HARVARD Kennedy School
JOHN F. KENNEDY SCHOOL OF GOVERNMENT

Confronting an Enemy with Unknown Preferences: Deterrer or Provocateur?

Faculty Research Working Paper Series

Artyom Jelnov
Ariel University

Yair Tauman
Stony Brook University

Richard J. Zeckhauser
Harvard Kennedy School

January 2018

RWP18-006

Confronting an Enemy with Unknown Preferences: Deterrer or Provocateur?

Artyom Jelnov [*] Yair Tauman [†] Richard Zeckhauser [‡§]

January 26, 2018

Abstract

Nation 1 is seeking to join the nuclear club. Nation 2, its enemy, would like to prevent this, and has the potential to destroy 1's bomb-making facilities. It is uncertain whether 1 has a bomb. So are its intentions. 1 could be seeking to deter an attack. Alternatively, if no bomb is present, 1 might wish to provoke one as a means to secure support at home and abroad. Lacking a bomb, 1 can avoid an attack by allowing inspections. If it refuses inspections, 2 must rely on its imperfect intelligence system to determine whether to attack. This game has a unique sequential equilibrium, possibly separating, possibly pooling. At that equilibrium there is a positive probability that: No bomb is built; 2's intelligence system accurately detects no bomb; 1 refuses inspections; nevertheless 2 attacks. Present and past experiences form Iraq, Iran, Syria and North Korea illustrate the analysis.

Forthcoming in the European Journal of Political Economy.

1 Introduction

Kim Jung Un launched an intermediate-range missile into the sea on February 12, 2017, the day after President Trump met with Prime Minister Abe of Japan. Prime Minister Abe and

[*] Ariel University, Israel, Israel, artyomj@ariel.ac.il. Significant part of this research was performed during the first author's stay at the Federmann Center for the Study of Rationality, Hebrew University of Jerusalem, Israel, and was supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° [249159] and by the Lady Davis fellowship.

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

[‡] Harvard University, USA,richard_zeckhauser@hks.harvard.edu

[§] The authors thank the Editor, two referees, and the participants at the Political Economy of Public Policy conference at Ariel University for helpful comments.

Jens Stottenberg, Secretary General of NATO, and other world leaders quickly condemned the test, essentially framing it as a provocation. As Stottenberg stated, North Korea "must refrain from further provocations." Such provocations have hardly stopped, despite warnings and implorations from many nations, including China, its predominant trading partner and only consequential friend. By summer 2017, North Korea had carried out further weapons tests, including intermediate range and intercontinental ballistic missiles, and a missile test from a submarine. Indeed, Kim threatened a nuclear strike at the heart of the United States. Why would the Supreme Leader of North Korea engage in such activities when he knows that his nation could be subject to devastating attack by the United States, possibly in conjunction with a nearby Asian ally, with regime change almost inevitable? One possible answer is that he thinks it highly unlikely that there would be an attack. Before Iran reached agreement with the six nations comprising P5+1 to curtail its nuclear weapons development, it had risked just such an attack, but that attack never happened. But confidence against attack would seem risky with an often threatening and difficult-to-predict President Trump. A second possible answer is that he believes that although an attack is possible, it would be designed as a pinpoint attack on his weapons facilities, that collateral damage would be kept to a minimum, and that his nation and regime would survive, although his military capabilities would be severely impaired. This second answer seems plausible. Regime change has turned out extremely poorly for the United States in Iraq, where the announced justification was wiping out weapons of mass destruction.

Kim Jung Un may have concluded that as opposed to stopping all missile and bomb testing, i.e., knuckling under, i.e., giving up his missiles and bombs, was inferior to risking a surgical strike on his weapons facilities. The North Korean economy is in a shambles, and significant food shortages are reported. Presumably these happenings have weakened the regime. An attack on his weapons facilities by his hated enemy might lead his populace to rally to his support. No doubt somewhat equivalent reasoning led the Iranian regime led by Ayatollah Ali Khameini not to knuckle under and thereby risk attack while it bought time for weapons development and a more favorable deal.

This paper studies a situation where a nation seeking weapons of mass destruction (WMDs), nation 1, is vulnerable to attack by a much stronger enemy, nation 2. The state of 1's weapons development is unclear, as it is for North Korea. No one outside that nation knows about the true state of its missile capabilities, or its ability to mount a nuclear weapon on an operational missile. Its July 2017 test of an ICBM was initially judged a success. Some sources ultimately deemed it a failure since fortuitous video footage from a Japanese TV station showed its reentry vehicle splintering into pieces. Similarly, no one knew where Iran stood in developing its nuclear weapons in the standoff before the treaty was signed.

The analysis posits that any attack would be targeted on weapons facilities, not on a nation or a regime. A war weary United States would not be seeking the extensive involvement that could bring about regime change. Moreover, the aftermaths of recent experiences in Iraq, Libya and Syria has dimmed the enthusiasm of even hawkish officials for deposing autocratic regimes.

This paper analyzes the interaction between two enemy nations. Henceforth we use the terminology *Player* rather than *nation.* 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. Player 1 is treated as male, and Player 2 as female.

We shall use the term "bomb" in our formulation to represent any WMD, or the combination of a WMD and a delivery system. The analysis could apply equally well if the critical uncertainty is not the possession of the weapon, but nation 1's intent to use it. Threats by various nations, such as Iran before its treaty and North Korea today, have become so widely employed that it is impossible for outsiders to discern such nations' true intentions. Here too, to gather strong support if attacked, nation 1 would have to be able to demonstrate that it never really intended to use its weapons, despite its past words.

Player 1 chooses whether or not to build the bomb. Player 2 chooses whether or not to attack. Hereafter, for expository ease, we will often refer to Player 1 as 1, and Player 2 as 2. 1 also has the capability to open his facility to reveal that he does not build the bomb, thereby avoiding any potential for an attack by 2. In determining whether or not to attack, 2 would like to assess whether 1 actively builds the bomb. To do so, she 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, 2 will decide whether or not (or with what probability) to attack. In most important respects, the set up this far parallels that of Jelnov, Tauman, and Zeckhauser (2017), hereafter JTZ.

We make a major departure from JTZ by introducing and focusing on a second critical uncertainty, the preferences, i.e., the type of Player 1. 1 may be a Deterrer, D, or a Provocateur, P, and this is his private information. The critical difference between these two types is that D's primary goal, whether or not he has built the bomb, is to avoid an attack. P, by contrast, prefers to be attacked when he did not build Such an attack would bring support to 1 and blame to 2, since it would be perceived both domestically and abroad to be unjustified. 1 cannot directly reveal his type, even if he would like to. However, he can open his facility to reveal his innocence.

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.

Provocateurs may welcome attacks for several reasons, illustrated in the following two examples. ISIS is known to have and/or be in pursuit of both chemical and nuclear weapons. It could exploit a misguided attack on facilities for such weapons – particularly, as seems inevitable, if that attack killed a number of civilians – as a device to recruit fighters and money. Had the United States attacked the alleged Iranian nuclear facilities prior to the P5+1-Iran agreement, and had no smoking gun for a bomb been found, it would have been widely criticized. Moreover, the Iranian regime would have benefited for at least two reasons: 1. World-wide support for lifting sanctions would have increased. 2. Iran would have been freer to pursue nuclear weapons in the future, since the U.S. would be much more reluctant to attack. In short, there are good reasons to be a Provocateur.

We first address the case where D's payoff from the outcome (NB,NA) is relatively high. He will then choose NBO to avoid an attack. Since P, who prefers the outcome (NB,A) to the outcome (NB,NA), will never open his facility, a separating equilibrium results, where D opens and P does not. Thus, a failure to open reveals to 2 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 in a separating equilibrium knows 1's type, several results from this case are surprising, at least without deep 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 her 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.

[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.

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 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.

We next examine the situation when D gets a relatively low payoff from (NB,NA). Then a pooling equilibrium may result. 2 with positive probability will not 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 Player 1 is D; 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 ex-post 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, 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.

1.1 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 (2003) between Iraq and the United States is perhaps the most salient real world case that both has these characteristics and that has played out relatively 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 a rational outcome, 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, is consistent with the possibility that Saddam was a Provocateur.[2]

A driving uncertainty in our model 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 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 in control in Iraq until that nation was attacked in 2003. Outsiders did not know what he was thinking. In a situation that roughly paralleled our model, prior to its agreement with the P5+1, Iran was a player 1. The Ayatollah Khameini, and perhaps a small circle around him, determined that nation's type. The Ayatollah had continued to make conflicting comments about Iran's intentions on nuclear weapons, about reaching agreement, on inspections, etc. His true preferences were 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

[2] The situation with ISIS bears some similarities. ISIS has launched chemical weapon attacks in Iraq. It also has made claims that it possesses a dirty bomb nuclear weapon, and has the financial resources and is seeking to purchase a traditional nuclear bomb.

audiences.

North Korea is the most inscrutable country in today's world that has or is near to having nuclear weapons and a weapons delivery capability. Its leader frequently proudly proclaims its weapons accomplishments, and Kim Jung Un often issues threats against South Korea, Japan and/or the United States. Kim promotes a powerful cult of personality at home, and it is clear that his declarations have as one obvious purpose shoring up support in a nation with a miserable economy. Whether he is seeking to deter an attack on his weapons, or whether he would welcome such an attack as a means to mobilize support at home, and possibly from China, his insecure ally, is unclear.

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.

1.2 Related literature

Our paper is closely related to JTZ, to Baliga and Sjöström (2008) and to Debs and Monteiro (2014). The sequence of events in this paper is similar to that of JTZ. The major innovation in this paper is that Player 1's type is not fixed. Rather, he has two alternate sets of preferences. Which set applies is not known to Player 2. This feature produces substantially different results. In JTZ, 1 always prefers to avoid an attack whether or not he builds the bomb. In the current paper this is true if and only if Player 1 is a Deterrer. As a result when the IS is relatively accurate Player 2 in JTZ acts cautiously. If she observes the signal nb she for sure does not attack 1, but even if she observes the signal b she still does not attack 1 with positive probability. In contrast, in this paper 2 may act aggressively regardless of IS precision if with positive probability Player 1 is a Provocateur. Namely, even if the IS is almost perfect, it is possible that 1 does not build the bomb, IS detects it correctly and sends the signal nb, yet 2 still mistakenly attacks 1 with positive probability. The other two papers also analyze a situation between a weak nation that possibly has a WMD and a strong nation that has the potential to attack and destroy it. Asymmetric information about whether the weak nation possesses such weapons is at the center of all three analyses. Each paper also features the potential for a mistaken attack, and illustrates using the 2003 invasion of Saddam Hussein's Iraq based on faulty intelligence.

The elegant model of Baliga and Sjöström, hereafter B&S, has incomplete information on both sides as to player types. Their weak nation can be either crazy, e.g., would give weapons of mass destruction (WMDs) to terrorists, or normal. Weak nations also differ in their expected costs of building a working WMD. The strong nation can be a peaceful dove (never attack), aggressive hawk (always attack), or merely an opportunistic type who will be deterred if she thinks the weak nation is normal and has the bomb. By refusing to open its weapons facilities, the weak nation can maintain strategic ambiguity. Thereby, it can still deter and avoid trying to build a WMD when the prospects for success are not high. B&S also provide an insightful analysis of the possible roles for direct communication between the players.

Our model is not as rich as B&S; nevertheless equivalent subtleties emerge. Beyond the weak nation's WMD possession, our model's only information asymmetry is on the type of the weak nation. One type, D, would prefer not to be attacked if it does not possess the bomb whereas the other type, P, prefers to be attacked in this circumstance. In our analysis, the strong nation wants to attack if the weak nation has a WMD, but not if it doesn't. The strong nation has an imperfect intelligence system to help it make that decision. The circumstances that favor attack in our model stand in strong contrast to those in B&S. Their opportunistic type the only type for whom possession matters wants to attack iff the weak nation doesn't have a WMD . Their model focuses on deterrence of the strong nation; ours emphasizes the strong nation's incentive to wipe out the WMD if it does exist. Not surprisingly, greater ambiguity in B&S makes the weak nations decision to build a WMD less likely, whereas in our model greater ambiguity as reflected in a less reliable intelligence system never makes the build decision less likely and makes it more likely over a significant range for intelligence reliability.

Debs and Monteiro, hereafter D&M, provide a detailed and insightful analysis of why the United States invaded Iraq, respectively a strong and a weak nation. They attribute that decision to the combination of Iraq's inability to commit not to develop nuclear weapons, and the United States' inability to definitively conclude that Iraq was not pursuing such an effort. D&M then couple their analysis with a game-theoretic model involving a strong and a weak nation. Their model shows that when a strong nation has the capability to wage a preventive war, and has highly capable intelligence, a weak nation will refrain from making a power-shifting military investment. That nation understands that a preventive war will be the result, making it worse off than if it had never invested. However, with lesser intelligence capabilities, the strong nation may launch a preventive war, as the United States did against Iraq, even though the weak nation neither had the feared weapon, nor was developing it. D&M also consider differences between the Iraq situation and the situation with Iran prior to the treaty.

Our model focuses on the weak nation's type, as defined by its preferences, assumed to be unknown to the strong nation. One of our two types,the Provocateur, prefers, assuming that it has no weapon, to receive rather than not to receive an unjustified attack. This would raise its standing in the world and with its domestic population, and will besmirch its enemy, the strong nation. In equilibrium, for some parameters of our model (i) the weak nation may not develop the weapon, (ii) the strong nation's imperfect intelligence may indicate that, (iii) yet the strong nation may still launch a preventive war even if the intelligence is of very high quality. Not knowing whether the weak nation would welcome an unjustified attack makes the world more conductive to preventive wars. Saddam Hussein's last-minute behavior, when an attack was imminent, appeared consistent with the weak nation's having such a preference. Unfortunately, given fallible intelligence, mistaken wars are a real possibility, as happened here.

D&M also address the question whether a weak nation, which may develop the WMD, will allow or not the public inspection. But in their model the information about the strong nation's preferences is not complete. The strong nation can attack the weak one not only in order to prevent it from building the bomb, but also for other reason, for example, to get control of a weak nation's resources.

Our model also relates to an earlier literature that addresses a major puzzle for rational theory: Why do nations go to wars that destroy value for all parties. In a classic paper, Fearon (1995) identifies three classes of possible explanations: the outcomes are indivisible, and no division satisfies both parties; there are asymmetries of information between the two parties, for example about preferences or capabilities; the nations may not be able to commit to a mutually preferable (Pareto-superior) bargain. A rich literature has developed following the second and third of these explanations. The asymmetric information explanation is well known to game theorists in a range of contexts, where such asymmetries produce inferior outcomes.[3] On the commitment explanation see, for example, Powell (2004) and Powell (2006). Much of this work assumes that there is a pie that has to be divided, and that conflict destroys a portion of the pie.

Our model also relates 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 (see also JTZ).

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

[3] Powell (1996) examines bargaining breakdowns, including value-destroying wars, given information asymmetries. His model adds private information about costs to the full-information Rubenstein bargaining model. See also Meirowitz and Sartori (2008), Bas and Coe (2012) and Arena and Wolford (2012).

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 what action 1 took. Another difference is that our Provocateur 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 he sends given the action they took.

Another related topic deals with audience costs (Fearon, 1994, Schultz, 1998, 1999, Debs and Weiss, 2016 and Moon and Souva, 2016), which introduce the flip side of information conveyance about preferences between nations whose interests are significantly opposed. For a detailed discussion see JTZ. These analyses on audience costs all focus, as do we, on the potential that informational asymmetries between rival nations have the potential to promote unwanted conflicts. Our analysis relates to but does not directly address the literature on audience costs, because our nation 1 is an autocratic, closed regime. One potential justification for such a nation holding Provocateur preferences is that an unjustified attack may strengthen the regime. Concern for such strengthening could be viewed as attending to audience costs in the context of a non-democratic regime. Despite the different contexts, our analysis produces an analogous finding to the works of Debs and Weiss and of Moon and Souva. High quality information – as possessed by the intelligence system of nation 2 – and an accurate signal may not avoid the potential for a situation escalating to a conflict despite no threat being present.[4]

We find that nation 2 becomes more prudent the more accurate is its information regarding nation 1's possession of weapons of mass destruction. Surprisingly, due to the feedback loop on nation 1's decisions, this result persists even if nation 2 receives information confirming its fears about nation 1's WMDs. The more reliable is that information the more hesitant is nation 2 to attack nation 1's weapon-making facilities.

Our model relates 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).

Wittman (1989) and O'Neill (1991) study an arms control verification system in an arms race. One of their results is similar to one of our results even though their models do not

[4] Note, our context also differs, since nation 2's immediate concern is wiping out weapons of mass destruction, not a territorial dispute.

allow players to open their facilities for inspection. Finally, this paper significantly extends the unpublished paper of Biran and Tauman (2009), which also has no possibility for Player 1 to allow inspection.

Brams and Kilgour (2017) study another aspect of the impact of intelligence on the enemy in international conflict games. In their Prediction Game two players can obtain a non-Nash equilibrium outcome, which is Pareto-optimal, if each player predicts (by collecting intelligence) that the second player seeks to deviate from a Pareto-inferior equilibrium.

The driving factor in our model is that one player does not know the other player's type, though the latter will voluntarily reveal its type in certain circumstances. This general framework could be profitably applied in the contest literature, which generally assumes that the competitors know each other's type. In many real world contexts, that assumption of knowledge may not be justified. Thus, for example, in a patent race, company A may not know company B's stage of development or what scientific path it is following. In some circumstances, B may wish to reveal its stage say by presenting at a professional meeting so as to dampen the efforts of A. Industrial espionage plays an equivalent role to the intelligence system in our model. And highly wasteful contests are the equivalent of our severely inferior outcome, when nation II attacks even though nation I has no WMDs, and that is what II's intelligence revealed. Long (2013) provides an insightful overview of the contests literature.

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,

2 D	NA	A
NB	$w_D, 1$	r_D, r_2
B	$1, 0$	$0, w_2$

2 P	NA	A
NB	$r_P, 1$	w_P, r_2
B	$1, 0$	$0, w_2$

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

Figure 2.1: The game and payoffs

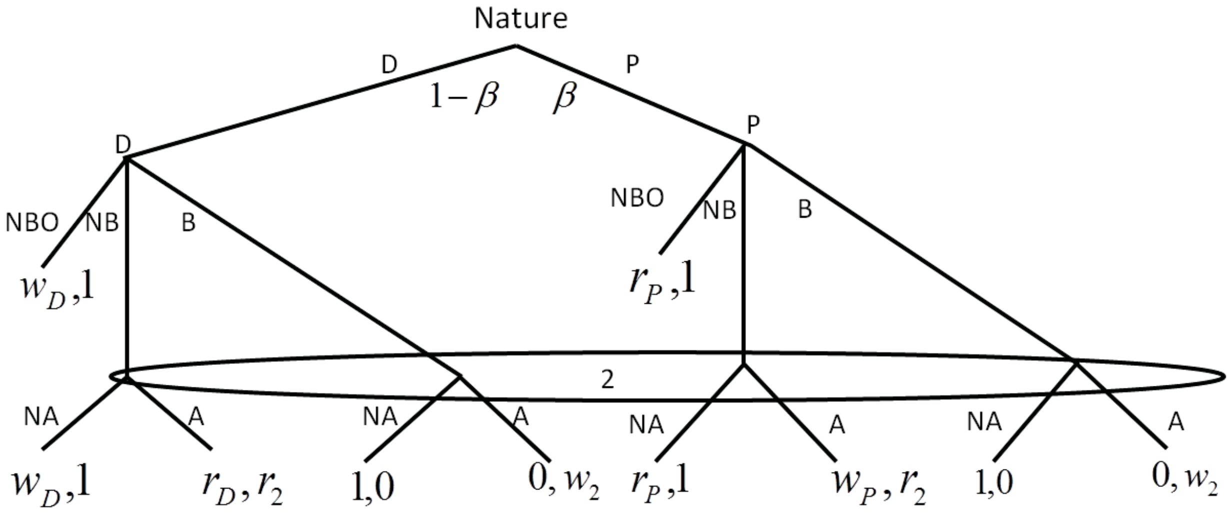

Figure 2.2: The game G_β

(NB,NA), since she will suffer a severe loss of legitimacy. His worst outcome is that 1 possesses the bomb undisturbed (B,NA).

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 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

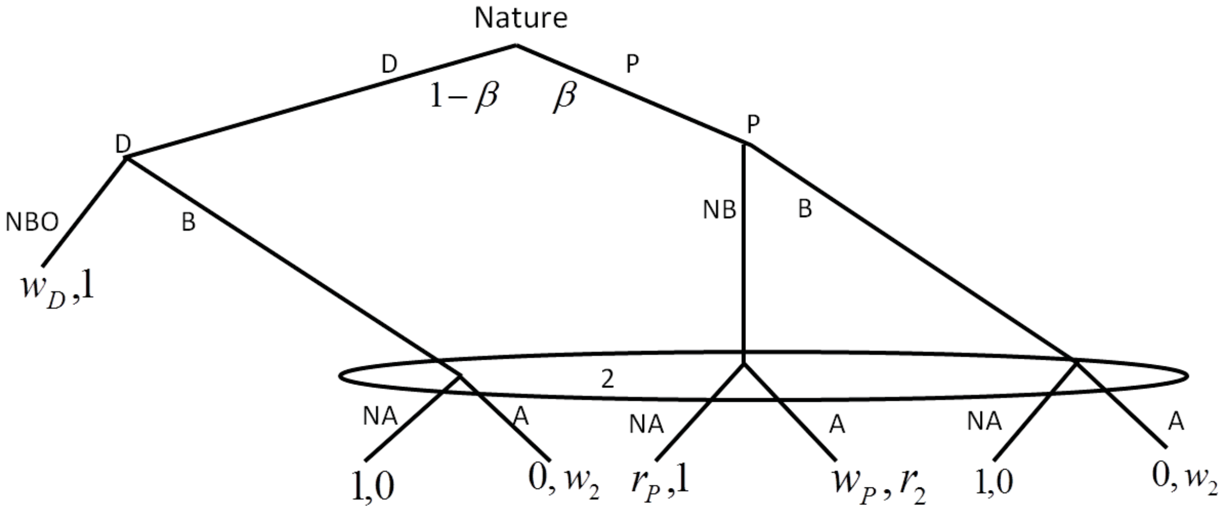

Figure 2.3: The reduced game G_β

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 strategies 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 $\beta = 1$). □

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 prefers to be unjustifiably attacked, 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. 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 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 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

Example Suppose $r_P = 0.1$, $w_P = 0.15$, $r_2 = 0.5$ and $w_2 = 0.75$. It can be verified that

$$\alpha = 0.8 \Rightarrow Prob((NB, A)) = 0.136$$

$$\alpha = 0.75 \Rightarrow Prob((NB, A)) = 0.2$$

$$\alpha = 0.9 \Rightarrow Prob((NB, A)) = 0.09$$

That is for some parameters and with relatively high IS precision with positive probability 1 does not build the bomb and 2 still attacks him.

Some 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 for $\alpha > \alpha_P$

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

where $Prob_\alpha(B|s)$ is the probability that 1 chooses B given $s \in \{nb, b\}$ when the precision of IS is α, 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 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. Namely, for all $\frac{1}{2} < \alpha < \alpha_P$ and $\alpha_P < \alpha' < 1$

$$Prob_{\alpha}(B|s) > Prob_{\alpha'}(B|s), s \in \{b, nb\}.$$

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,NA) 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 3.2 (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.

Example Suppose $r_P = 0.05$, $w_P = 0.8$, $w_D = 0.05$, namely, w_D is very small, $r_D < 0.05$, $r_2 = 0.5$ and $w_2 = 0.75$. Suppose also $\alpha = 0.9$ and $\beta = 0.14$. It can be verified that

$$x_D = 0.023$$

$$x_P = 1$$

$$y(A|nb) = 0.5$$

$$Prob(NB|nb) = 0.41$$

$$Prob(NB, nb, A) = 0.072$$

That is for some parameters and with relatively high IS precision with positive probability 1 does not build the bomb, the IS sends the "nb" signal and 2 still attacks him. Note that only P can be unjustifiably attacked. If D does not open up his facilities he chooses B for sure.

Proposition 4 asserts that for all α (except $\alpha = \alpha_P$), 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.

That is, irrespective of the quality of IS, with positive probability 2 attacks 1 of either type even if 1 does not build the bomb and IS sends the correct signal 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.

(ii) Suppose that $\alpha > \alpha_P$. Player 2 will not attack P if the signal is nb and will randomize between A and NA if the signal is b.

(iii) 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.

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

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 $\alpha \in (\frac{1}{2}, 1]$.

3 Conclusion

As weapons of mass destruction have spread, the world has confronted a new variety of threat: A weak, authoritarian nation is able to acquire or appear to be acquiring such weapons. Such weapons, if possessed, would threaten much stronger, democratic nations, which in turn had the capability to attack those weapons. Our model dealt with one set of specific preferences for the two nations. But the more general lesson is that when nations are unsure of each other's preferences, some counterintuitive and at times mutually non-beneficial outcomes may occur. Thus one nation, in deciding whether to build nuclear weapons at great expense, may not know whether its more powerful enemy would attack if it did not possess them. To some extent, this was Saddam's problem. He did not have the weapons. He knew that the United States had inaccurate, indeed fraudulent, evidence that he actually did have such weapons. Saddam did not know whether the United States knew that evidence was false. If he thought it knew, he may have been confident - perhaps too confident - that the United States would not attack. If no attack would come, he could avoid having to suffer the humiliation of opening his nuclear facilities.

The specific model studied in this paper highlighted two major uncertainties. First, what is the stage of the weak nation's weapons development. Second, is that nation seeking to deter an attack on its weapons and weapons facilities, or would it welcome such an attack, particularly if such weapons do not exist, as a means to curry support from both domestic and international audiences.

A strong nation will use its intelligence capabilities to assess the weak nation's progress in acquiring WMDs. Such intelligence is imperfect. A weak nation can also open its facilities to inspection to reveal its lack of WMDs. But intelligence on the preferences of the weak nation is no doubt much more imperfect, and it is often be impossible for the weak nation to convincingly convey its preferences. These dual uncertainties create situations where costly "accidents" can happen, namely an attack where given full information both the weak and the strong nation would prefer no attack.

References

P. Arena and S. Wolford. Arms, intelligence, and war. *International Studies Quarterly*, 56(2): 351–365, 2012.

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.

S. Baliga and T. Sjöström. Strategic ambiguity and arms proliferation. *Journal of Political Economy*, 116(6):1023–1057, 2008.

M. A. Bas and A. J. Coe. Arms diffusion and war. *Journal of Conflict Resolution*, 56(4): 651–674, 2012.

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

S. J. Brams and D.M. Kilgour. Stabilizing unstable outcomes in prediction games. *Working paper*, 2017.

A. Debs and N. P. Monteiro. Known unknowns: Power shifts, uncertainty, and war. *International Organization*, 68(01):1–31, 2014.

A. Debs and J. C. Weiss. Circumstances, domestic audiences, and reputational incentives in international crisis bargaining. *Journal of Conflict Resolution*, 60(3):403–433, 2016.

J. D. Fearon. Domestic political audiences and the escalation of international disputes. *American Political Science Review*, 88(03):577–592, 1994.

J. D. Fearon. Rationalist explanations for war. *International organization*, 49(03):379–414, 1995.

A. Jelnov, Y. Tauman, and R. Zeckhauser. Attacking the unknown weapons of a potential bomb builder: The impact of intelligence on the strategic interaction. *Games and Economic Behavior*, 2017.

N.V. Long. The theory of contests: A unified model and review of the literature. *European Journal of Political Economy*, 32:161–181, 2013.

A. Meirowitz and A. E. Sartori. Strategic uncertainty as a cause of war. *Quarterly Journal of Political Science*, 3(4):327–352, 2008.

C. Moon and M. Souva. Audience costs, information, and credible commitment problems. *Journal of Conflict Resolution*, 60(3):434–458, 2016.

B. O'Neill. Why a good verification system can give ambiguous evidence. *YCISS working paper*, 1991.

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.

R. Powell. Bargaining in the shadow of power. *Games and Economic Behavior*, 15(2):255–289, 1996.

R. Powell. The inefficient use of power: Costly conflict with complete information. *American Political Science Review*, 98(02):231–241, 2004.

R. Powell. War as a commitment problem. *International organization*, 60(01):169–203, 2006.

T. C. Schelling. *The Strategy of Conflict*. Harvard University Press, 1960.

K. A. Schultz. Domestic opposition and signaling in international crises. *American Political Science Review*, 92(04):829–844, 1998.

K. A. Schultz. Do democratic institutions constrain or inform? Contrasting two institutional perspectives on democracy and war. *International Organization*, 53(02):233–266, 1999.

D. Wittman. Arms control verification and other games involving imperfect detection. *American Political Science Review*, 83(03):923–945, 1989.

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

By (1), $q = w_D$. 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 3.1 describes the sequence of events and the possible outcomes of the game.

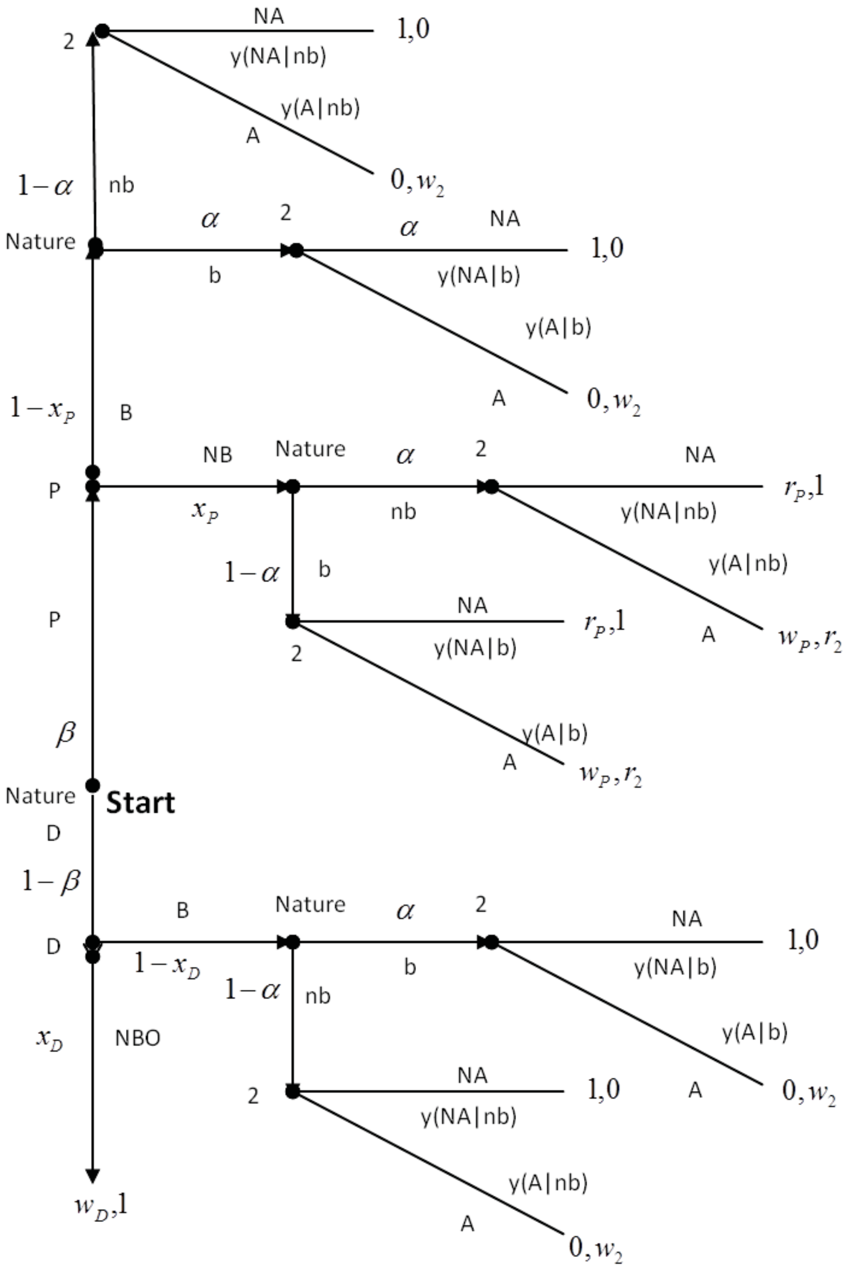

Figure 3.1: The reduced game of $G_{\alpha,\beta}$

D weakly prefers B on NBO (we write $B \succeq_D NBO$) iff (see Figure 3.1)

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

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 3.1) 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 3.1) 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

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

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

Lemma 3A In every se $y(A|b) > 0$ 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),

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

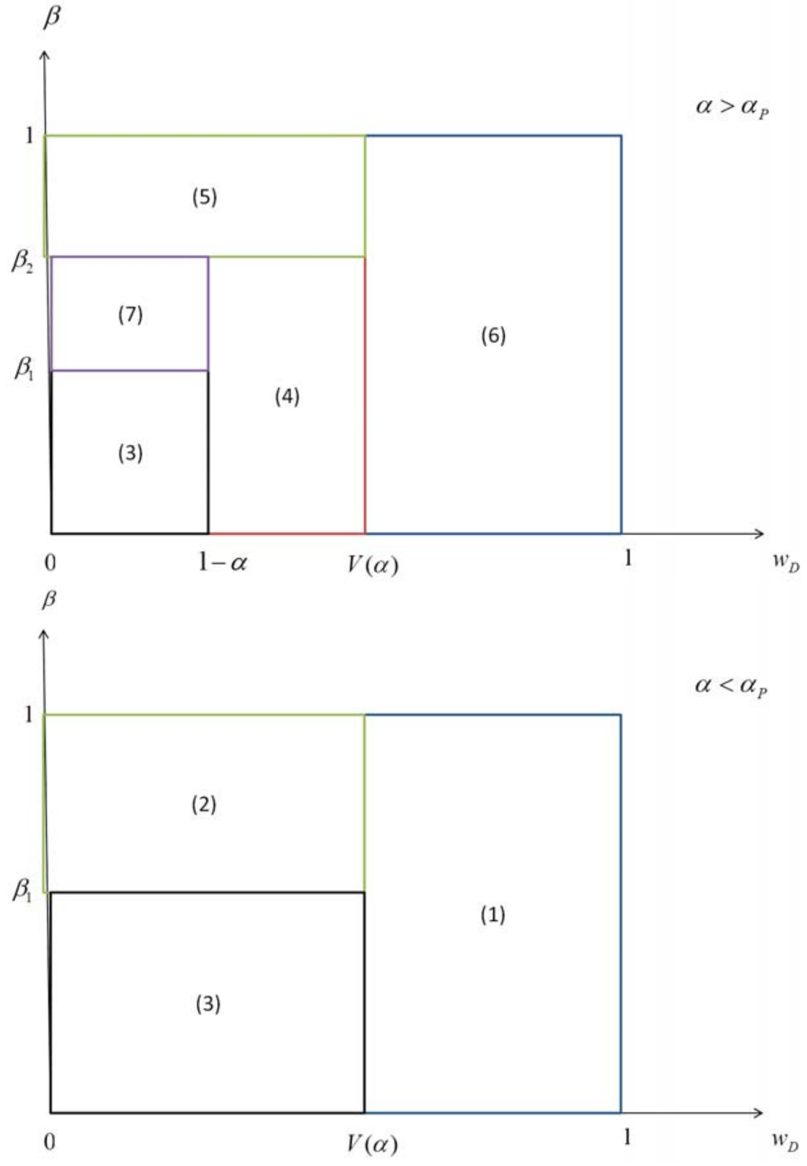

Figure 3.2: Areas of equilibrium outcomes

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 3.2 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

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

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

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

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

$$\Pi^*_2(\alpha) = x^*_P(\alpha),$$

and it is increasing in α.

□