 1$ . It follows that there are no such equilibria with  $t = 0$ , for, if  $t = 0$ ,

$$\frac{\partial E_R}{\partial s} = -r_2 + u[r_2 + r_3 - 1] \leq \max\{-r_2, r_3 - 1\} < 0$$

by (A12). Now suppose that (see Figure 2)

$$x(1 - r_2) + yr_3 < 1.$$

Then

$$yr_3 - (1 - x)(1 - r_2) = yr_3 + x(1 - r_2) - (1 - r_2) < 1 - (1 - r_2) = r_2,$$

so that (A12) yields

$$\frac{\partial E_R}{\partial s} < -r_2 + ur_2 + tr_2 = -r_2 + (u + t)r_2 \leq 0$$

since  $u + t \leq 1$ . Again, no equilibrium with  $0 < s < 1$  can exist.

Thus, to find all equilibria not of types I or II, we may assume that (see Figure 2)

$$x(1 - r_2) + yr_3 \geq 1; \quad (\text{A13})$$

further, we need only consider strategies for COL with  $t > 0$ . First, consider the case  $u = 0$ ,  $t = 1$ . Substitution in (A12) shows that

$$-r_2 + [yr_3 - (1 - x)(1 - r_2)] = 0,$$

which is equivalent to

$$x(1 - r_2) + yr_3 = 1,$$

is necessary. From (A2),

$$E_C(u, t; s) = sc_3 + (1 - s)c_2 + u[s(1 - c_3) - (1 - s)c_2] + t[sy(1 - c_3) - (1 - s)(1 - x)c_2]$$

which we write as

$$E_C(u, t; s) = H + uK + tL, \quad (\text{A14})$$

with  $H$ ,  $K$ , and  $L$  defined appropriately. By calculus, the choice  $u = 0$ ,  $t = 1$  maximizes (A14) iff

$$L \geq 0, L \geq K.$$

The condition  $L \geq 0$  is easily seen to be equivalent to

$$s \geq s_\ell(x, y) = \frac{(1 - x)c_2}{y(1 - c_3) + (1 - x)c_2},$$

and the condition  $L \geq K$  is equivalent to

$$s \leq s_u(x, y) = \frac{xc_2}{(1 - y)(1 - c_3) + xc_2}.$$

It is obvious that  $0 \leq s_\ell < 1$ , with equality iff  $x = 1$ , and  $0 < s_u \leq 1$ , with equality iff  $y = 1$ . Finally,

$$s_u - s_\ell = c_2(1 - c_3)[x + y - 1] > 0$$

since  $x + y > x(1 - r_2) + yr_3 = 1$ .

It can be verified that these necessary conditions are also sufficient for an equilibrium with  $u = 0$ ,  $t = 1$ . In summary, we have

---

III. Constant-Detection Equilibrium. There is an equilibrium with

$0 < s < 1$ ,  $u = 0$ , and  $t = 1$  iff  $x(1 - r_2) + yr_3 = 1$ . In this case, such equilibria are precisely those strategy combinations satisfying

$$s_\ell \leq s \leq s_u; t = 1, u = 0.$$

Payoffs at any equilibrium of type III are

$$(yr_3, xc_2 + s[c_3 + y(1 - c_2) - xc_2]). \quad \square$$


---

Next, under the assumption that (A13) holds, we search for equilibria with  $0 < t < 1$  and  $u = 0$ , and, of course,  $0 < s < 1$ . Substitution in (A12) shows that

$$-r_2 + t[yr_3 - (1 - x)(1 - r_2)] = 0,$$

which is equivalent to

$$t = t_\ell(x, y) = \frac{r_2}{yr_3 - (1 - x)(1 - r_2)}.$$

[As noted above, (A13) ensures that the denominator of this fraction is at least  $r_2$  so that  $0 < t_\ell \leq 1$ , with equality precisely when equality obtains in (A13).]

Now consider (A14). If the choice  $0 < t < 1$  and  $u = 0$  maximizes COL's expected payoff, it is necessary that  $L = 0$  and  $K \leq 0$ . As above,  $L = 0$  iff  $s = s_\ell(x, y)$  and, assuming this to be true,  $K \leq 0$  iff

$$-c_2 + s_\ell[1 - c_3 + c_2] \leq 0.$$

This relationship can be shown to be equivalent to

$$c_2(1 - c_3)[1 - x - y] \leq 0,$$

which is certainly true, since, as noted earlier, (A13) ensures that  $x + y > 1$ .

Because the above necessary conditions are also sufficient, we have

---

IV. Never-Accept (Unconditionally) Equilibrium. There is an equilibrium with

$0 < s < 1$ ,  $u = 0$ , and  $0 < t < 1$  iff  $x(1 - r_2) + yr_3 > 1$ . In this case, such equilibria are precisely those strategy combinations satisfying

$$s = s_\ell; u = 0, t = t_\ell.$$

Payoffs at any equilibrium of type IV are

$$\left( \frac{yr_2r_3}{yr_3 - (1-x)(1-r_2)}, \frac{yc_2(1-c_3) + (1-x)c_2c_3}{y(1-c_3) + (1-x)c_3} \right). \square$$

Analogously, we now assume (A13) and search for equilibria with  $u = 1 - t$ ,  $0 < t < 1$ , and, of course,  $0 < s < 1$ . Substitution in (A12) and simplification shows that

$$t = t_u(x, y) = \frac{1 - r_3}{x(1 - r_2) - (1 - y)r_3}$$

is a necessary condition for such an equilibrium. [It is easy to verify that  $0 < t_u \leq 1$ , with equality precisely when equality holds in (A13).] The choice of  $0 < t < 1$  and  $u = 1 - t$  maximizes (A14) only if  $L = K$  and  $L \geq 0$ . As observed earlier,  $L = K$  iff  $s = s_u(x, y)$ . If  $s = s_u$ ,  $L \geq 0$  iff

$$-(1-x)c_2 + s_u[y(1-c_3) + (1-x)c_2] \geq 0,$$

which is equivalent to

$$c_2(1-c_3)[x+y-1] \geq 0;$$

the latter inequality is a consequence of (A13).

Because these necessary conditions are also sufficient, we have

**V. Never-Challenge (Unconditionally) Equilibrium.** There is an equilibrium with  $0 < s < 1$ ,  $u = 1 - t$ , and  $0 < t < 1$  iff  $x(1 - r_2) + yr_3 > 1$ . In this case, such equilibria are precisely those strategy combinations satisfying

$$s = s_u; u = 1 - t_u, t = t_u.$$

Payoffs at any equilibrium of type V are

$$\left( \frac{x(1-r_2)r_3 - (1-y)r_3}{x(1-r_2) - (1-y)r_3}, \frac{xc_2}{(1-y)(1-c_3) + xc_3} \right). \square$$

We now show that we have identified all Nash equilibria of VG with

detection. If there is an equilibrium not already described, it can exist only when (A13) holds, and it must satisfy  $0 < s < 1$ ,  $0 < u < 1 - t$ , and  $0 < t < 1$ . Consideration of (A14) shows that a necessary condition for such an equilibrium is  $L = 0$  and  $K = 0$ .

We have already noted that  $L = 0$  iff  $s = s_\ell$ , and  $L = K$  iff  $s = s_u$ . Since  $s_u = s_\ell$  can be shown to be equivalent to

$$c_2(1 - c_3)[x + y - 1] = 0,$$

and since  $x + y > 1$  is a consequence of (A13), we conclude that no such equilibrium can exist. Therefore, the only equilibria of VG with detection are those given by I-V.

### Comparisons

When  $x(1 - r_2) + yr_3 < 1$ , VG with detection is simple: ROW's strategy  $s = 0$  (noncompliance) is dominant, and there is a unique (except when  $x = 1$ ) equilibrium (IIa) with payoffs  $(r_2, c_2)$ . On the other hand, the situation is more complicated when  $x(1 - r_2) + yr_3 > 1$ , for even with  $x < 1$  and  $y < 1$  there are two distinct new equilibria, IV and V, as well as the possibility of inducement by COL. We next compare these latter possibilities.

Assume that  $x(1 - r_2) + yr_3 > 1$ ,  $x < 1$ , and  $y < 1$ . First observe that  $s_u > s_\ell$  by a calculation similar to the one above. Also  $t_u > t_\ell$  iff  $1 - r_2 - r_3 > 0$ , and  $t_u = t_\ell$  iff  $1 - r_2 - r_3 = 0$ . Finally, comparison of (A9) with the definition of  $t_u$  shows that COL's optimal inducement strategy is to consult its detector (D) just slightly more often than at the type V equilibrium, where the proportion is  $t_u$ .

We now compare the players' payoffs at type IV and type V to show that V dominates IV, and that both dominate IIa. For COL,  $E_C(\text{IV}) \geq c_2$ , with equality iff  $x = 1$ , is obvious. The inequality  $E_C(\text{V}) > E_C(\text{IV})$  is

equivalent to

$$xc_2[y(1-c_3) + (1-x)c_2] > c_2[(1-y)(1-c_3) + xc_2][y(1-c_3) + (1-x)c_3],$$

which can be shown to hold iff

$$x[y(1-c_2) + (1-x)c_2] > (1-y)[y(1-c_3) + (1-x)c_3].$$

Now  $x > 1 - y$  since (A13) holds, and

$$[y(1-c_2) + (1-x)c_2] - [y(1-c_3) + (1-x)c_3] = (c_3 - c_2)(x+y-1) > 0,$$

again because of (A13). This proves that (A15) holds, and we have

$$E_C(V) > E_C(IV) > E_C(II) = c_2.$$

We now prove analogous inequalities for ROW's payoffs at these three equilibria. First, it is easy to show that  $E_R(IV) \geq E_R(IIa) = r_2$ , with equality iff  $x = 1$ . The inequality  $E_R(V) > E_R(IV)$  can be shown to be equivalent to

$$Q(y) = m - ny + py^2 > 0,$$

where  $m = (1-x)[1-x(1-r_2)]$ ,  $n = (1-x)(1+r_3) - xr_2$ , and  $p = r_3$ .

It is easy to verify that, if  $\frac{1-r_3}{1-r_2} < x < 1$ ,  $Q\left(\frac{1-x+xr_2}{r_3}\right) = 0$ . The minimum of  $Q(y)$  occurs at  $y = \frac{n}{2p}$ , and again, it can be checked that  $\frac{n}{2p} < \frac{1-x+xr_2}{r_3}$ , for  $\frac{1-r_3}{1-r_2} < x < 1$ . This completes the proof that  $E_R(V) > E_R(IV) > E_R(II)$  for  $x(1-r_2) + yr_3 > 1$ ,  $x < 1$ , and  $y < 1$ .

Therefore, of the three equilibria that exist in the region  $x(1-r_2) + yr_3 > 1$ ,  $x < 1$ , and  $y < 1$ ,  $V$  is strictly preferred by both players. It is immediate that  $E_R(V) < r_3$ , and it can be shown that  $E_C(V) \geq c_3$  iff  $xc_2 \geq (1-y)c_3$ .

We now compare the payoffs at the dominant equilibrium,  $V$ , with the payoffs under optimal inducement by COL. We have already noted that COL's strategies are only marginally different in these two situations.

Comparison of  $E_R(V)$  with (A11) shows that ROW receives slightly less under optimal inducement by COL than at  $V$ . [This also proves that  $E_R(V) < r_3$ .]

Using (A10), it can be shown that COL's expected payoff under optimal inducement,  $E_C^*$ , exceeds  $E_C(V)$  exactly when

$$x[1 - r_2 - c_2 + c_2 r_3] + y[1 - c_3 + c_3 r_3] > 1 - c_3 + c_3 r_3. \quad (\text{A16})$$

It is easy to verify that the line defined by equality in (A16) lies below and to the left of  $x(1 - r_2) + yr_3 = 1$  (see Figure 2). Therefore, the region where  $E_C^* > E_C(V)$  is the entire triangle defined by  $x(1 - r_2) + yr_3 > 1$ ,  $x < 1$ , and  $y < 1$ , which is the shaded area shown in Figure 2. Optimal inducement by COL thus increases COL's expected payoff significantly over Equilibrium V. Also, recall that  $E_C^* > c_3$  whenever inducement by COL is possible.

## FOOTNOTES

1. This game (VG) was originally suggested to Steven Brams by Danny Kleinman in 1983; we are pleased to acknowledge its source.
2. The probabilities of error are  $1 - x$  and  $1 - y$ ; in statistics, an error of the latter kind (incorrectly signaling a violation) is a type 1 error, and an error of the former kind (incorrectly signaling compliance) is a type 2 error.
3. We have ignored a fourth pure strategy for COL -- consulting the detector and assuming it is incorrect by choosing A if it signals a violation and  $\bar{A}$  otherwise. Unless  $x$  and  $y$  are small, this fourth strategy would simply complicate the analysis without adding any significant features to the game. In fact, our analysis will focus on relatively large values of  $x$  and  $y$ .
4. Lynn (1985) gives details on bilateral and multilateral agreements and assesses compliance with them.

## REFERENCES

- Avenhaus, Rudolf (1986). Safeguard Systems Analysis. New York: Plenum (forthcoming).
- Brams, Steven J. (1985). Superpower Games: Applying Game Theory to Superpower Conflict. New Haven, CT: Yale University Press.
- Brams, Steven J., and Morton D. Davis (1985). "The Verification Problem in Arms Control: A Game-Theoretic Analysis." In Interaction and Communication in Global Politics, ed. Claudio Cioffi-Revilla, Richard L. Merritt, and Dina A. Zinnes. London: Sage, pp. 133-154.
- Brams, Steven J., and D. Marc Kilgour (1986). "Notes on Arms-Control Verification: A Game-Theoretic Analysis." In Modeling and Analysis of Arms Control Problems, ed. Rudolf Avenhaus and Reiner K. Huber. Berlin: Springer-Verlag (forthcoming).
- Fichtner, John (1986). "Solution Concepts for Two Person Games Which Model the Verification Problem in Arms Control." In Modeling and Analysis of Arms Control Problems, ed. Rudolf Avenhaus and Reiner K. Huber. Berlin: Springer-Verlag (forthcoming).
- Fishburn, Peter C. (1974). "Lexicographic Orders, Decision Rules and Utilities: A Survey." Management Science 20: 1442-1471.
- Lynn, William J. (1985). "Existing U.S.-Soviet Confidence-Building Measures." In Preventing Nuclear War: A Realistic Approach, ed. Barry M. Blechman. Bloomington, IN: Indiana University Press, pp. 24-51.
- Maschler, Michael (1967). "The Inspector's Non-Constant-Sum Game: Its Dependence on a System of Detectors." Naval Research Logistics Quarterly 14: 275-290.

- Maschler, Michael (1966). "A Price Leadership Method for Solving the Inspector's Non-Constant-Sum Game." Naval Research Logistics Quarterly 13: 11-33.
- Nash, John (1951). "Non-cooperative Games." Annals of Mathematics 54: 286-295.
- Potter, William C. (1985). "Introduction." In Verification and Arms Control, ed. William C. Potter. Lexington, MA: Lexington, pp. 1-5.
- Rapoport, Anatol (1966). Two-Person Game Theory: The Essential Ideas. Ann Arbor, MI: University of Michigan Press.
- Richelson, Jeffrey (1985). "Technical Collection and Arms Control." In Verification and Arms Control, ed. William C. Potter. Lexington, MA: Lexington, pp. 169-216.
- von Stackelberg, H. (1934). Marktform und Gleichgewicht. Berlin: Julius Springer.