Attacking the Unknown Weapons of a Possible Provocateur: How Intelligence Affects the Strategic Interaction

Artyom Jelnov * Yair Tauman † Richard Zeckhauser ‡

March 11, 2015

Abstract

We consider the interaction of two enemy nations. Nation 1 wants to develop a nuclear bomb (or other weapons of mass destruction). Nation 2 wants to prevent such a development through the deterrence of a threatened attack, or an actual attack if it thought the bomb was produced. 2 has an intelligence system that imperfectly indicates the presence of a bomb. 1, if lacking the bomb, can open its facilities to prevent an attack. A further uncertainty is that nation 2 does not know nation 1's type. He could be a Deterrer, whose prime goal is to avoid an attack, or he could a Provocateur who prefers an unjustified attack if he does not possess the bomb, so as to build support from inside his nation and the outside world. The game has a unique sequential equilibrium. The qualitative nature of that equilibrium depends on parameters, on preferences and information conditions.

A number of initially counterintuitive results emerge. For example, it may sometimes be rational (an equilibrium strategy) for 2 to attack even though 1 does

*Hebrew University of Jerusalem, Israel and Ariel University, Israel, artyomj@ariel.ac.il

†The Interdisciplinary Center, Herzliya, Israel and Stony Brook University, NY, USA,amty21@gmail.com

‡Harvard University, USA,Richard_Zeckhauser@hks.harvard.edu

not have a bomb, and even though 2's high quality intelligence system indicates that a bomb is not present. Fortunately, intuitive explanations can be provided for all such results.

Illustrations of the model's implications are provided from the experiences of the West (as nation 2) with Saddam Hussein (as nation 1).

1 Introduction

This paper analyzes the interaction between two enemy countries (players). Player 1 wants to possess weapons of mass destruction, or for simplicity the bomb, and has the capability to build it. Player 2 wants to prevent Player 1 from securing the bomb, and is capable of and willing to destroy Player 1's bomb(s) if it exists. It is not clear, however, whether Player 1 possesses the bomb, and Player 2 would lose significant value if she attacked and the bomb did not exist. Uncertainty about that existence drives this analysis.

Player 1 chooses whether or not to build the bomb. Player 2 chooses whether or not to attack. Player 1 also has the capability to open its facility to reveal that it does not possess the bomb, thereby avoiding any potential for an attack by Player 2. In determining whether or not to attack, Player 2 would like to assess whether a bomb was present. To do so, it employs a spying or intelligence system (IS). The system has precision α, $\frac{1}{2} < \alpha < 1$, where α is common knowledge. In other word, the IS will correctly detect the presence or absence of a bomb, each with probability α, and incorrectly, each with probability $1 - \alpha$. Thus, the IS will yield either signal b, bomb present, or signal nb, no bomb present. Based on the signal it receives from the IS, Player 2 will decide whether or not (or with what probability) to attack.

There is a second critical uncertainty, the type of Player 1. Player 1 may be a Deterrer, D, or a Provocateur, P. The critical difference between these two possibilities is that if D does not possess the bomb, he prefers not to be attacked, whereas P prefers to be attacked in this circumstance. The ex ante likelihood that Player 1 is a Provocateur is β,

and β is common knowledge. Player 1 cannot directly reveal his type, even if he would like to. However, he can open his facility to reveal the absence of a bomb. Hereafter, for expository ease, we will refer to Player 1 as 1, with types D and P, and Player 2 as 2. Player 1 is treated as male, and Player 2 as female.

There are four possible outcomes, depending on whether or not the bomb is built, and whether or not 2 attacks. 1's potential actions are build, B, or not build, NB. He can also not build and open his facility, NBO, to reveal that fact. 2's potential strategies are attack, A, or not attack, NA. Given that the players will sometimes employ mixed strategies, their payoffs have to be cardinal (von Neumann-Morgenstern) utilities. Those utilities are common knowledge. The players are assumed to be sophisticated game theorists, and to maximize their expected utilities.

Player 2's preference ordering from best to worst is (NB,NA), (B,A), (NB,A) and (B,NA). Both D and P have (B,NA) as their best outcome and (B,A) as their worst outcome. They differ, however, on the ordering of the other two outcomes. D wishes to avoid an attack if he doesn't build the bomb; hence (NB,NA) is superior to (NB,A). *D*'s payoff is the same for (NB,NA) as it is for (NBO,NA). That is, there is no embarrassment from opening his facility. (If he chooses NBO, there is no chance of an attack.) Similarly, Player 2 receives the same payoff from (NB,NA) and from (NBO,NA). P, the Provocateur, will get sufficient benefit from being attacked that he prefers (NB,A) to (NB,NA). One implication is that he will never employ the strategy NBO, since it avoids the potential for attack when he chooses NB.

The phenomena underlying this preference for P could be either or both that it bolsters support from his constituents and sympathizers, and that it undermines the legitimacy and support of Player 2, its enemy. We observe, in examples well beyond this model, that Provocateurs often welcome attacks for these reasons. Thus, insurgents such as FARC may want to have government forces attack, which usually involves imposing casualties on the surrounding peasant populations, as a means to build support among the peasantry. ISIS engages in outrageous activities to provoke attacks from both Muslim nations and the West, in part because it serves as a device to recruit fighters and money. Hamas

launches generally ineffective rocket attacks on Israel partly as a means to elicit Israeli retaliation, which in turn impairs Israel's legitimacy.

1.1 Introduction to the separating equilibrium

We first address the case where D's payoff from (NB,NA) is relatively high. He will then choose NBO to avoid an attack. Since P will never open his facility, a separating equilibrium results, where D opens and P does not. Thus, if the facility is not opened, 2 knows that she is playing against P. But she does not know whether P has chosen B, to build the bomb. The IS provides probabilistic information on this choice.

Even though 2 knows 1's type, several results from this case are surprising, at least without further reflection. When IS is sufficiently accurate (α exceeds a critical threshold), 2 will not attack P if the signal is nb. No surprise. However, if the signal is b, 2 will not attack with significant probability even though his worst outcome is that 1 has the bomb and she does not attack, (B,NA). Interestingly, if IS is worse (α is below the critical threshold), 2 acts much more aggressively. If the signal is b, she attacks for sure; if it is nb, she attacks with positive probability [1].

Let us provide some intuition for these results. When IS is relatively precise, α is high, 1 knows that if he builds the bomb, 2 will detect that with relatively high probability, and if she concludes that a bomb is present, she will attack. The result is that 1 only builds the bomb with a sufficiently small probability that if 2 gets the unexpected signal b, she will employ Bayesian analysis and conclude that a bomb is unlikely. She will be indifferent between attacking or not given this signal. 2 attacks sufficiently often to hold down the probability that 1 builds the bomb. Indeed, the more precise is IS, the less likely the bomb is to be built, and the less likely an attack given signal b.

When IS is less reliable, matters turn almost topsy turvy. Player 1 builds the bomb with significant probability, knowing that there is a good chance that the bomb will go undetected. The initially surprising result is that if signal b is obtained, it is more likely

[1] In effect, 2 employs a threat that leaves something due to chance. Schelling [1960] first examined the potential for a threat with probabilistic implementation in a quite different context.

to be reliable, since the higher probability of bomb built more than compensates for the lesser reliability of IS. Player 2, observing b, thus attacks for sure so as to avoid (B,NA), her worst outcome. Moreover, she even attacks some of the times when she observes nb. Such attacks have the twin benefits of wiping out a quite possible bomb and also holding down the probability that a bomb is built.

1.2 Introduction to the pooling equilibrium

We next examine the situation when D gets a relatively low payoff from (NB,NA). Then a pooling equilibrium will result. 2 will never know whether she is playing against D or P. The nature of the equilibrium will then depend on 2's assessment about 1's type. If she believes that D is quite likely, then for any value of α, and in every sequential equilibrium, 2 will attack with positive probability even if she receives the signal nb. D randomizes between the two pure strategies of B and NBO. P in some sense free rides on the inability of 2 to distinguish between him and a D playing B, when facilities are not opened. Thus, P chooses NB, and hopes to be attacked. Consider 2's decision when facilities are not opened. She assumes that there is a high probability that D is player 1; if so, it is certain that he has chosen b. 2 attacks. Even if IS is highly accurate (though not perfect), the following scenario is possible (has positive probability) with both players choosing rationally: 1 chooses not to build the bomb, the IS correctly signals nb, yet 2 ignores the signal and unjustifiably attacks 1.

Matters differ greatly if 2 thinks it quite likely that she faces the Provocateur. In this case, even if IS is highly accurate, then both D and P behave more aggressively. D builds the bomb for sure, and P does so with positive (but less than 1) probability. 2, by contrast, tempers her aggression. She does not attack if the signal is nb, and attacks with probability less than 1 if the signal is b. It should be noted that the imperfect nature of IS plays a role in the mixed strategy of 2, here and in other contexts. D would not build the bomb if he knew he would be attacked with high probability. However, the chance of an erroneous signal reduces the likelihood of attack. Note also, that D also free rides here on the ex ante possible presence of P. P will not build the bomb as 2 knows, which

makes it more likely that a b signal is erroneous, which in turn makes A less attractive to 2 when she gets the signal b.

1.3 Real world applications

Three uncertainties are critical in our model: Player 1's type (payoff structure) is not known to 2; whether 1 possesses weapons of mass destruction is not known; and intelligence is imperfect. The second Gulf War (2004) between Iraq and the United States is perhaps the most salient real world case that both has these characteristics and that has played out recently. Iraq lacked weapons of mass destruction; yet it was attacked by the US. Many critics have decried the folly of the US, but at least in the context of this model it is possible that attacking was the rational strategy, even if intelligence was good and even if it correctly signaled no weapons of mass destruction. The strategy of Saddam Hussein, who refused to provide accurate information on his weapons and fully cooperate on opening his facilities even when he had the chance at the last minute when attack was imminent and his demise was virtually certain, however, can only be explained using factors beyond our model. For example, it may have been due to his relationship with Iraqi players who believed that he had weapons of mass destruction.[2] In short, Saddam may have been an unknown type, U, neither a D nor a P.

A driving uncertainty in our model, as mentioned, is the inability of Player 2 to determine 1's type, except of course when there is a separating equilibrium. Closed and authoritarian societies, particularly those with a single figure at the head, are more likely to function with undisclosed types than are open and democratic societies. In most instances, a foreign power can assess the intentions of the United States, India or Israel all open democratic societies – nearly as well as can the leaders of those nations. The foreign powers can listen to speeches, read the press, and even use espionage to assess

[2]Many of the critiques center on the desire of the Bush Administration to depose Saddam Hussein whether or not he had weapons of mass destruction, in effect departing from the assumption of this model that for 2 (NB,A) is inferior to (NB,NA). These critiques also often assert that the intelligence was purposefully misinterpreted to be more conclusive than it was.

preferences. To be sure, preferences may change if a regime changes, which leads to some uncertainties, but outsiders are not significantly handicapped in assessing the likelihood of such changes.

With closed and authoritarian societies, matters are quite different. Saddam Hussein was calling the shots in Iraq until that nation was attacked in 2004, and outsiders did not know what he was thinking. In a situation that closely parallels our model, Iran is a Player 1, with the Ayatollah Khamenei, and perhaps a small circle around him, determining that nation's type. The Ayatollah has continued to make conflicting comments about Iran's intentions on nuclear weapons, about reaching agreement on inspections, etc. His true preferences are impossible to read. Though there is not a question about weapons of mass destruction, Vladimir Putin is also expert in keeping the West guessing about his true intentions on a range of issues. That Western analysts have disagreed strongly about his intentions illustrates his success. Note, frequently the confusing tactics of authoritarian leaders of closed societies are the result of their need to speak to multiple audiences. Putin surely wants to look tough to his domestic audience, even though he might wish to look more accommodating overseas. One illustration of this multiple audience phenomenon is the decision by leaders to use different phrasing and even different subjects when speaking in their home language, usually both are much more pugnacious, than when speaking in English to overseas audiences.

We should also note that the payoffs in our model are intended to take into account that Players 1 and 2 are simultaneously engaged with players other than each other. Thus, if part of these players' payoffs come from domestic audiences, or from other external players, those parts are included in the payoffs of our model. Thus, for example, a considerable portion of the payoff to the Provocateur from being attacked unjustifiably arises because he gains with players other than Player 2, such as outside sympathizers.

1.4 Related literature

Our model relates most closely to the literature on inspection games. Those games apply to situations where an inspector verifies whether an agent(s) adheres to specified rules.

Applications include situations such as arms control and disarmament, environmental regulation, and financial auditing. Avenhaus et al. [2002] provides an extensive survey of this literature. In such games, verification typically involves sampling data generated by the activities of the agents. The agents therefore alter their activities in ways that conceal the true situation. The inspector employs Bayesian methods to assess whether an agent has abided by the rules, and identifies a violation depending on the signal received. However, once agents adopt strategic behavior, the inspector can no longer assume that he has merely observed a random sample. To illustrate, the plant subject to environmental regulation might store effluents and have occasional big releases, hoping to escape the random inspection.

Inspection games are similar to our game in their sequence of moves. First, an agent decides whether or not to adhere to the rules. If he chooses NOT, he selects the violation procedure that sends the optimal noisy signal of his action. The inspector observes the signal and decides whether or not to sound an alarm. The analogy to our game is clear. The agent is like our Player 1, and the inspector is like our Player 2. Moreover, there are multiple types of agents, with different preferences. Thus, in the environmental inspection case, some agents will be able to cheaply adhere to the regulations. They will definitely comply. Given many cheap adherers in the population, agents with high costs are advantaged. The inspector, employing Bayesian methods, will start with a prior that is favorable to them. Some signals that would merit an alarm in a less favorable population, will not lead to an alarm.

One important difference between our model and a typical inspection game is that in the latter, by auditing the agent, the inspector can detect with certainty whether or not he adhered to the rules, before possibly taking tough measures against him. In our model Player 2's tough action of attacking (and destroying) Player 1's facility is taken under uncertainty since she can't detect with certainty 1's action. Another difference is that our type P agent prefers to be unjustifiably punished by 2. This can't happen in inspection games. Finally, the inspector in inspection games is allowed to alter the nature of the inspection, e.g., make it more intense or more frequent, to deter bad behavior. In addition

agents in inspection games can manipulate signals by the action they take. Thus, the quality of the inspection is the product of an equilibrium. In our game, by contrast, the quality of the inspection is intrinsic to the IS, and Player 1 has no ability to influence the signals they send given the action they took. Of course, future versions of our model could allow for the quality of the IS, the α, to be endogenously determined. Player 2 would then choose the optimal α by comparing marginal benefit with marginal cost. The benefit however may also depend on possible actions by Player 1 to conceal the information of whether and where he built the bomb.

Our model relates more broadly, though more distantly, to a great variety of models in the arms building, nuclear deterrence, and arms control fields, and more generally to military strategy. O'Neill [1994] provides an extensive survey of this literature. The classic work that examines how one player's action affects another's behavior, including in particular the role of threats (equivalent to the threat of attack in our model), is Schelling's "The Strategy of Conflict" [1960].

Most of the literature on game theory applied to military affairs deals with attackers and defenders. In 1917, an aged Thomas Edison, best known for his work producing technological breakthroughs, analyzed the problem of how to enable transport ships to break through the dangers represented by German U-boats, and thus secure safe passage to British ports. General discussions utilizing military examples can be found in Dresher [1961, 1968], Thomas [1964], Shubik [1983, 1985], Finn and Kent [1985] and O'Neill [1993]. Some papers deal with missile attack and defense. The best known conception of the subject is the Prim-Read theory, which is based on Read Jr [1958, 1961] and Karr [1981]. In a simple version of the model, an attacker sends missile warheads to destroy some fixed targets, while the defender tries to protect the targets using interceptors that are themselves missiles .

Most intelligence assessments are made by human beings, who examine disparate pieces of data and conclude, for example, whether the enemy possesses a certain capability. It is important to recognize that IS is a machine, and not a human assessor. Earlier work does look at the value of information in strategic conflicts. See Kamien et al. [1990].

Finally, this paper significantly extends the unpublished paper of Biran and Tauman [2009], hereafter BT. In BT, Player 1 is always a Deterrer. Moreover, he has no ability to open his facility for inspection, which makes (NB,A) much more likely in a range of circumstances. The lack of types in BT also helps lead it to quite different results from those found here.

2 The Model

There are two players, who are enemies. Player 1 has the capability to build a nuclear bomb, and would like to possess one. Player 2 would regard such a bomb as a severe threat, and has the capability to attack and destroy it if it exists. Player 1 moves first and can build, B, or not build, NB, the bomb. That move is secret. However, if he chooses NB, he can also open his facility, thereby choosing NBO, in order to prove no bomb. Once Player 1 has moved, Player 2 must decide whether to Attack, A, or not attack, NA.

Player 1 can be one of two types, a Deterrer, D, or a Provocateur, P. The likelihood that he is a P is β, which is common knowledge. The Provocateur would like to be attacked unjustly, i.e., when he does not have the bomb. However, Player 2 has no way to know Player 1's type, though he does learn it if 1 chooses NBO. (Hereafter the players will usually be denoted as 1 and 2.) Either D or P regards the outcome (B,NA) as best and (B,A) as worst of the four possibilities. However, D regards (NB,NA) as superior to (NB,A), whereas P, prefers (NB,A) to (NB,NA).

Player 2's best outcome is when no bomb is built and no attack is made, (NB,NA). His second best is a bomb is built and destroyed, (B,A). Third down is an unjustified attack, (NB,NA), since she will suffer a severe loss of legitimacy. His worst outcome is that 1 possesses the bomb undisturbed (B,NA).

Player 1

Player 2	Deterrer, D: NB	Deterrer, D: B	Provocateur, P: NB	Provocateur, P: B
NA	$1, w_D$	$0, 1$	$1, r_P$	$0, 1$
A	r_2, r_D	$w_2, 0$	r_2, w_P	$w_2, 0$

Assumption $0 < r_i < w_i < 1$, $i = D, P, 2$.

Figure 2.1: The game and payoffs

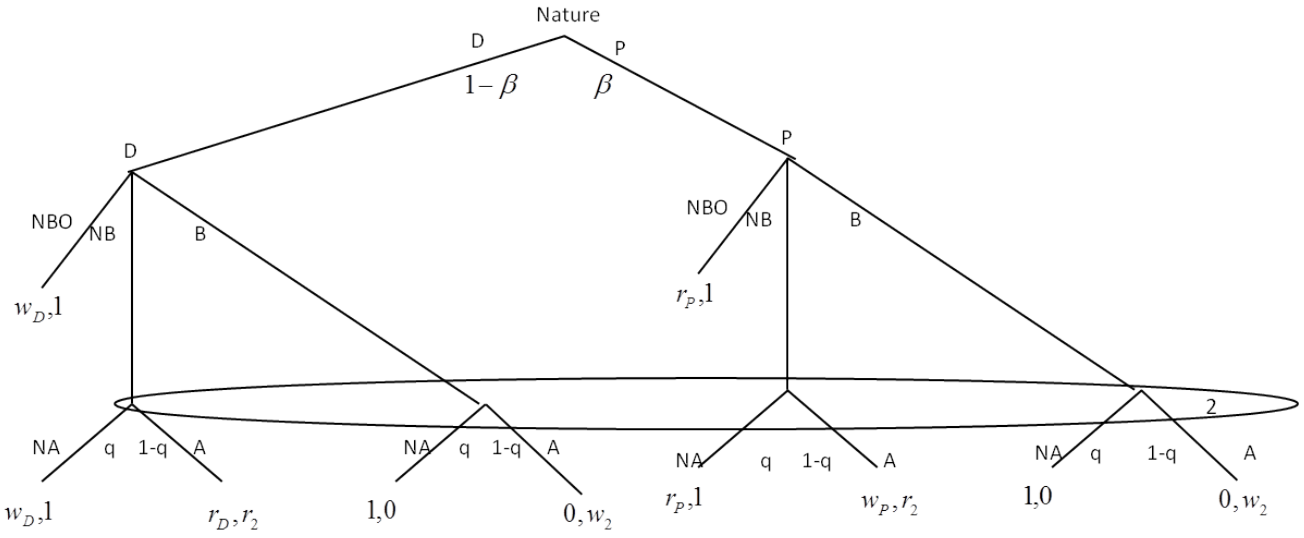

Figure 2.2: The game G_β

2.1 The case without IS

The game matrix, indicating only ordinal payoffs, is shown as Figure 2.1 below. Note, since players will often be employing mixed strategies, it will often be necessary to know the cardinal values of payoffs, as von Neumann-Morgenstern utilities.

The game is shown in tree form as game G_β in Figure 2.2. That game has two weakly dominated strategies: NB for D is inferior to NBO; NBO for P is inferior to NB. They are eliminated, producing the reduced tree form of the game presented in Figure 2.3.

Consider first situations where 1's type is known, namely $\beta = 0$ (he is a Deterrer) and $\beta = 1$ (he is a Provocateur). If 1 is a known Deterrer, should he not open his facility, 2 will know for sure he has chosen B. Thus 2 will attack with certainty, leading to D's

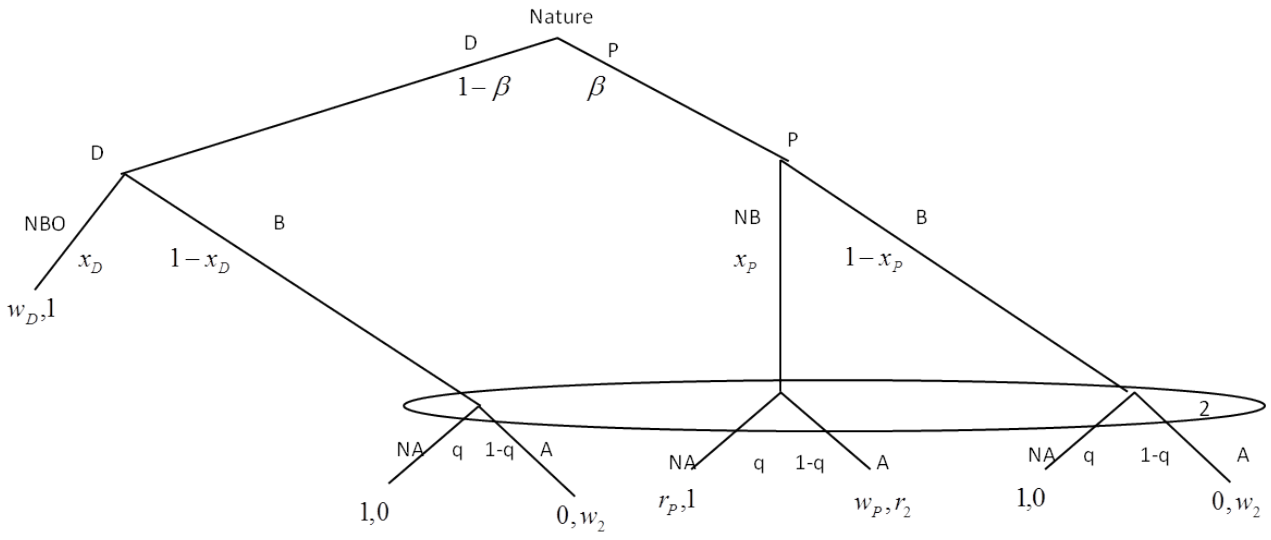

Figure 2.3: The reduced game G_β

worst outcome. Thus, the B strategy is struck from D's arsenal, and the equilibrium is (NBO,NA). 2 gets his best outcome; D gets his second best. The situation when 1 is a known Provocateur is more complex, thus leading to mixed strategies. Denote by x_P the probability that P chooses the strategy NB, and by q the probability that 2 chooses the strategy NA.

Lemma 1 Suppose that Player 1 is a Provocateur. The game has a unique subgame perfect Nash equilibrium. It satisfies:

$$x_P = \frac{w_2}{1 - r_2 + w_2},$$

$$q = \frac{w_P}{1 - r_P + w_P}$$

The equilibrium strategies of both P and 2 are quite intuitive. Note that w_2 and r_2 are the payoffs of 2 if she attacks P (respectively, with or without justification). The greater is either w_2 or r_2, or both, the greater is 2's incentive to attack P. This implies in turn that the probability P will not build the bomb, x_P, will increase. Similarly, r_P and w_P are the payoffs to P if he does not build the bomb. The higher is either r_P or w_p (or both) the greater is the incentive to P to not build the bomb, and as a result the higher is the probability, q, of 2 of not attacking P.

Proof: For P the strategy NBO is dominated by NB. P randomizes his two pure strate-

gies B and NB. Thus $0 < x_P < 1$ and P is indifferent between NB and B. Similarly $0 < q < 1$ and 2 is indifferent between her two pure strategies A and NA. The proof is now straightforward (see Figure 2.3 for $\beta = 1$). $\square$

Consider next the incomplete information case where 2 does not know the type of 1, but where she assigns probability β, $0 < \beta < 1$, that 1 is a Provocateur.

Proposition 1: Suppose $w_D \neq \frac{w_P}{1-r_P+w_P}$. The game G_β has a unique sequential equilibrium. It satisfies the following.

(i) 2 mixes her two pure strategies NA and A.

(ii) If $w_D > \frac{w_P}{1-r_P+w_P}$, D chooses NBO with probability 1, and P mixes B and NB.

(iii) If $w_D < \frac{w_P}{1-r_P+w_P}$ and $\beta > \frac{w_2}{1-r_2+w_2}$, D builds the bomb (B) with probability 1, and P mixes B and NB.

(iv) If $w_D < \frac{w_P}{1-r_P+w_P}$ and $\beta < \frac{w_2}{1-r_2+w_2}$, D mixes NBO and B, and P chooses NB with probability 1.

Proof See Appendix.

Remark: The case $w_D = \frac{w_P}{1-r_P+w_P}$ produces multiple sequential equilibrium outcomes.

Let us provide some intuition for this result, which revolves around the payoffs for the three players, D, P and 2, for their second-best, the w_i values, and their third-best, the r_i values, outcomes. If D's payoff, w_D, when he opens his facilities (NBO) and avoids an attack is sufficiently high, he will choose this strategy with probability 1. If it is less high, he will build the bomb (B) with positive probability. Suppose that 2 is relatively confident that 1 is of type D (part iv). Then if 1 does not open his facilities, D chooses B for sure. Given this parlay, D highly likely and D chooses B, 2 will conclude that the likelihood is great that 1 chose B. In response, 2 attacks with high probability. The best reply strategy of the Provocateur, who fears to be wrongly identified, is to not build the bomb.

By contrast, if 2 assigns a high probability β that 1 is a Provocateur, and if in addition 1 does not allow inspection (as is consistent with relatively small w_D), then 2 even raises her belief that 1 is a P. Since P does well with an unjustified attack, he chooses NB with significant probability. In response, 2 reduces the probability of an attack as compared to lower values of β. In short, if you are 2 and believe that you likely face a Provocateur, seek to avoid an unjustified attack. With the attack threat reduced, D's best reply is to build the bomb for sure. In this case, D acts more aggressively than P.

Proposition 1 (part ii) shows that that in the unique equilibrium there is a positive probability that 2 will attack the Provocateur, even though he has not built the bomb. That is what being a Provocateur is all about. This surprising outcome could apply to the outcome in the Second Gulf War (2004), where the US attacked although Saddam Hussein did not have weapons of mass destruction. We shall save further discussion of this illustration to the next section, where we consider the case with IS. After all, intelligence supposedly played a major role in justifying that attack.

2.2 The case with IS

Player 2 has an Intelligence System, IS, with quality α. The system is not relevant if 1 opens his facility to reveal NB. If 1 chooses B, the IS will send the correct signal b with probability α, and nb with probability $1-\alpha$. If 1 chooses NB, and does not open his facilities, the IS will send the correct signal nb with probability α, and b with probability $1-\alpha$.

Player 2 gets the signal, and then decides whether to attack, A, or not, NA. Without loss of generality, we assume that $\frac{1}{2} < \alpha < 1$. Its value is common knowledge. Thus 1 knows that he is being spied upon, and he knows the reliability of 2's intelligence . The game proceeds as follows. (1) The values of α and β were revealed as common knowledge. (2) 1 chooses between B and NB, and if NB whether he should choose variant NBO and open his facilities. (3) If 1 chooses NBO, 2 chooses NA, and the game ends. (4) If 1 does not open his facilities, his choice of B or NB sends a signal via IS to 2. (5) 2 draws inferences from the signal and chooses whether to attack A, or not attack NA. Steps (2)

and (5) may involve mixed strategies. This describes a game $G_{\alpha,\beta}$.

2.2.1 Player 1 type known

Note, if it is known that 1 is type D, $\beta = 0$, then the equilibrium turns out to be the equivalent of the no-IS case. Player 1 chooses the pure strategy NBO and player 2 chooses the pure strategy NA. This neat use of pure strategies vanishes, however, if 1 is known to be a Provocateur, $\beta = 1$. Denote the critical value for α, call it α_P, if 2 faces a Provocateur:

$$\alpha_P = \frac{1 - w_P}{1 - w_P + r_P}$$

Note, that $\alpha_P < \frac{1}{2}$ iff $1 - r_P < w_P$. To interpret the last inequality suppose hypothetically that 2 does not posses an IS capability. There are two possible ex-post mistakes that P may commit. The first one (type I) occurs when P does not build the bomb even though 2 decides not to attack him. The second mistake (type II) occurs when P builds the bomb even though 2 decides to attack him. The cost associated with a type I mistake is $1 - r_P$, whereas it is w_P for a type II mistake. If the cost for P of a type I mistake is smaller than that of a type II mistake then $\alpha_P < \frac{1}{2}$. In this case $\alpha > \alpha_P$ for all $\frac{1}{2} < \alpha < 1$.

Proposition 2: Let $1/2 < \alpha < 1$ and suppose Player 1 is a Provocateur. The game has a unique subgame perfect Nash equilibrium. It satisfies:

(i) Suppose that $\alpha > \alpha_P$. Player 2 will not attack P if the signal is nb and will randomize between A and NA P if the signal is b. The probability that P is building the bomb is decreasing to zero in α.

(ii) Suppose that $\alpha < \alpha_P$. Player 2 attacks P with probability 1 if the signal is b. She randomizes between attacking and not attacking P if the signal is nb. The probability that P is building the bomb is increasing in α, for $\frac{1}{2} < \alpha < \alpha_P$.

(iii) The expected payoff of P is decreasing in α, the expected payoff of Player 2 is increasing in α, for $\frac{1}{2} < \alpha < \alpha_P$ and for $\alpha_P < \alpha < 1$.

Proof: See Appendix

Many elements of Proposition 2 seem counter intuitive, since 2 reacts more aggressively

the lower is the quality of IS. Yet intuition returns when we recognize that 2 basically has a Bayesian decision problem. He must take into account how a Provocateur will respond to weaker IS. He then factors in the actions P would take with what probability. Thus, 2 may correctly conclude that when his intelligence is less reliable, P is more likely to build a bomb, which makes it more attractive to attack. Let us now elaborate a few of the specifics underlying this general lesson.

As with Proposition 1, there are critical values for α. If the precision of IS is relatively low ($\alpha < \alpha_P$) and the signal is less reliable, 2 attacks P with no hesitation (namely with probability 1) if the signal of IS is b. 2 still attacks P with positive probability even if the signal is nb. While if IS is of high quality ($\alpha > \alpha_P$), and therefore more reliable, 2 hesitates to act even when the signal is b. She attacks P with probability smaller than 1. If the signal is nb, she never attacks P. Let us provide some intuition for these results. Suppose first $\alpha_P < \alpha < 1$ (the precision of the IS is relatively high). In this case P knows that with relatively high probability his action will be correctly detected by IS. Thus, he chooses to build the bomb with low probability, which decreases to zero as α increases to 1. Consequently, for large α, Player 2 does not expect the signal b; when it does appear, she concludes it is likely that the IS sent the wrong signal. Player 2 updates her belief about the probability that Player 1 chose B, when the signal is b. It can be shown (see the proof of Proposition 1) that

$$Prob(B|b, \alpha > \alpha_P) = \frac{1 - r_2}{1 - r_2 + w_2},$$

which is bounded away from 1. This is the reason why Player 2 acts with caution even if the IS is highly accurate and the signal is b.

Suppose next that $\frac{1}{2} < \alpha < \alpha_P$, that IS is somewhat informative, but it is not too accurate. The Provocateur now builds the bomb with significant probability, expecting that there is a reasonable chance that it will not be detected. Here the signal b hardly surprises 2, and she attacks for sure. 2 even attacks some of the time when she receives the signal nb. First, she knows that with weak IS and a strong incentive for P to select B, there is a reasonable chance that the true state is B. Second, by attacking some of the

time despite nb, she reduces P's incentive to build the bomb. Remember, 2 far prefers to be embarrassed by an unjustified attack (NB,A) than to leave a bomb in the hands of Player 1 (B,NA).

2's conclusion that weak intelligence raises the likelihood that P has a bomb applies to both signals, b and nb.

$$Prob(B|s, \alpha < \alpha_P) > Prob(B|s, \alpha > \alpha_P), s \in \{b, nb\}.$$

The logical and correct implication is that 2 attacks with higher probability when $\alpha < \alpha_P$. Indeed, in the extreme case where IS is perfect, $\alpha = 1$, P never builds a bomb and 2 never attacks.

When 1's type is known, 2's expected payoff is higher if his intelligence is better. With D, the result is trivial, since the equilibrium is (NBO,A) regardless of α; 2 gets his best outcome and a payoff of 1. With P, the greater is α, the more accurate is 2's information; hence she responds more effectively, which in turn makes P less likely to build the bomb, as 2 would like. 2's payoff increases. The value of α matters not to D. P prefers α to be lower, since he benefits from either of 2's errors.

2.2.2 Player 1 type not known

A prime motivation for this analysis is to understand what happens when Player 2 does not know Player 1's type. The remainder of the paper is devoted to this case. We continue to assume that 2 has intelligence about 1's move, but that it is imperfect. Thus, $\frac{1}{2} < \alpha < 1$. The nature of the equilibrium will depend on the likelihood that 1 is a Provocateur, namely the value of β.

Let

$$\beta_1 = \frac{(1-\alpha)w_2}{(1-\alpha)w_2 + \alpha(1-r_2)},$$

and

$$\beta_2 = \frac{\alpha w_2}{(1-\alpha)(1-r_2) + \alpha w_2}.$$

Let

$$V_1(\alpha) = \frac{(1-\alpha)w_P}{\alpha(w_P - r_P) + 1 - \alpha},$$

and

$$V_2(\alpha) = \frac{(1-\alpha)(w_P - r_P) + \alpha r_P}{(1-\alpha)(w_P - r_P) + \alpha},$$

and let

$$V(\alpha) = \max\{V_1(\alpha), V_2(\alpha)\} = \begin{cases} V_2(\alpha) & , \alpha > \alpha_P \\ V_1(\alpha) & , \alpha < \alpha_P. \end{cases}$$

Note that $1-\alpha > V_1(\alpha) > V_2(\alpha)$ if $\frac{1}{2} < \alpha < \alpha_P$ and $1-\alpha < V_1(\alpha) < V_2(\alpha)$ if $\alpha_P < \alpha < 1$.

Proposition 3 Consider the game $G_{\alpha,\beta}$ for all $\frac{1}{2} < \alpha < 1$. Suppose that (i) $\alpha \neq \alpha_p$, (ii) $\beta \neq \beta_1$ and $\beta \neq \beta_2$ (iii)$w_D \neq 1-\alpha$ and $w_D \neq V(\alpha)$. Then $G_{\alpha,\beta}$ has a unique sequential equilibrium.

The proof of the proposition and the characterization of the equilibrium strategies as a function of the parameters α, β and w_D are given in the appendix. See the diagrams in Figure 4.2 (in the appendix). The restrictions $\alpha \neq \alpha_P$, etc. of Proposition 3 were made to avoid multiple equilibria.

We direct attention to three special cases. The first two address situations where w_D is small, a Deterrer gets a low payoff from (NB,NA). Propositions 4 and 5 respectively deal with relatively low and high probabilities that 1 is a Provocateur. The third case examines situations where w_D is large. Proposition 6 examines it for all values of β.

Proposition 4: Consider the game $G_{\alpha,\beta}$ for all $\frac{1}{2} < \alpha < 1$, $\alpha \neq \alpha_P$. Suppose that $w_D < \min\{1-\alpha, V_1(\alpha)\}$ and $\beta < \beta_1$. Then in any sequential equilibrium of $G_{\alpha,\beta}$, Player 2 attacks if the signal she receives is b and with positive probability even if she gets signal nb. D builds the bomb with positive probability, while P never builds the bomb .

Proof: See Appendix.

Proposition 4 asserts that for α in the relevant range, if w_D and β are relatively small, 2 acts aggressively. She attacks for sure given signal b, and with positive probability even if nb. The intuition is clear. If 1 refuses to open his facilities, he is either a D who built the bomb (otherwise he would open his facilities), or he is a P. However, 2 assigns a high initial probability that 1 is a D, implying a high probability of a bomb. Signal b reinforces this belief, and 2 attacks for sure.

Signal nb raises some doubts because of the likelihood of 1 being a D, and for sure

D has built the bomb if he did not open up his facilities for inspection. These doubts are stronger for lesser α since it makes it more likely that 1 will have chosen B. The best reply action of 2 to the signal nb is to attack 1 with positive probability that decreases with α.

Meanwhile, P is analyzing the interaction of 2 and a hypothetical D. He sees 2's aggressive stance, and completely refrains from building the bomb. He sits back and welcomes the likely unjustified attack from 2, since he prefers (NB,A) to (NB,NA). This equilibrium only applies because the likelihood of P is below a critical threshold, $\beta < \beta_1$.

Proposition 5 addresses the case where P is much more likely.

Proposition 5: Suppose that $\alpha_P < \alpha < 1$, $w_D < V_2(\alpha)$ and $\beta > \beta_2$. Then in any sequential equilibrium of $G_{\alpha,\beta}$ Player 2 acts cautiously. She attacks with positive probability only if the signal is b. Player 1 of type D builds the bomb with probability 1, and the Provocateur builds the bomb with positive probability, but less than 1.

Proof: See Appendix.

When P is relatively likely, Player 2 is restrained for the same reason that he was in Proposition 2, when the enemy was P for sure. Here, unlike Proposition 2, D can take advantage of 2's restraint. Hence, D builds the bomb for sure.

The next proposition deals with the case where the payoff w_D is sufficiently large. Not surprisingly, the result is similar to part (ii) of Proposition 1.

Proposition 6: Let $\frac{1}{2} < \alpha < 1$, $\alpha \neq \alpha_P$, and suppose $w_D > V(\alpha)$. Then the only equilibrium is a separating one:

(i) D opens up his facilities for inspection and the Provocateur builds the bomb with positive probability. This is the best outcome for Player 2 and the second best outcome for D.

(ii) The expected payoff of P is decreasing in $\alpha \in (\frac{1}{2}, \alpha_P) \cup (\alpha_P, 1]$.

Proposition 6 asserts that if the payoff w_D is relatively high then irrespective of β, D will open up to inspection, thus preventing any chance of attack. Therefore if 1 does not permit inspection he reveals his type P and the analysis of Proposition 2 ($\beta = 1$)

applies. Thus, if $\alpha_P < \alpha < 1$, Player 2 does not attack if the signal is nb and with positive probability does not even attack if the signal is b. However, if $\frac{1}{2} < \alpha < \alpha_P$ and $w_D > V(\alpha)$, then 2 attacks 1 if the signal is b, and with positive probability she attacks if the signal is nb. Finally, if $\alpha_P > \frac{1}{2}$ (namely, $1 > w_P + r_P$) the expected payoff of 2 is increasing in $\alpha \in (\frac{1}{2}, \alpha_P)$ and in $\alpha \in (\alpha_P, 1]$. If $\alpha_P < \frac{1}{2}$ the expected payoff of 2 is increasing in $\alpha \in (\frac{1}{2}, 1]$.

2.2.3 A brief distillation and transition

Using game theory models, we have sorted our way through models of how two players, a first mover who seeks to possess weapons of mass destruction, and a second mover who seeks to deter their creation or failing that to destroy them, will interact. The subtleties in the game arise because the first mover's type is unknown, and the second mover's intelligence about the existence of weapons is imperfect. Throughout, positing that both players were rational, we identified initially surprising results. Upon reflection, those results yielded to intuition.

The goal of our game theory models was to provide insight into real world situations. The real world does not disappoint. It throws up results that replicate many of the phenomena observed in our models, and that are at least as intriguing. Let us return therefore to the unfinished business of assessing Saddam Hussein's type, and compare his behavior in 2003 with that of the Ayatollah Khamenei a decade later.

3 A Tale of Two Tyrants

Two uncertainties lie at the heart of this game-theoretic analysis: (1) whether player 1 possesses weapons of mass destruction, and (2) whether player 1 is a Deterrer or a Provocateur, that is his type as defined by his payoffs. Though not explicitly assumed, the model is particularly germane when player 1 is a closed and secretive regime, and payoffs may be harder to discern. The regimes of Saddam Hussein in Iraq and the Ayatollah Ali Khamenei in Iran, the former vanquished and dead, the latter the current supreme leader,

exemplify player 1's with these characteristics.[3] The Second Gulf War was substantially justified on the premise that Saddam possessed such weapons, a perception that was reinforced when he refused to provide accurate information on his weapons and fully cooperate on opening his facilities even when a coalition invasion was imminent. The current standoff between Western nations and Iran also revolves around the question of complete inspections, in this instance to indicate the status of its nuclear weapons potential. In the current environment, punitive economic sanctions are the primary threat to Iran, though military threats, from Israel as well as from Western nations, are also consequential.

Uncertainty 1 - the status of weapons - certainly applies to both the Saddam and Khamenei cases. Uncertainty 2 - the structure of payoffs to the players are also prominent. Both Saddam and the Ayatollah, it is clear, respectively carried and carry bold ambitions to be leaders among Middle Eastern Muslim nations. Part of their strategies has been to issue numerous threats to Israel, both to appeal to their citizens, and individuals beyond their borders. However, the true intentions and payoffs of these two individuals have always proved hard to read, and Western analysts have differed in their assessments.

At least to the extent that interviews with former regime members and massive numbers of captured Baathist state documents can reveal, we now have a much better understanding of what Saddam's goals, and his beliefs about Western intentions. That information contains many surprises.[4] One illustration: In early 2003, even with U.S. forces massed on his border, he thought that they would not invade, and that if they did, they would suffer a combination of significant losses and international pressures that

[3]This section is drawn entirely from secondary sources. The authors also thank Matthew Bunn and Jeffrey Friedman for helpful discussions on this section. Neither is responsible for any errors or misinterpretations.

[4]It would be less surprising to students of international affairs who are familiar with the pioneering work on the role of misperceptions about beliefs due to Jervis [1976]. The role of misperceptions in conflicts between the United States and Saddam Hussein is explored insightfully by Duelfer and Dyson [2011].

would force them to withdraw. Briefly later, in the midst of a military disaster, he still thought his side was winning. Evidently, he believed his regime's propaganda.[5] Thus, his refusal to open his facilities to inspection, apart from revealing to internal supporters and enemies that he lacked weapons of mass destruction, was strongly influenced by his belief that he would not be overthrown. The potential for such bizarre beliefs makes it hard to draw sound conclusions from an enemy's actions.

There is evidence directly relevant to our model, and the possible types for player 1. Retrospective analyses make it clear that Saddam was a strategist who sought weapons of mass destruction to serve him as a Deterrer, but he also clearly took actions that exemplified his role as a Provocateur. Saddam sought nuclear weapons so that he could pursue his vision of leading Iraq and his Arab allies in the fight against Israel in a war of attrition employing conventional weapons. His nuclear bomb would serve as a deterrent to Israel's use of nuclear weapons, which he would otherwise expect them to employ against Iraq. In January-February 1991, Saddam launched SCUD missiles, ineffectively against Israeli population centers (and also against Saudi Arabia). He expected the attacks would "inspire the Arab masses to rally behind Iraq, thereby forcing their governments to distance themselves from the U.S.-led coalition." (Brands and Palkki [2011, p. 157]). Saddam also thought he could launch these attacks without inviting a devastating response, relying on the deterrence provided by his vast stores of chemical and biological weapons.[6] Israel, with strong pressures from the United States, did not respond to the SCUDs. Had Israel responded significantly, it surely would have brought fierce Arab condemnation, and possibly would have fractured the coalition. Had that happened, Saddam would have succeeded in his role as a Provocateur.

The current-day situation with the Ayatollah Ali Khamenei differs from the 2003 confrontation with Saddam Hussein on weapons concerns. With the Ayatollah it is the potential for nuclear weapons whereas for Saddam it was the existence of biological and chemical weapons. Nevertheless, the two cases bear many similarities with respect to the

[5]This discussion is based on Woods et al. [2006].

[6]See Brands and Palkki [2011] for much further discussion on these issues.

fundamental properties of our model. Khamenei's preferences, his type, as were Saddam's, are virtually impossible for his player 2, the Western allies and Israel, to assess. So too, it is hard to discern the time schedule on which Iran could build nuclear weapons.[7] We know that it is advanced, but not how advanced, and not whether it is currently pushing further advancement. One of the most tantalizing uncertainties about Khamenei is his true view on nuclear weapons. His alleged fatwa against them has been widely publicized by his regime, in particular in international forums. But the fatwa itself has never been produced, and it is not clear whether it exists.[8] What is clear is that the threat of Iran to influence in the region would be greatly enhanced if it had nuclear weapons, and its ability to deter its potential attackers would be substantially enhanced as well.

It is also clear that Iran is suffering severely economically, and recent elections and protests have revealed severe dissatisfactions with the regime, particularly among the elite. Nothing would do more to bolster support for the regime than an attack on facilities that were shown clearly not to be close producing a nuclear weapon. In short, Khamenei would reap significant benefits were Iran attacked unjustifiably. Thus, the preferences of a Provocateur are a clear possibility. However, that hardly means that he would favor

[7]A third significant uncertainty relates to Iran's beliefs about Western and Israeli intentions. Jervis [2014, p. 17] relates the misperceptions problem to the conflict between Iran and Western nations. "... the state [player 2 in our model] needs to understand not only what the other side has done but also why it did so." Jervis's main point about types is that the West does not know whether Iran's intentions are primarily offensive or defensive.

Miller and Bunn [2014] focus on American perceptions of Iran from the beginning of the Islamic Republic. They explain why America has strong evidence to support the belief that Iran is unchangeably hostile and extremist, and also irrational. However, they also point to evidence that at times Iran has put state interest ahead of religious ideology, implying rationality, and that some of its aggressive moves may have stemmed from defensive concerns. The latter view makes more sense if one posits that the Iranian leadership has misleading perceptions of the intentions of the United States and Israel. In short, their presentation of Iran is almost like the famed optical illusion where it is possible to see either an old hag or a beautiful young woman in the same image.

[8]See, for example, Ariane Tabatabai,

http://www.middleeast-armscontrol.com/2013/02/28/dont-misunderstand-khameneis-nuclear-fatwa/

and Mehdi Khalaji, http://www.washingtoninstitute.org/uploads/Documents/pubs/PolicyFocus115.pdf.

such an attack; he may be a Deterrer. If and when documents from this regime come to light, and when officials can be interviewed in retrospect, as happened after the downfall of Saddam Hussein, we may come to better understand the Ayatollah's perceptions of his own payoffs.

In the interim, given uncertainties about Iran's stage of weapon intentions and development, and about the preferences driving the regime, our model is directly relevant, as are its possible uncomfortable conclusions.

4 Summary and Conclusion

There are two ways of denying a first-mover enemy the potential for possessing nuclear weapons: deterring them from producing them, and destroying them if they are produced. Economic sanctions or other penalties may play a role in the deterrence. However, this paper considers deterrence as coming solely from the threat of having the weapons destroyed. We start with three major challenges to the destruction strategy: First, the second mover may not have an assured capability for destroying the weapons. Second, she may not know for sure whether the weapons have been developed, particularly if the first mover is a closed society. Third, attacking the weapons facility, most importantly if the weapons have not yet been built, has the potential to greatly damage the attacker's legitimacy and enhance the position of the enemy with its internal community and the world more generally. Our analysis assumes the first challenge has been surmounted: destruction is possible. It focuses on the second and third challenges, and posits that the second mover has an imperfect intelligence capability to assess whether the weapons have been built.

This paper considers still a fourth challenge: the motives of the potential bomb builder, as illuminated by his payoffs, may be hard to fathom. That was certainly the case with Saddam Hussein and the situation surrounding the Second Gulf War, and exists with Iran and North Korea today. We consider two possible types for the bomb seeker, Deterrer and Provocateur. When the first mover's motives are uncertain, sequences of

moves may appear that seem bizarre, but in fact are part of the equilibrium strategies of rational players. Thus, in real life, in 2003 Saddam Hussein did not fully cooperate with inspections, even though he did not possess the weapons that the United States feared. The US attacked on intelligence that was at best faulty, and might have done so even if its intelligence indicated there were no weapons. Of course, factors beyond our model may have played a role in the Second Gulf War. For example, there is evidence that even at the last moment Saddam did not believe that the US would invade, [9] and that he may have refused inspections because he wanted his own army and the Iranians to believe he had the feared weapons.

Our model focused on the two enemies, but concerns about domestic players and outside audiences may substantially influence the payoffs of those two players, and hence their actions. The real world is much richer than our model or any possible model. Nevertheless, the prime lessons of our model surely apply: You rarely know the motives of enemies that are closed societies. Despite that uncertainty, strategic thinking still yields significant insights. Those insights are more powerful still if they start, as did this analysis, with the recognition that the enemy's payoffs are not known.

Our analysis addressed situations between international enemies. But the information structure, roughly speaking, arises in an array of game-theoretic situations. The critical elements are that neither the choice nor the type of the first mover, 1, is automatically disclosed, though the first mover could and sometimes will credibly reveal his move. The second mover, 2, wishes to deter the hostile (*build*) move by the first mover through his threat of retaliation (*attack*). But she would prefer not to attack if 1 chooses an innocuous move (*not build*). Player 2 also has an imperfect intelligence system to assess 1's move.

As an example, consider a labor-management situation. Management is the first

[9]This suggests that future work should consider situations where both player 1's and player 2's type is unknown to the other. A further generalization would have player i's probabilistic assessment of player j's type be private information. In the Second Gulf War, the United States assumed, incorrectly, that Saddam Hussein knew he was about to be invaded, yet still was unwilling to open his facilities. This reinforced any view that Iraq possessed weapons of mass destruction. For the seminal work on such misperceptions see Jervis [1976].

mover. Labor fears that management could undertake expensive preparations that would enable it to transfer lower-end production to China. That would be an unfriendly move, and the union would like to prevent it. The contract will be coming to an end in the near future, and the union has announced that a job-security provision will be in its demands for the next contract to protect against the China move. The union would like to strike at the contract's expiration if and only if management had made the expensive preparations. The union is following the Chinese business press, expecting that it might learn of any deal made by management to potentially move lower-end production. However, such intelligence is highly imperfect, and subject to type I and type II errors. There is an additional factor. The union does not know whether management is *tough* or *soft*. If management is *tough*, it could *not build*, hope for a strike, and then reveal its move. That would give it legal protection and public relations cover to lay off some surplus workers. If management is *soft*, it would suffer from a strike regardless of its action. Indeed, a *soft* management might even accept a version of the job-security provision at the outset, thereby revealing it had no Chinese-production possibility. This would preserve morale, possibly get a better wage deal, and avoid any potential for a strike. Though the enmity between the parties in this labor-management example is less severe than in the nuclear bomb model, the two situations are directly parallel.

In many real world situations, the payoffs to another player are unknown. That is particularly likely when the other player is a closed and/or secretive regime, as is frequently the case in international affairs. We analyze the equilibrium under the assumption that the prior distribution on that player's type is common knowledge. It defines the strategies the players use in the critical situation where an open nation is seeking to prevent - through deterrence or an attack on facilities - a closed nation from possessing a nuclear weapon.

References

R. Avenhaus, B. Von Stengel, and S. Zamir. Inspection games. *R. J. Aumann and S. Hart (eds), Handbook of game theory with economic applications*, 3:1947–1987, North–Holland, Amsterdam, 2002.

D. Biran and Y. Tauman. The decision to attack a nuclear facility: The role of intelligence. *Unpublished manuscript*, 2009.

H. Brands and D. Palkki. Saddam, Israel, and the bomb: nuclear alarmism justified? *International Security*, 36(1):133–166, 2011.

S. Chassang and G. P. Miquel. Conflict and deterrence under strategic risk. *The Quarterly Journal of Economics*, 125(4):1821–1858, 2010.

M. Dresher. *Some Military Applications of the Theory of Games*. RAND Corporation, 1961.

M. Dresher. Mathematical models of conflicts. *E. S. Quade and W. I. Boucher (eds), Systems Analysis and Policy Planning: Applications in Defence*, pages 228–240, RAND Corporation, 1968.

C. A. Duelfer and S. B. Dyson. Chronic misperception and international conflict: The US-Iraq experience. *International Security*, 36(1):73–100, 2011.

M. V. Finn and G. A. Kent. Simple analytic solutions to complex military problems. Technical report, DTIC Document, 1985.

N. Gennaioli and H. Voth. State capacity and military conflict. *Review of Economic Studies*, Forthcoming.

E. D. Gould and E. F. Klor. Does terrorism work? *The Quarterly Journal of Economics*, 125(4):1459–1510, 2010.

R. Jervis. *Perception and misperception in international politics*. Princeton University Press, 1976.

R. Jervis. The United States and Iran: Perceptions and policy traps. *A. Maleki and J. Tirman (eds), U.S. – Iran Misperceptions*, pages 15 – 36, New York: Bloomsbury Academic, 2014.

M. I. Kamien, Y. Tauman, and S. Zamir. On the value of information in a strategic conflict. *Games and Economic Behavior*, 2(2):129–153, 1990.

A. F. Karr. Nationwide defense against nuclear weapons: Properties of prim-read deployments. Technical report, DTIC Document, 1981.

S.E. Miller and M. Bunn. Interpreting the implacable foe: American perceptions of Iran. *A. Maleki and J. Tirman (eds), U.S. – Iran Misperceptions*, pages 57 – 88, New York: Bloomsbury Academic, 2014.

B. O'Neill. Operations research and strategic nuclear war. *International Military Defense Encyclopedia*, Pergammon-Brassey, 1993.

B. O'Neill. Game theory models of peace and war. *R. J. Aumann and S. Hart (eds), Handbook of game theory with economic applications*, 2:995–1053, North–Holland, Amsterdam, 1994.

W.T. Read Jr. *Tactics and Deployment for Anti-Missile Defense.* Bell Telephone Laboratories, Whippany, NJ, 1958.

W.T. Read Jr. Strategy for active defense. *The American Economic Review*, pages 465–471, 1961.

T. C. Schelling. *The strategy of conflict.* Harvard University Press, 1960.

M. Shubik. Game theory, the language of strategy', 1-28. *M. Shubik (ed.), Mathematics of Conflict*, North–Holland, Amsterdam, 1983.

M. Shubik. *The uses, value and limitations of game theoretic methods in defense analysis.* Defense Technical Information Center, 1985.

C. Thomas. Some past applications of game theory to problems of the United States Air Force. *Proceedings of a conference under the aegis of the NATO Scientific Affairs Committee, Toulon, France*, pages 250–267, 1964.

K. Woods, J. Lacey, and W. Murray. Saddam's delusions: the view from the inside. *Foreign Affairs*, pages 2–26, 2006.

Appendix

Proof of Proposition 1

Consider the game G_β where 2 does not use IS.

Lemma 1A: Whether the type of 1 is a private information or commonly known, in equilibrium (i) 2 strictly mixes her two strategies A and NA (ii) P does not play a pure B.

Proof: (i) If 2 plays a pure A 1's best reply (of any type) is NB. But then 2 is best off deviating to NA. If 2 plays pure NA then 1's best reply is B and 2 is best off deviating to A.

(ii) Suppose P chooses a pure B. If 1 does not open up his facilities for inspection 2 knows that 1 of any type builds the bomb with probability 1. Hence 2's best reply is a pure A, contradiction (i). □

Suppose that D chooses NBO with a probability x_D, $0 \leq x_D \leq 1$, and suppose P chooses NB with probability x_P. D prefers NBO on B (see Figure 2.3) iff

$$w_D \geq q. \tag{1}$$

P prefers NB on B iff

$$r_P q + w_p(1-q) \geq q. \tag{2}$$

Denote by N the event "1 does not open up his facilities for inspection". When N occurs, Player 2 assigns probablity

$$Prob(B|N) = \frac{\beta(1-x_P) + (1-\beta)(1-x_D)}{\beta + (1-\beta)(1-x_D)}, \tag{3}$$

that 1 is building the bomb. If 1 does not open up his facilities for inspection, 2 prefers A on NA iff

$$Prob(B|N)w_2 + (1 - Prob(B|N))r_2 \geq 1 - Prob(B|N)$$

By Lemma 1A $0 < q < 1$ hence

$$Prob(B|N)w_2 + (1 - Prob(B|N))r_2 = 1 - Prob(B|N) \tag{4}$$

By assumption $w_D \neq \frac{w_P}{1-r_P+w_P}$, thus, from (1) and (2) there is no equilibrium where both $0 < x_D < 1$ and $0 < x_P < 1$. This together wih Lemma 1A imply that there are only three possible equilibrium profiles: $(0 < x_D < 1, x_P = 1, 0 < q < 1)$, $(x_D = 1, 0 < x_P \leq 1, 0 < q < 1)$ and $(x_D = 0, 0 < x_P \leq 1, 0 < q < 1)$.

Case 1 $(0 < x_D < 1, x_P = 1, 0 < q < 1)$

In this case D is indifferent between NBO and B, and P strictly prefers NB. By (1) and (2) this implies

$$w_D < \frac{w_P}{1 - r_P + w_P}$$

By (3), in this case

$$Prob(B|N) = \frac{(1-\beta)(1-x_D)}{\beta + (1-\beta)(1-x_D)},$$

and since $0 < q < 1$, 2 is indifferent between NA and A when 1 does not allow inspection. Therefore, by (4)

$$x_D = \frac{(1-\beta)w_2 - \beta(1-r_2)}{(1-\beta)w_2}.$$

Note that $0 < x_D$ iff $\beta < \frac{w_2}{1+w_2-r_2}$.

Case 2$(x_D = 1, 0 < x_P \leq 1, 0 < q < 1)$

By (3), $Prob(B|N) = 1 - x_P$, and from (4)

$$x_P = \frac{w_2}{1 + w_2 - r_2}$$

Since $x_D = 1$ D strictly prefers NBO on B and P is indifferent between B and NB. From (1) and (2)

$$w_D > \frac{w_P}{1 - r_P + w_P}$$

Since $0 < x_p < 1$ (2) holds as equality and

$$q = \frac{w_P}{1 - r_P + w_P}.$$

Case 3($x_D = 0, 0 < x_P \leq 1, 0 < q < 1$)

By (3), $Prob(B|N) = 1 - \beta x_P$, and from (4)

$$x_P = \frac{w_2}{\beta(1 + w_2 - r_2)},$$

and $0 < x_P < 1$ iff $\beta > \frac{w_2}{1+w_2-r_2}$. Hence for $\beta > \frac{w_2}{1+w_2-r_2}$ P is indifferent berween NB and B, and from (2)

$$q = \frac{w_P}{1 - r_P + w_P},$$

By (1) D strictly prefers B on NBO iff

$$w_D < \frac{w_P}{1 - r_P + w_P}$$

□

Denote

$$\beta_1 = \frac{(1-\alpha)w_2}{(1-\alpha)w_2 + \alpha(1 - r_2)}$$

and

$$\beta_2 = \frac{\alpha w_2}{(1-\alpha)(1 - r_2) + \alpha w_2}$$

Note that $\beta_2 > \beta_1$ for $\alpha > \frac{1}{2}$.

Proof of Propositions 3,4,5 and 6

(i) In every sequential equilibrium (se) 2's best reply to NBO is NA. If 1 does not open up his facilities for inspection then 2 chooses A with positive probability (otherwise both NP and P are best off choosing pure B). Since D is better off when 2 chooses NA (whether or not he builds the bomb) he strictly prefers NBO on NB and hence he plays NB with zero probability. Consequently, D chooses either B or NBO or a mixture of the two. Similarly, P who prefers to be attacked when he does not build the bomb prefers NB on NBO and hence he either plays B or NB or a mixture of the two. Let $(x_D, x_P, y(A|b), y(A|nb))$ be a se profile where x_D is the probability that D chooses NBO and x_P is the probability that P chooses NB (with probability $1 - x_D$ and $1 - x_P$, D and P respectively builds

the bomb). Similarly $y(A|t)$ is the probability that 2 attacks 1 if she receives the signal $t \in \{b, nb\}$. Figure 4.1 describes the sequence of events and the possible outcomes of the game.

D weakly prefers B on NBO (we write $B \succeq_D NBO$) iff (see Figure 4.1)

$$\alpha[y(NA|b) \cdot 1 + y(A|b) \cdot 0] + (1-\alpha)[y(NA|nb) \cdot 1 + y(A|nb) \cdot 0] \geq w_D$$

Equivalently,

$$\alpha y(NA|b) + (1-\alpha) y(NA|nb) \geq w_D. \tag{5}$$

Inequality must hold if $0 < x_D < 1$. Similarly, $B \succeq_P NB$ iff

$$\alpha y(NA|b)+(1-\alpha)y(NA|nb) \geq \alpha[y(NA|nb)r_P+(1-y(NA|nb))w_P]+(1-\alpha)[y(NA|b)r_P+(1-y(NA|b)w_P]. \tag{6}$$

In addition (see Figure 4.1) we have

$$Prob_2(NB|b) = \frac{(1-\alpha)\beta x_P}{(1-\alpha)\beta x_P + \alpha[(1-\beta)(1-x_D) + \beta(1-x_P))]}, \tag{7}$$

and

$$Prob_2(NB|nb) = \frac{\alpha\beta x_P}{\alpha\beta x_P + (1-\alpha)[(1-\beta)(1-x_D) + \beta(1-x_P))]}, \tag{8}$$

where $Prob_2(NB|t)$ is the probability that 2 assigns to the event that 1 chose NB if she receives the signal $t \in \{b, nb\}$. Denote $A \succeq_t NA$ the case where 2 weakly prefers A on NA when observing the signal t. It is easy to verify (see Figure 4.1) that

$$A \succeq_t NA \text{ iff } Prob_2(NB|t)r_2 + Prob_2(B|t)w_2 \geq Prob_2(NB|t).$$

Equivalently

$$A \succeq_t NA \text{ iff } Prob_2(NB|t) \leq \frac{w_2}{1 - r_2 + w_2},\ t \in \{b, nb\} \tag{9}$$

and equality holds if $0 < y(A|t) < 1$.

Lemma 2A There is no se where $0 < y(A|t) < 1$ for both $t = b$ and $t = nb$.

Proof: Suppose to the contrary that $0 < y(A|t) < 1$ for both $t = b$ and $t = nb$. Then (9) holds as equality for both b and nb, implying that $Prob(NB|b) = Prob(NB|nb)$. By (7) and (8) we have

$$\frac{(1-\alpha)\beta x_P}{(1-\alpha)\beta x_P + \alpha[(1-\beta)(1-x_D) + \beta(1-x_P))]} = \frac{\alpha\beta x_P}{\alpha\beta x_P + (1-\alpha)[(1-\beta)(1-x_D) + \beta(1-x_P))]},$$

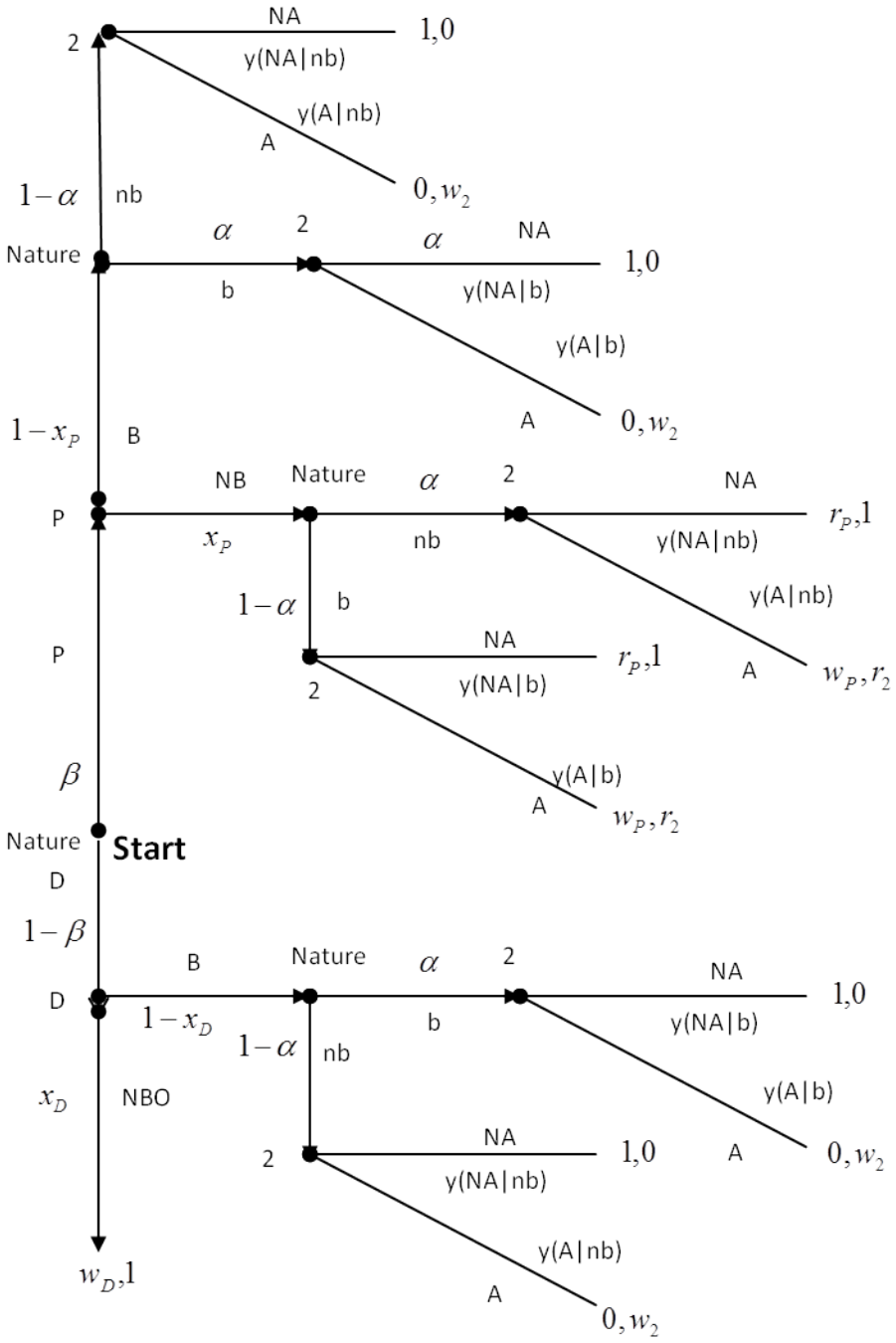

Figure 4.1: The reduced game of $G_{\alpha,\beta}$

but the last equality holds only for $\alpha = \frac{1}{2}$, contradicting our assumption $\alpha > \frac{1}{2}$. □

Lemma 3A In every se $y(NA|b) > 0$ or $y(A|nb) < 1$.

Proof: Suppose to the contrary that $y(A|b) = 0$. If in addition $y(A|nb) = 0$ then 2 does not attack 1 irrespectively of the signal she observes. In this case B is the best reply of both D and P. But then 2 is better off deviating to A, a contradiction. Thus $y(A|nb) > 0$ must hold. This implies $A \succeq_{nb} NA$. By (8) and (9) we have:

$$\frac{\alpha\beta x_P}{\alpha\beta x_P + (1-\alpha)[(1-\beta)(1-x_D) + \beta(1-x_P))]} \leq \frac{w_2}{1 - r_2 + w_2} \tag{10}$$

Next, since $y(A|b) = 0$ $NA \succeq_b A$ and by (7) and (9)

$$\frac{(1-\alpha)\beta x_P}{(1-\alpha)\beta x_P + \alpha[(1-\beta)(1-x_D) + \beta(1-x_P)]} \geq \frac{w_2}{1 - r_2 + w_2}. \tag{11}$$

By (10) and (11)

$$\frac{(1-\alpha)\beta x_P}{(1-\alpha)\beta x_P + \alpha[(1-\beta)(1-x_D) + \beta(1-x_P)]} \geq \frac{\alpha\beta x_P}{\alpha\beta x_P + (1-\alpha)[(1-\beta)(1-x_D) + \beta(1-x_P))]}.$$

It is easy to verify that the last inequality holds iff $\alpha \leq \frac{1}{2}$, a contradiction.

Similarly it can be shown that $y(A|nb) < 1$. □

Corollary 1A In every se either $y(A|b) = 1$ or $y(A|nb) = 0$.

Proof: Follows directly from Lemmas 2A and 3A.□

Lemma 4A In every se $x_P > 0$

Proof: Suppose to the contrary that $x_p = 0$. Namely, P builds the bomb with certainty. If D does not permit inspection he too for sure builds the bomb (since NB for D is dominated by NBO). Hence, in equilibrium if 1 does not permit inspection 2 knows that irrespective of his type 1 builds the bomb with certainty. Her best reply strategy in this case is a pure A irrespective of the signal of IS. But then P is best off deviating to NB. □

Lemma 5A: Suppose $x_D = 1$. Then $0 < x_P < 1$.

Proof: Suppose to the contrary that P plays a pure strategy. By Lemma 4A $x_D = x_P = 1$. Thus 2's best reply strategy is a pure NA irrespective of the signal received. But then both D and P are better off deviating to b. □

To avoid multiple equilibria for some specific values of α,β and w_D we conveniently assume

Assumption: $\alpha \neq \alpha_P$, $\beta \notin \{\beta_1, \beta_2\}$, $w_D \notin \{1-\alpha, V_1(\alpha), V_2(\alpha)\}$.

Lemma6A The following profiles are not se profiles

(i) $(0 < x_D < 1, 0 < x_P \leq 1, y(A|b) = 1, y(A|nb) = 0)$

(ii) $(0 < x_D < 1, 0 < x_P < 1, y(A|b) = 1, y(A|nb) > 0)$

(iii) $(0 < x_D < 1, 0 < x_P < 1, y(A|b) < 1, y(A|nb) = 0)$

(iv) $(0 \leq x_D \leq 1, 0 < x_P < 1, y(A|b) = 1, y(A|nb) = 0)$

(v) $(x_D = 0, x_P = 1, y(A|nb) = 1, 0 < y(A|nb) < 1)$

(vi) $(x_D = 0, x_P = 1, y(A|nb) < 1, y(A|nb) = 0)$

Proof: Consider profile (i). Substituting $y(A|b) = 1$ and $y(A|nb) = 0$ in (5) and since D is indifferent between B and NBO, we have $w_D = 1 - \alpha$, a contradiction to our assumption.

Consider next profiles (ii) and (iii). D is indifferent between B and NBO, and P is indifferent between B and NB. Substituting $y(A|b) = 1$ or $y(A|nb) = 0$ in (5) and (6) (as equalities), our assumptions $w_D \neq V_1(\alpha)$ and $w_D \neq V_2(\alpha)$ are violated. Consider now profile (iv). P is indifferent between B and NB. Substitution $y(A|b) = 1$ and $y(A|nb) = 0$ in (6) implies $\alpha = \alpha_P$, a contradiction.

Consider next profile (v). Then 2, when receiving the signal nb, is indifferent between A and NA. Hence (9) holds as equality. Since $x_D = 0$ and $x_P = 1$, (8) and (9) imply $\beta = \beta_1$, contradicting the assumption $\beta \neq \beta_1$. Similarly, (7) implies the impossibility (vi) to be a se profile.

□

Corollary 2A: The following seven profiles are the only candidates for se profiles

(1) $(x_D = 1, 0 < x_P < 1, y(A|b) = 1, 0 < y(A|nb) < 1)$

(2) $(x_D = 0, 0 < x_P < 1, y(A|b) = 1, 0 < y(A|nb) < 1)$

(3) $(0 < x_D < 1, x_P = 1, y(A|b) = 1, 0 < y(A|nb) < 1)$

(4) $(0 < x_D < 1, x_P = 1, 0 < y(A|b) < 1, y(A|nb) = 0)$

(5) $(x_D = 0, 0 < x_P < 1, 0 < y(A|b) < 1, y(A|nb) = 0)$

(6) $(x_D = 1, 0 < x_P < 1, 0 < y(A|b) < 1, y(A|nb) = 0)$

(7) $(x_D = 0, x_P = 1, y(A|b) = 1, y(A|nb) = 0)$

Proof: Follows directly from Corollary 1A and from Lemmas 4A,5A and 6A.

□

Case 1 $(x_D = 1, 0 < x_P < 1, y(A|b) = 1, 0 < y(A|nb) < 1)$

By (6)

$$y(A|nb) = \frac{1 - \alpha(1 - w_P + r_P) - w_P}{1 - \alpha(1 - w_P + r_P)},$$

$y(A|nb) > 0$ implies $\alpha < \alpha_P$. (5) impiles $w_D > V_1(\alpha)$. By (9)

$$x_P(\alpha) = \frac{(1-\alpha)w_2}{\alpha(1-r_2) + (1-\alpha)w_2}.$$

Corollary 2.1A: Suppose $w_D > V_1(\alpha)$ and $\alpha < \alpha_P$. Then $V_1(\alpha) = V(\alpha)$ and

$$(x_D = 1, x_P = \frac{(1-\alpha)w_2}{\alpha(1-r_2) + (1-\alpha)w_2}, y(A|b) = 1, y(A|nb) = \frac{1 - \alpha(1 - w_P + r_P) - w_P}{1 - \alpha(1 - w_P + r_P)})$$

is a se profile.

Case 2 $(x_D = 0, 0 < x_P < 1, y(A|b) = 1, 0 < y(A|nb) < 1)$

By (6)

$$y(A|nb) = \frac{1 - w_P - \alpha(1 + r_P - w_P)}{1 - \alpha + \alpha(w_P - r_P)}$$

and it implies $\alpha < \alpha_P$. By (5) $w_D < V_1(\alpha)$. By (9)

$$x_P = \frac{(1-\alpha)w_2}{\beta[(1-\alpha)w_2 + \alpha(1-r_2)]},$$

in particular $\beta > \beta_1$, since $x_P < 1$.

Corollary 2.2A: Suppose $w_D < V_1(\alpha)$, $\beta > \beta_1$ and $\alpha < \alpha_P$. Then $V_1(\alpha) = V(\alpha)$ and

$$(x_D = 0, x_P = \frac{(1-\alpha)w_2}{\beta[(1-\alpha)w_2 + \alpha(1-r_2)]}, y(A|b) = 1, y(A|nb) = \frac{1 - w_P - \alpha(1 + r_P - w_P)}{1 - \alpha(1 + r_P - w_P)})$$

is a se profile.

Case 3 $(0 < x_D < 1, x_P = 1, y(A|b) = 1, 0 < y(A|nb) < 1)$

From (5)

$$y(A|nb) = 1 - \frac{w_D}{1-\alpha},$$

which implies

$$\alpha < 1 - w_D.$$

By (6) $w_D < V_1(\alpha)$ and hence it is assumed that $w_D < 1 - \alpha$. By (9)

$$x_D = \frac{w_2(1-\alpha)(1-\beta) - \alpha\beta(1-r_2)}{w_2(1-\alpha)(1-\beta)}$$

which implies $\beta < \beta_1$.

Corollary 2.3A: Suppose $w_D < V_1(\alpha)$ and $\beta < \beta_1$. Then

$$(x_D = \frac{w_2(1-\alpha)(1-\beta) - \alpha\beta(1-r_2)}{w_2(1-\alpha)(1-\beta)}, x_P = 1, y(A|b) = 1, y(A|nb) = 1 - \frac{w_D}{1-\alpha})$$

is a se profile.

Case 4 $(0 < x_D < 1, x_P = 1, 0 < y(A|b) < 1, y(A|nb) = 0)$

From (5)

$$y(A|b) = \frac{1 - w_D}{\alpha},$$

which implies $w_D > 1 - \alpha$. From (6),

$$w_D < V_2(\alpha)$$

Note, that $w_D < V_2(\alpha)$ and $w_D > 1 - \alpha$ is possible only if $\alpha > \alpha_P$. From (9),

$$x_D = \frac{\alpha(1-\beta)w_2 - \beta(1-\alpha)(1-r_2)}{\alpha(1-\beta)w_2}$$

which implies $\beta < \beta_2$.

Corollary 2.4A: Suppose $1-\alpha < w_D < V_2(\alpha)$, $\alpha > \alpha_P$ and $\beta < \beta_2$. Then $V_2(\alpha) = V(\alpha)$ and

$$(x_D = \frac{\alpha(1-\beta)w_2 - \beta(1-\alpha)(1-r_2)}{\alpha(1-\beta)w_2}, x_P = 1, y(A|b) = \frac{1-w_D}{\alpha}, y(A|nb) = 0)$$

is a se profile.

Case 5 $(x_D = 0, 0 < x_P < 1, 0 < y(A|b) < 1, y(A|nb) = 0)$

From (6)

$$y(A|b) = \frac{1 - r_P}{w_P - r_P + \alpha(1 + r_P - w_P)}$$

which implies $\alpha > \alpha_P$.

From (5), $w_D < V_2(\alpha)$.

By (9)

$$x_P = \frac{\alpha w_2}{\beta[\alpha w_2 + (1-\alpha)(1-r_2)]}$$

which implies $\beta > \beta_2$.

Corollary 2.5A: Suppose $w_D < V_2(\alpha)$, $\alpha > \alpha_P$ and $\beta > \beta_2$. Then $V_2(\alpha) = V(\alpha)$ and

$$(x_D = 0, x_P = \frac{\alpha w_2}{\beta[\alpha w_2 + (1-\alpha)(1-r_2)]}, y(A|b) = \frac{1-r_P}{w_P - r_P + \alpha(1+r_P-w_P)}, y(A|nb) = 0)$$

is a se profile.

Case 6 $(x_D = 1, 0 < x_P < 1, 0 < y(A|b) < 1, y(A|nb) = 0)$

From (6)

$$y(A|b) = \frac{1-r_P}{\alpha(1+r_P-w_P) + w_P - r_P}. \tag{12}$$

which implies $\alpha > \alpha_P$.

From (5)

$$\frac{1-w_D}{\alpha} < y(A|b),$$

which implies $w_D > 1-\alpha$. This together with (12) requires

$$\frac{1-w_D}{\alpha} < \frac{1-r_P}{\alpha(1+r_P-w_P) + w_P - r_P}.$$

The last inequation holds for $w_D > V_2(\alpha)$. Note, that for $\alpha > \alpha_P$, $V_2(\alpha) > 1-\alpha$.

By (9)

$$x_P = \frac{\alpha w_2}{\alpha w_2 + (1-\alpha)(1-r_2)}$$

and $0 < x_P < 1$.

Corollary 2.6A: Suppose $w_D > V_2(\alpha)$ and $\alpha > \alpha_P$. Then $V_2(\alpha) = V(\alpha)$ and

$$(x_D = 1, x_P = \frac{\alpha w_2}{\alpha w_2 + (1-\alpha)(1-r_2)}, y(A|b) = \frac{1-r_P}{\alpha(1+r_P-w_P) + w_P - r_P}, y(A|nb) = 0)$$

is a se profile.

Case 7 $(x_D = 0, x_P = 1, y(A|b) = 1, y(A|nb) = 0)$

In this case, from (5) $w_D < 1 - \alpha$ is required.

From (9),

$$\beta_1 < \beta < \beta_2$$

Corollary 2.7A: Suppose $w_D < 1 - \alpha$, $\beta_1 < \beta < \beta_2$. Then

$$(x_D = 0, x_P = 1, y(A|b) = 1, y(A|nb) = 0)$$

is a se profile.

Figure 4.2 describes the regions of the seven sequential equilibrium profiles.

□

Proof of Proposition 2

We analyze the game $G_{\alpha,\beta}$ with $\beta = 1$.

The extreme is $\alpha = 1$. In this case Player 1's action is completely detected and Player 2's action is based on the action taken by Player 1. In the (unique) Nash equilibrium of this game Player 1 does not build the bomb (NB) and Player 2 chooses a pure NA. This is the best outcome for 2 and the third best outcome for P.

Suppose next that $\frac{1}{2} < \alpha < 1$. Observe that in any sgpe P does not play a pure strategy. If he chooses B with probability 1, Player 2 attacks him with probability 1 irrespective of the signal, but then P is better off deviating to NB. If P plays NB, then 2 does not attack him irrespective of the signal. But then P can improve upon by deviating to B. This observation together with Corollary 1, imply that it is sufficient to consider only two sgpe profiles: $(x_D, x_P, y(A|b) = 1, y(A|nb) < 1)$ and $(x_D, x_P, 0 < y(A|b), y(A|nb) = 0)$, where $0 < x_P < 1$.

Recall that

$$\alpha_P = \frac{1 - w_P}{1 - w_P + r_P}$$

Case 1 The strategy profile is $(x_D, x_P, 0 < y(A|b), y(A|nb) = 0)$ and $0 < x_P < 1$. Since $0 < x_P < 1$, P is indifferent between B and NB. Therefore (6) holds as an equality and

$$\alpha(1 - y(A|b)) + (1 - \alpha) = \alpha r_P + (1 - \alpha)[(1 - y(A|b))r_P + y(A|b)w_P], \qquad (13)$$

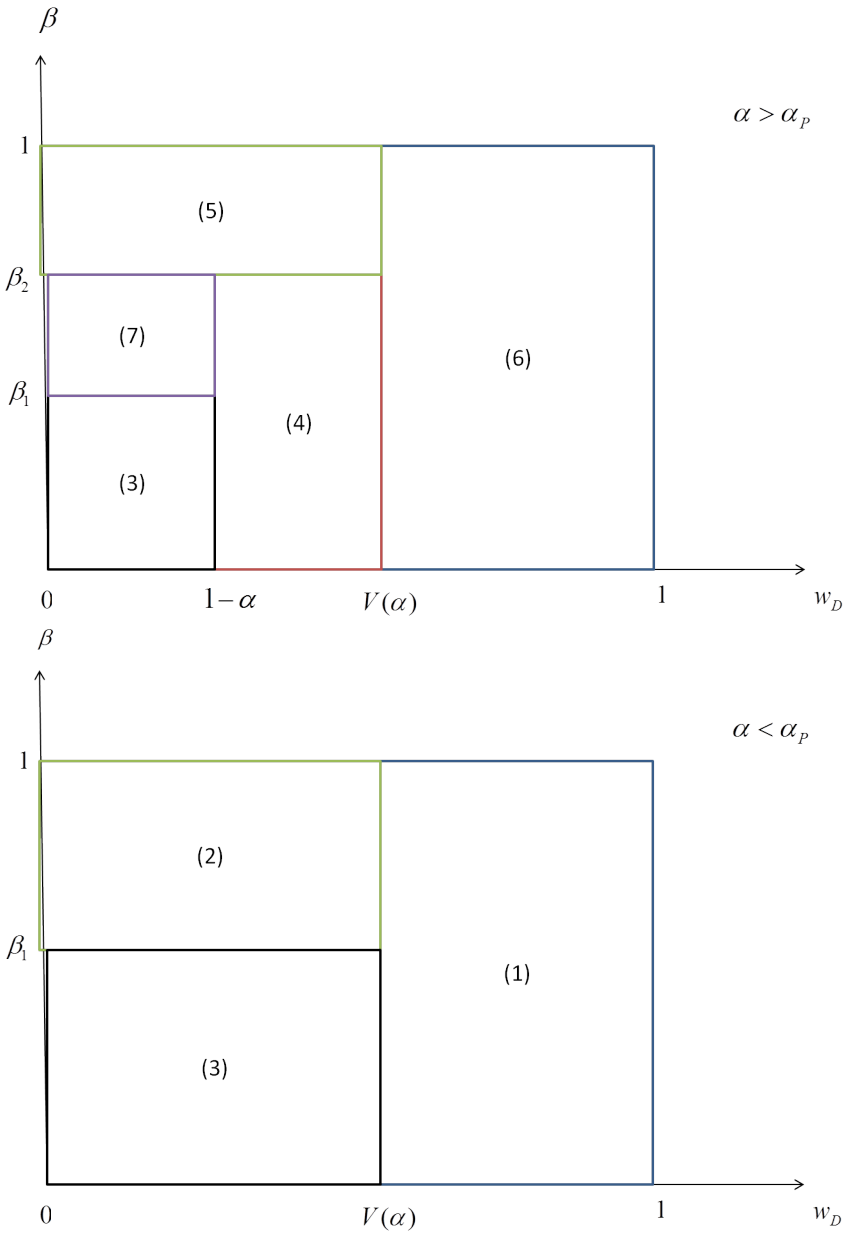

Figure 4.2: Areas of equilibrium outcomes

Its solution is

$$y^*(A|b) = \frac{1 - r_P}{\alpha(1 + r_P - w_P) + w_P - r_P}. \tag{14}$$

By (14) $y^*(A|b) > 0$ and $y^*(A|b) \leq 1$ iff $\alpha \geq \alpha_P$. Let $\alpha > \alpha_P$, then $0 < y^*(A|b) < 1$. Therefore, (9) holds as an equality for signal b. Namely, by (7) (for $\beta = 1$) and (9)

$$\frac{(1-\alpha)x_P}{(1-\alpha)x_P + \alpha(1-x_P)} = \frac{w_2}{1 - r_2 + w_2}.$$

Solving for x_P we obtain

$$x_P^*(\alpha) = \frac{\alpha w_2}{\alpha w_2 + (1-\alpha)(1-r_2)}$$

The expected payoff of P is

$$\Pi_P^*(\alpha) = \frac{(2\alpha - 1)r_P + (1-\alpha)w_P}{\alpha(1 - w_P + r_P) - (r_P - w_P)}.$$

and it decreases in α. If $\alpha = \alpha_P$ then $y^*(A|b) = 1$. Since P is indifferent between B and NB and since by (13) in both cases he obtains $1 - \alpha_P$, any x_P can be supported in equilibrium as long as the inequality

$$\frac{\alpha x_P}{\alpha x_P + (1-\alpha)(1-x_P)} \geq \frac{w_2}{1 - r_2 + w_2} \tag{15}$$

holds. The inequality (15) is derived by (8), (9) and $y(A|nb) = 0$.

Equivalently,

$$x_P \geq \frac{(1-\alpha)w_2}{\alpha(1-r_2) + (1-\alpha)w_2}.$$

Case 2 The strategy profile is $(x_D, x_P, y(A|b) = 1, y(A|nb) < 1)$ and $0 < x_P < 1$.

Again, by (6)

$$(1-\alpha)(1 - y(A|nb)) = \alpha[(1 - y(A|nb))r_P + y(A|nb)w_P] + (1-\alpha)w_P$$

and its solution is

$$\hat{y}(A|nb) = \frac{1 - \alpha(1 - w_P + r_P) - w_P}{1 - \alpha(1 - w_P + r_P)}.$$

Notice that $0 < \hat{y}(A|nb)$ iff $\alpha \leq \alpha_P$. Thus this case is relevant only if $\alpha_P > \frac{1}{2}$. If $\alpha < \alpha_P$, $0 < \hat{y}(A|nb) < 1$, and by (8) and (9)

$$\hat{x}_P(\alpha) = \frac{(1-\alpha)w_2}{\alpha(1-r_2) + (1-\alpha)w_2}.$$

The payoff of P is

$$\hat{\Pi}_P(\alpha)(1-\alpha)\hat{y}(A|nb) = \frac{(1-\alpha)w_P}{1-\alpha(1+r_P-w_P)},$$

and it is decreasing in α. If $\alpha = \alpha_P$, $y^*(A|nb) = 0$. Similar to Case 1, P can choose then any x_P, if it satisfies

$$x_P \leq \frac{\alpha w_2}{\alpha w_2 + (1-\alpha)(1-r_2)}. \tag{16}$$

Inequality (16) is derived by (7), (9) and $y(A|b) = 1$. For $\frac{1}{2} < \alpha < \alpha_P$, $Prob(B) = 1 - \hat{x}_P(\alpha)$, and it is increasing in α. For $\alpha_P < \alpha < 1$, $Prob(B) = 1 - x_P^*(\alpha)$, and it is decreasing in α. Finally, it can be verified that if $\alpha < \alpha_P$, the payoff of 2 is

$$\hat{\Pi}_2(\alpha) = 1 - \beta + \beta[\hat{x}_P(\alpha)r_2 + (1-\hat{x}_P(\alpha))w_2]$$

and it is increasing in α. If $\alpha > \alpha_P$, the payoff of 2 is

$$\Pi_2^*(\alpha) = 1 - \beta + \beta x_P^*(\alpha),$$

and it is increasing in α.

□