Games and Economic Behavior 104 (2017) 177–189

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Attacking the unknown weapons of a potential bomb builder: The impact of intelligence on the strategic interaction

Artyom Jelnov [a], Yair Tauman [b,c], Richard Zeckhauser [d]

[a] Ariel University, Israel
[b] The Interdisciplinary Center, Herzliya, Israel
[c] Stony Brook University, NY, USA
[d] Harvard University, USA

ARTICLE INFO

Article history:
Received 19 May 2015
Available online 27 January 2017

JEL classification:
D74
D82
C70

Keywords:
Intelligence system
Deterrence
Weapons of mass destruction
Incomplete information
Pooling and separating equilibria

ABSTRACT

Nation 1 wants to develop a nuclear bomb (or other weapons of mass destruction). Nation 2, its enemy, wants to prevent this, either by requiring that 1 open his facilities, or through a pinpoint strike if her imperfect intelligence system (IS) indicates a bomb is present or imminent. If 1 refuses full inspection, 2 can attack 1 or not. 1's cost for allowing inspection, private information, can be either high, H, or low, L. The game's unique sequential equilibrium will be separating or pooling, depending on the precision of IS. The equilibrium is fully characterized. Surprisingly for less accurate IS, 2 behaves aggressively – her appetite to attack is strong. Highly accurate IS dampens that appetite. The following tragic outcome arises in equilibrium with positive probability: 1 does not develop the bomb; 2's IS correctly signals 1's decision; 1, regardless of type, refuses to open its facilities; 2 attacks 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 having or being close to securing the bomb. Player 2 is capable of and willing to destroy Player 1's potential bomb-making facilities as well as its bomb(s) if they exist. It is not clear, however, whether Player 1 is close to a bomb, and Player 2 would lose significant value if she attacked and bomb making was not at least at an advanced stage. Uncertainty about the status of Player 1's bomb drives this analysis.

In what follows, for ease of exposition, when we say Player 1 builds or has the bomb, that concept should be understood to include being at an advanced stage in producing a bomb, for example having it within a year or two. Similarly, when we discuss destroying the bomb, that concept should include destroying any facilities that could produce a bomb imminently.

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 his facility to reveal that he 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, she employs an imperfect 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 bomb making is at an advanced stage, or not, each with probability α,

E-mail addresses: artyomj@ariel.ac.il (A. Jelnov), amty21@gmail.com (Y. Tauman), richard_zeckhauser@hks.harvard.edu (R. Zeckhauser).

http://dx.doi.org/10.1016/j.geb.2017.01.006
0899-8256/

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. In most important respects, the set up this far parallels that of Debs and Monteiro (2014). (See 1.1 Related Literature below.)

We depart from Debs and Monteiro by focusing the analysis on a second critical uncertainty, the preferences, i.e., the type of Player 1. Opening his facilities for a full inspection can be very costly for one nation and less costly for another nation. For instance, just before the second Gulf War (2003) many even within Iraq's ruling circle believed that Saddam possessed WMDs. Revealing the truth would weaken Saddam not only internally but also in the Arab world. Moreover, in Saddam's view, this could encourage his two enemies, Iran and Israel, to attack him. The cost of fully complying with inspections was high to Saddam. Moreover, Saddam did not believe coalition ground forces would ever reach into the heart of Iraq. He expected that his regime would survive whatever conflict ensued. As a result, while allowing UN inspectors on Iraqi soil, Saddam did not fully cooperate with them in an attempt to create the illusion that he might well have had WMDs. Other regimes may prefer to come clean about WMDs by complying fully with inspections to avoid either sanctions or attack. It seems that Iran's net cost of compliance with inspection was relatively small taking into account the continuation of tough sanctions and, consequently, possible unrest as an alternative. Assad gave up his chemical weapons in response to international threats if he refused. While probably the cost of doing so were not low, the net cost was low, given what he thought would happen under the alternative.

Player 1 can be one of two types, depending on the cost to him of allowing inspections. He can be High cost, H, or Low cost, L, where the ex ante likelihood of H is the common knowledge parameter β. Player 1's type is neither known, nor revealed publicly, but he can open his facilities to show no bomb is present. Hereafter, we refer to Player 1 as 1, with types H and L, and Player 2 as 2.

1 chooses whether to build, B, or not build, NB. If he doesn't build, he may choose to open, NBO, to reveal that fact. If 1 chooses not-to-open, 2 can attack, A, or not attack, NA.

Player 2's preference ordering from best to worst is (NB,NA), (B,A), (NB,A) and (B,NA), and she is indifferent between NBO and (NB,NA). Both L and H have (B,NA) as their best outcome; and (B,A) as their worst outcome; they both prefer the outcome (NB,NA) to (NB,A). L, given his lower cost of complying with inspection, may prefer to open to avoid attack. H always prefers attack to opening. Thus NBO is better than (NB,A) for such an L, and the reverse for H. Thus, NB dominates NBO for H.

The analysis in this paper examines what happens as two parameters shift. The first is L's cost of inspection, which may or may not make it worthwhile to open. The second is the capability of the IS system. The results are often surprising, at least without further reflection.

1.1. Related literature

Our paper is closely related to two prior works, one by Baliga and Sjöström (2008), the other by Debs and Monteiro (2014). Those papers, as do we, 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 lies 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 imperfect information on both sides as to player types. Their weak nation can be 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 paper has fewer moving parts than 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 would prefer to open his facilities to avoid an attack while the other type would refuse to open even if it knew it would be attacked. He would lose too much face or legitimacy with internal and/or external constituencies if he opens. 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, opposite to ours, focuses on deterrence of the strong nation; while we emphasize the strong nation's incentive to wipe out the WMD if it does exist. Not surprisingly, greater ambiguity in B&S makes the weak nation's 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.[1] One of our two types for the weak nation would regard opening his facilities as the inferior action, whether or not it had a nuclear weapon, and whatever the probability of attack is it refuses to open. To "capitulate" and open would diminish its standing with its domestic population and with some external constituencies. This assessment by the West over Iran's leadership's preference was a possibility until 2015, when it signed the treaty with the G5+1. It was impossible for external parties to discern the preferences of Iran's leadership. In equilibrium, for some parameters of our model with positive probability (i) the weak nation will not develop the weapon, (ii) it will still refuse international inspection, and (iii) the strong nation will still launch a preventive war even if its intelligence is of relatively high quality and sends the correct signal. 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 for other reasons, 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; and 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.[2] 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 analysis also focuses on asymmetric information, and its potential for leading to a highly inefficient outcome. However, only a tortured interpretation would construe it as a bargaining or pie-division model. Rather, it follows the pioneering tradition of Schelling (1960), who stressed the non-zero sum nature of potential international conflicts, showing that "enemies" have some elements of conflict and some of cooperation.[3] A major departure in our model is having one type for the weak nation being passionately unwilling to open, with the possible or likely result of an unwarranted attack that hurts both the weak and strong nation. The attacks in our model are implicitly in accord with the potential attacks by strong nations on the WMDs of weak nations in the current era. Such attacks make conscientious efforts to limit fatality numbers and keep collateral damage low. They are often described as pin-prick attacks.[4] By contrast, Schelling's work, which implicitly and explicitly focused of potential nuclear wars that would hit population centers, stressed that both sides would suffer devastating losses.

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. In such games, verification typically involves sampling data generated by the activities of the agents. The agents therefore alter their activities in ways that seek to 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 secure an advantage. The

[1] Given our focus on preference types, pooling and separating equilibria play a significant role in our model. There is no equivalent in the D&M model, since their is of complete information about players' preferences.

[2] 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).

[3] Schelling's work also pioneered the investigation of commitment in game-theoretic situations, though his focus was mostly on how to make threats or promises credible.

[4] The Israeli attack on the Osirak reactor led to 11 fatalities.

inspector, employing Bayesian methods, will start with a prior that indicates that they are likely adherers. 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 what action 1 took. Another difference is that one of our types for player 1 prefers not opening to opening even if it means getting attacked. 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. Of course, future versions of our model could allow for the quality of the IS, the α in our model, 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 the bomb is built.

Kirstein (2014) studied inspection games in sports to detect an athlete who cheats in a contest (for instance, taking forbidden drugs). An imperfect signal about athlete's decision is sent to an enforcer. Following the signal, the enforcer decides whether to use sanctions against the athlete or not. An analogy to our set up is clear. However, there are important differences. In Kirstein (2014) players' preferences are different than in our model. For example, for player 1 (the athlete) to behave well and to be punished is the worst outcome, while in our model, the worst outcome for Player 1 is not to behave well (to build the bomb) and to be punished. Player 1 in our model can prove his innocence by opening his facilities for inspection, in Kirstein (2014) this option does not exist. Kirstein (2014) deals with a complete information environment while our model is of incomplete information about Player 1's preferences. Finally, the game in Kirstein (2014) results with multiple equilibrium points while in our model it is unique.

Audience costs (Fearon, 1994) introduce the flip side of information conveyance about preferences between nations whose interests are significantly opposed. Consider a nation where there is competition for political leadership. Within it, political leaders suffer significant costs from domestic audiences when they issue a challenge to the status quo and then back down. These audience costs enable those leaders to make more credible commitments. These commitments enhance the reliability of the information on the preferences on which those leaders will act. Fearon shows that they may force a government into a war that it would prefer not to fight.

Recent work has extended, and in so doing has added nuance to the traditional implications of audience costs on the likelihood of conflict between democratic nations. Such costs can make war less likely, since escalation followed by backing down becomes much more expensive to a leader. Perhaps more important, an opposition party possesses information about the likely consequences of war, and its potential to signal such information to the electorate makes the ruling party more prudent in issuing challenges and reduces the information asymmetries between the rival states (Schultz, 1998). As Schelling (1960) noted, such asymmetries can frequently foster conflicts. Schultz (1999) conducts an empirical analysis of 1,654 military disputes from 1816 to 1980, and documents that there were lesser informational asymmetries about democracies' intentions. Hence, they bluffed less when issuing threats, and their rivals were less likely to resist such threats.

Debs and Weiss (2016) build on this work, but introduce two additional uncertainties the domestic audience faces: Are circumstances favorable for using force?; and, Is the leader competent at using force? Their analytic model shows that greater knowledge about such uncertainties may increase not diminish the prospects for conflict.

Moon and Souva (2016) also extend our understanding of audience costs and the likelihood of conflict. They consider that likelihood when credible commitment problems as well as information-asymmetry problems are present. When credible commitment is a problem, they argue analytically and show empirically, information-only models may lead to the wrong conclusion: "reducing uncertainty about resolve through the use of costly signals may not reduce the likelihood of conflict escalating" (Moon and Souva, 2016, p. 435).[5]

These analyses all focus, as do we, on the potential that informational asymmetries between rival nations have 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. 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.[6]

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.

[5] Their analysis employs the 211 cases of territorial disputes from 1919 to 1995, as identified by Huth and Allee (2002).

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

L \ 2	NA	A
NBO	$w_1 - c_L, 1$	
NB	$w_1, 1$	r_1, r_2
B	$1, 0$	$0, w_2$

H \ 2	NA	A
NBO	$w_1 - c_H, 1$	
NB	$w_1, 1$	r_1, r_2
B	$1, 0$	$0, w_2$

$0 < r_1 < w_1 < 1$
$0 < c_L < w_1 - r_1 < c_H$
$0 < r_2 < w_2 < 1$

Fig. 2.1. The game payoffs.

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).

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 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 the possibility for Player 1 to allow inspection.

2. The model

There are two players. 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 decide to 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 attempt to build the bomb. We first deal with the case where Player 2 has no intelligence capability over Player 1 and she only observes whether 1 opens or does not open his facilities. If Player 1 refuses to open, Player 2 must decide whether to Attack, A, or not attack, NA.

Player 1 can be one of two types, H (high cost) or L (low cost). The costs c_L and c_H can be either internal or external (or both) associated with opening the facilities for a full inspection. The likelihood that 1 is of type H is β, which is common knowledge. Both L or H regard the outcome (B,NA) as best and (B,A) as worst of the five possibilities. Both of them regard (NB,NA) as superior to (NB,A). However, the outcome following the strategy NBO is worse for H even than (NB,A). Hence NBO strategy is strictly dominated by NB and H never opens. In contrast, the outcome following NBO is better for L than (NB,A), but for H it is worse than (NB,NA). For H the NBO strategy is strictly dominated by NB, hence H never opens. The payoffs of the players of the five possible outcomes is given in Fig. 2.1.

2.1. The case where Country 2 has no intelligence

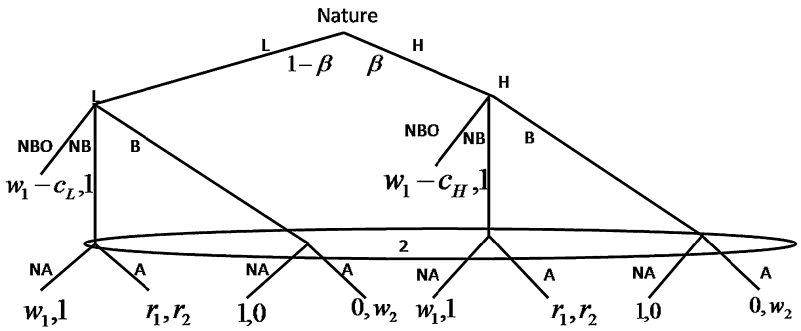
We first analyze the conflict between 1 and 2 under the assumption that 2 has no intelligence on 1. This is the game G_β in a tree form as described in Fig. 2.2. The reduced game is described in Fig. 2.3 and is obtained after eliminating the dominated strategy NBO of H.

The result of this case is simple and intuitive.

Proposition 1. *In every sequential equilibrium of the game G_β:*

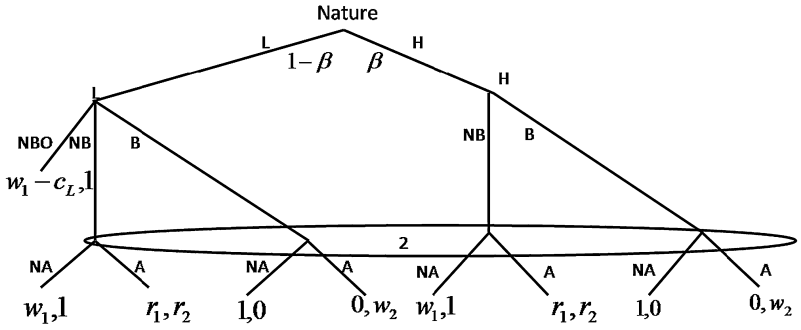
(i) *2 attacks 1 with probability* $\frac{1-w_1}{1-w_1+r_1}$.
(ii) *If* $c_L < \frac{(1-w_1)(w_1-r_1)}{1-w_1+r_1}$, *L chooses NBO with probability* 1, *and H mixes his two pure strategies B and NB.*
(iii) *If* $c_L > \frac{(1-w_1)(w_1-r_1)}{1-w_1+r_1}$, *both L and H are indifferent between B and NB, and they assign zero probability to NBO.*

Proof. See Appendix A. □



$$0 < r_1 < w_1 < 1$$
$$0 < c_L < w_1 - r_1 < c_H$$
$$0 < r_2 < w_2 < 1$$

Fig. 2.2. The game G_β.



$$0 < r_1 < w_1 < 1$$
$$0 < c_L < w_1 - r_1$$
$$0 < r_2 < w_2 < 1$$

Fig. 2.3. The reduced game G_β.

The equilibrium of G_β depends on the magnitude of c_L. If the cost to L of not building the bomb and allowing full inspection is relatively small, the unique sequential equilibrium is separating. L allows full inspection and avoids attack while H mixes each B and NB (i.e., chooses each one of them with positive probability). If c_L is relatively high, every sequential equilibrium is pooling and it can be shown that there is a continuum of them. Yet the probability that 2 assigns to 1 building the bomb (B) is the same across all equilibrium points. Both L and H refuse inspection and each chooses B and NB with positive probability. Player 2 is mixing A and NA irrespective of c_L. It attacks 1 with a probability that increases to 1 when w_1 and r_1 decrease to zero.

Our next goal is to find out how the ability of 2 to use intelligence on 1 affects these results.

2.2. The case where Country 2 has intelligence

Suppose next that Player 2 has an Intelligence System, IS, with quality α. Namely, if 1 chooses B, the IS will send the correct signal b with probability α, and the incorrect signal nb with probability $1-\alpha$. Similarly, if 1 chooses NB, and does not open his facilities, the IS will send the correct signal nb with probability α, and the signal b with probability $1-\alpha$.

Player 2, based on the signal observed decides whether to attack 1 (A) or not (NA). Suppose $\frac{1}{2} < \alpha < 1$ and assume that the value of α is common knowledge. In particular, 1 knows that he is being spied upon, and he knows the reliability of 2's intelligence. Once α and β become common knowledge the game proceeds as follows. (1) 1 chooses (depending on his type) between B, NB and NBO. (2) If 1 chooses NBO, 2 chooses NA, and the game ends. (3) If 1 does not open his facilities,

his choice of B or NB generates a signal for 2 via IS. (4) 2 draws inferences from the signal and chooses whether to attack A, or not attack NA. Payoffs for each outcome are equal to those of G_β. This describes a Bayesian game $G_{\alpha,\beta}$.

Observe that Player 1 has two possible types of ex-post regret. The first (type 1 regret) is regretting not building the bomb (B), given the choice NA of 2. The penalty 1 occurs in this case is $1-w_1$. The second (type 2 regret) is 1's regret of choosing B and not NB, given that 2 chooses A. The penalty to 1 in this case is $r_1-0=r_1$. The equilibrium of $G_{\alpha,\beta}$ depends on the relative size of these two penalties.

Suppose Player 1's main goal is to build the bomb and he is ready to take a considerable risk to achieve this goal. His utility for every other outcome is relatively small, in particular w_1 (and certainly r_1) are relatively small, implying that $1-w_1>r_1$. Namely, the penalty associated with type 1 regret exceeds the one associated with type 2 regret. We show that Player 2 in this case acts surprisingly aggressively for any $\alpha\in(\frac{1}{2},\bar{\alpha})$ where

$$\bar{\alpha}=\frac{1-r_1}{1-r_1+w_1}. \tag{1}$$

This interval is larger the smaller is w_1 and it may contain highly accurate IS. Player 2 attacks 1 for sure if the signal is b and with significant probability even if the signal is nb. For $\alpha\in(\bar{\alpha},1)$ (very accurate IS) Player 2 acts much less aggressively. She for sure does not attack 1 if she observes the signal nb and with positive probability she refrains from attacking 1 even if the signal is b. We provide the intuition of these results after stating them formally.

If however 1 assigns a relatively high utility w_1 to the outcome (NB,NA) then type 2 regret applies ($1-w_1<r_1$). In this case, irrespective of the precision α of IS, 2 acts less aggressively.

Let $\hat{c}:[\frac{1}{2},1]\to\mathbb{R}$ be defined by

$$\hat{c}(\alpha)=\begin{cases}\frac{(w_1-r_1)(1-\alpha-\alpha w_1)}{1-\alpha-\alpha w_1+\alpha r_1} & ,\frac{1}{2}<\alpha<\bar{\alpha}\\ \frac{(w_1-r_1)(1-\alpha)(1-w_1)}{\alpha(1+w_1-r_1)-(w_1-r_1)} & ,\bar{\alpha}<\alpha<1.\end{cases} \tag{2}$$

Proposition 2. *Every sequential equilibrium of $G_{\alpha,\beta}$ belongs to one of the following two sets: non-aggressive and aggressive.*

(i) *The non-aggressive set consist of all equilibrium points where Player 2 for sure does not attack 1 if she observes the signal nb, and only with positive probability (which is decreasing in α) if she observes the signal b. If $c_L<\hat{c}(\alpha)$ the equilibrium is separating: L opens his facility for inspection (NBO) while H mixes B and NB. If $c_L>\hat{c}(\alpha)$ every equilibrium is semi-pooling: both L and H do not open their facilities and they mix B and NB (possibly with different mixtures). The expected equilibrium payoffs of L, H and 2 are nondecreasing in α.*

(ii) *The aggressive set consists of all equilibrium points where Player 2 attacks 1 for sure if the signal is b, and with positive probability (though decreasing in α) if the signal is nb. Similar to the non-aggressive group if $c_L<\hat{c}(\alpha)$ every equilibrium is separating and if $c_L>\hat{c}(\alpha)$ every equilibrium is semi-pooling. Also in this set of equilibria, the expected payoffs of the players are nondecreasing in α.*

(iii) *For every c_L, $0<c_L<w_1-r_1$, the equilibrium is semi-pooling for IS with a sufficiently high quality.*

The proof of part (iii) follows immediately from the definition of $\hat{c}(\alpha)$, which is a decreasing and approaching zero, as $\alpha\to1$. The proof of parts (i) and (ii) appear in the Appendix.

Apply Proposition 2 to the Second Gulf War, and suppose the cost of Saddam Hussein to fully cooperate with the UN inspectors was small. Still for sufficiently accurate IS this cost was higher than the threshold $\hat{c}(\alpha)$ (part (iii) of the proposition) and even as the L-type player his equilibrium strategy was to mimic the H-type and to refuse a full inspection (parts (i) and (ii) of the proposition), as he actually did.

Proposition 3.

(i) *Suppose the penalty associated with type 2 regret exceeds that of type 1 ($r_1\geq1-w_1$). Then every sequential equilibrium of $G_{\alpha,\beta}$ belongs to the non-aggressive group of equilibrium.*

(ii) *Suppose $r_1<1-w_1$. Then every sequential equilibrium of $G_{\alpha,\beta}$ belongs to the aggressive group if $\alpha<\bar{\alpha}$ and to the non-aggressive group if $\alpha>\bar{\alpha}$.*

The proof of the proposition appears in the Appendix.

Some elements of Proposition 3 seem counter intuitive at least at the first glance, since 2's appetite to attack is higher the lower is the quality of IS. Yet intuition returns when we recognize that 2 takes into account how 1 will respond to weaker IS. She will correctly conclude that when her intelligence is less reliable, 1 is more likely to build a bomb, which makes it more attractive for her to attack. There is a critical value $\bar{\alpha}$ of α given by (1). If the precision of IS is relatively low ($\alpha<\bar{\alpha}$) and the signal is less reliable, 2 attacks 1 with no hesitation (namely with probability 1) if the signal of IS is b. 2 still attacks with positive probability even if the signal is nb. While if IS is of high quality ($\alpha>\bar{\alpha}$), and therefore more reliable, 2 hesitates to attack even when the signal is b. She attacks 1 with probability smaller than 1. If the signal

is nb, she never attacks 1. Let us provide some intuition for these results. Suppose first $\bar{\alpha} < \alpha < 1$ (the precision of the IS is relatively high). In this case 1 knows that with relatively high probability his action will be correctly detected by IS. Thus, he chooses to build the bomb with a 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 2) that for $\alpha > \bar{\alpha}$ the probability that 2 assigns to 1 choosing B given the signal b is

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

and it 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 $\frac{1}{2} < \alpha < \bar{\alpha}$. Namely, IS is somewhat informative, but it is not too accurate. 1 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 1 to select B, there is a reasonable chance that B is chosen. Second, by attacking some of the time despite nb, she reduces 1'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 1 has a bomb applies to both signals, b and nb. Namely, for all $\frac{1}{2} < \alpha < \bar{\alpha}$ and $\bar{\alpha} < \alpha' < 1$

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

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

2's expected payoff is higher if her intelligence is better. The greater is α, the more accurate is 2's information; hence she responds more effectively, which in turn makes 1 less likely to build the bomb, as 2 would like. 2's payoff increases. Much more surprising is the fact that the expected payoff of 1 is also increasing in α. This result comes because the probability of an attack on 1 decreases when α increases. 1 is obviously better off if the probability of attack on him decreases. A surprising implication is that it would be in 1's interest to subsidize 2's IS quality, for example, by leaking an information to 2. Note, that if 1 does not build the bomb he can supply a perfect information to 2 by opening his facilities for inspection, but this action is costly to 1 in a way that information leaks may not be.

Recall, $\hat{c}(\alpha)$ is the threshold, where L opens his facilities for inspection if $c_L < \hat{c}(\alpha)$. If 1 does not build the bomb, the higher is the quality of IS, the higher is the probability that 2 obtains the correct signal about 1's action, and 2 then will attack 1 with lower probability. Therefore, the incentive of 1 to open his facilities (a costly action) is lower.

Let $P^{SE}(B)$ and $P^{SP}(B)$ be the probability that 1 builds the bomb in a separating and semi-pooling equilibrium, respectively. It is shown in the Appendix that

$$P^{SP}(B) = \begin{cases} \frac{\alpha(1-r_2)}{(1-\alpha)w_2+\alpha(1-r_2)} & , \frac{1}{2} < \alpha < \bar{\alpha} \\ \frac{(1-\alpha)(1-r_2)}{\alpha w_2+(1-\alpha)(1-r_2)} & , \bar{\alpha} < \alpha < 1. \end{cases} \tag{3}$$

Since in a separating equilibrium L (and not H) opens his facilities we have

$$P^{SE}(B) = \beta P^{SP}(B). \tag{4}$$

By (3) and (4), the probability that 1 builds a bomb is increasing for $\frac{1}{2} < \alpha < \bar{\alpha}$ and decreasing for $\bar{\alpha} < \alpha < 1$. While the latter is intuitive, the former is not. The better is α the better is the probability of correctly detecting 1's action. So why in $(\frac{1}{2}, \bar{\alpha})$ is the probability that 1 builds the bomb increasing in α? Recall that in this region the equilibrium is aggressive and $P(A|b) = 1$ for all $\alpha \in (\frac{1}{2}, \bar{\alpha})$. While the likelihood of b given B is increasing in α ($P(b|B) = \alpha$)), the probability of an attack given b remains unchanged, namely, $P(A|b) = 1$ for all $\alpha \in (\frac{1}{2}, \bar{\alpha})$. In addition there is a significant probability that IS makes a mistake and sends the wrong signal nb (given B), since the IS in this region is not very accurate. As α increases the IS becomes more reliable and $P(NA|nb)$ is increasing (see (15) in the Appendix) making the best outcome (B,NA) for 1 more likely. Hence 1 plays with higher probability B, as α increases.

It is shown in the Appendix that the probability of an attack on 1 given his decision to build the bomb is

$$P(A|B) = \begin{cases} \alpha + \frac{(1-\alpha)[1-\alpha-\alpha w_1+\alpha r_1-r_1]}{1-\alpha(1+w_1-r_1)} & , \frac{1}{2} < \alpha < \bar{\alpha} \\ \frac{\alpha(1-w_1)}{\alpha(1+w_1-r_1)-(w_1-r_1)} & , \bar{\alpha} < \alpha < 1. \end{cases} \tag{5}$$

Surprisingly $P(A|B)$ is decreasing in α. Namely, even though it is of the best interest of 2 to attack and eliminate the facilities of 1 if he builds the bomb, yet the higher the precision of IS to detect 1's action, the lower is the probability of 2 to launch an attack on 1. Let us try to explain why this happens. Consider the case where $\bar{\alpha} < \alpha < 1$. In this case every equilibrium is non-aggressive and 1, as expected, builds the bomb with a probability which is decreasing to zero, as

α approaches 1. 2 therefore expects to obtain the signal nb. When 2 observes b she assigns significant probability that the signal is mistaken and as a result she refrains with significant probability from attacking 1. This probability is increasing to w_1, as α approaches 1. As for the case where $\frac{1}{2} < \alpha < \bar{\alpha}$, the equilibrium is aggressive and even though $P(b|B) = \alpha$ is increasing in α, 2 attacks 1 with certainty if she observes b, irrespective of α. Given B, the probability of the signal nb, $P(nb|B) = 1 - \alpha$, is decreasing in α and in addition 2 attacks 1 when observing the signal nb with a probability which is also decreasing in α (see (15) in the Appendix). Consequently, $P(A|B)$ is decreasing in α for $\frac{1}{2} < \alpha < \bar{\alpha}$.

The probability of a mistaken attack is

$$P(A|NB) = \begin{cases} 1 - \alpha + \frac{\alpha[1-r_1-\alpha(1+w_1-r_1)]}{1-\alpha(1+w_1-r_1)} & , \frac{1}{2} < \alpha < \bar{\alpha} \\ \frac{(1-\alpha)(1-w_1)}{\alpha(1+w_1-r_1+r_1)-(w_1-r_1)} & , \bar{\alpha} < \alpha < 1 \end{cases} \tag{6}$$

and as expected $P(A|NB)$ is decreasing in α. Note that (5) and (6) apply only to the case where 1 does not open his facilities for inspection. By definition $P(A|NBO) = 0$. By (6) a mistaken attack happens with positive probability which is decreasing to zero as α approaches 1.

Finally, a mistaken attack can occur in equilibrium even if 1 chooses not to build the bomb and the IS detects that fact correctly. By Propositions 2 and 3 this can not happen if $\bar{\alpha} < \alpha < 1$, since in a non-aggressive equilibrium 2 does not attack following the signal nb. But this happens with positive probability if $\frac{1}{2} < \alpha < \bar{\alpha}$.

Example. Suppose, $r_1 = 0.1$, $w_1 = 0.15$, $c_L = 0.015$ and $c_H > 0.1$. Also suppose $r_2 = 0.2$, $w_2 = 0.7$, $\beta = 0.9$ and the precision of IS $\alpha = 0.8$. Then $\bar{\alpha} = \frac{1-r_1}{1-r_1+w_1} = 0.818$ and $\hat{c}(\alpha) = 0.333$. Since $c_L < \hat{c}(\alpha)$ the equilibrium is separating and belongs to the aggressive group (Proposition 2). L chooses NBO and avoids attack. By (5), (6) and (4) it can be verified that H chooses NB with probability 0.18. The overall probability that 2 attacks is the weighted average of attack when H builds the bomb and when H does not build the bomb. It is

$$P(A) = 0.9[0.82(0.8 \times 1 + 0.2 \times 0.375) + 0.18(0.2 \times 1 + 0.8 \times 0.375)] = 0.73.$$

The disturbing major finding that emerges from this example is the sizeable likelihood of a thoroughly mistaken attack. The conditional probability of an attack given that 1 chooses NB, thus revealing himself to be H, and IS correctly signals nb is $P(A|NB, nb) = 0.375$. That is, 37.5% of the time when no bomb is built, and intelligence accurately reveals that fact, 2 still attacks. Arguably, this was the situation with Saddam Hussein, taking his refusal to cooperate fully to represent not opening, just before what turned out to be the mistaken attack that led to his demise. It is to the strategic interaction of potential attackers with the regimes of both Saddam Hussein and the Ayatollah Khameini that we now turn.

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 occurs high or low cost should be open his facilities for inspection. The regimes of Saddam Hussein in Iraq and of the Ayatollah Ali Khamenei in Iran, the former vanquished and dead, the latter the current supreme leader, exemplify player 1's for whom these uncertainties were present.[7] 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 past weapons, to completely comply with his disarmament obligations, or to fully cooperate on opening his facilities even when an imminent coalition invasion was likely. Saddam regarded revealing to internal supporters and his enemies Iran and Israel that he lacked weapons of mass destruction to be very costly especially since he believed that no matter what he would not be overthrown.

The recent situation with the Ayatollah Ali Khamenei differed from the 2003 confrontation with Saddam Hussein on weapons concerns. With the Ayatollah it was the potential for nuclear weapons whereas for Saddam it was more the existence of biological and chemical weapons. Nevertheless, the two cases bear several similarities. 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 was hard to discern the time schedule on which Iran could have built nuclear weapons.[8,9] The West knew that Iran's capabilities were advanced, but not how advanced, and not whether it was currently pushing further advancement. One of

[7] 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.

[8] 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.

[9] Miller and Bunn (2014) focus on American perceptions of Iran from the beginning of the Islamic Republic. They explain why America had 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.

the most tantalizing uncertainties about Khamenei was his true view on nuclear weapons. His alleged fatwa against them had been widely publicized by his regime, in particular in international forums. It looks as if the net cost of allowing full inspection was not especially high since Iran was suffering severely economically in 2015, and recent elections and protests had revealed severe dissatisfactions with the regime, particularly among the elite.

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 and has chosen not to reveal whether he has such weapons. 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 her 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 outcome ranking of the potential bomb builder may be hard to fathom. That was the case with Iran and with Saddam Hussein and exists with North Korea today. We consider two possible types for the bomb seeker, with two different costs associated with opening their facilities for inspection. 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 strong evidence that even at the last moment Saddam did not believe that the US would invade,[10] and that he may have refused full compliance 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: A nation rarely knows the exact motives of an enemy. Despite that uncertainty, strategic thinking still yields significant insights.

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.

Appendix A

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) H is indifferent between playing B or NB.*

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

(ii) Suppose H obtains higher payoff by playing B than by playing NB. The payoffs of L are identical to those of H for every outcome, except the one following NBO. Since in equilibrium both L and H have the same assessment of 2's strategy, L's payoff after B is higher that his payoff after NB. So, in equilibrium if 1 does not allow inspection 2 knows 1 chose B and her best reply is a pure A, a contradiction. The same contradiction is obtained if L chooses NBO and H chooses B.

Suppose H obtains higher payoff by playing NB. Then the same is true for L. But then L is better off playing NBO since L strictly prefers NBO on NB. In this case the type of H is identified if 1 does not allow inspection and the best reply strategy of 2 is a pure NA, a contradiction. □

[10] 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 he still was unwilling to fully cooperate with the inspectors. This reinforced any view that Iraq possessed weapons of mass destruction. For the seminal work on such misperceptions see Jervis (1976).

Suppose that 2 attacks 1 with probability $0 < q < 1$. L weakly prefers NBO on B (see Fig. 2.3) iff

$$w_1 - c_L \geq 1 - q. \tag{7}$$

H prefers NB on B iff

$$r_1 q + w_1(1-q) \geq 1 - q. \tag{8}$$

By part (ii) of Lemma 1A (8) holds as equality. Thus,

$$q = \frac{1 - w_1}{1 - w_1 + r_1}$$

and the choice of L follows from (7). □

Proof of Propositions 2 and 3. Similar to Lemma 1A, H does not play in equilibrium a pure B or NB.

Let $y(NA|s)$, $s \in \{b, nb\}$ be the probability that 2 chooses NA given the signal s.

L weakly prefers B to NBO (we write $B \succeq_L NBO$) iff

$$\alpha[y(NA|b) \cdot 1 + y(A|b) \cdot 0] + (1-\alpha)[y(NA|nb) \cdot 1 + y(A|nb) \cdot 0] \geq w_1 - c_L.$$

Equivalently, $B \succeq_L NBO$ iff

$$\alpha y(NA|b) + (1-\alpha) y(NA|nb) \geq w_1 - c_L. \tag{9}$$

Similarly, H weakly prefers B on NB ($B \succeq_H NB$) iff

$$\begin{aligned} &\alpha y(NA|b) + (1-\alpha) y(NA|nb) \\ &\quad \geq \alpha[y(NA|nb) w_1 + (1 - y(NA|nb)) r_1] + (1-\alpha)[y(NA|b) w_1 + (1 - y(NA|b) r_1]. \end{aligned} \tag{10}$$

Since H in equilibrium is indifferent between B and NB, equality holds in (10). Suppose L and H build the bomb with probability x_L and x_H, respectively. Then

$$P_B = \beta x_H + (1-\beta) x_L$$

is the probability that 1 chooses B. Let $P_2(NB|s)$ be the probability that 2 assigns to the event that 1 chose NB if she receives the signal $s \in \{b, nb\}$.

$$P_2(NB|b) = \frac{(1 - P_B)(1-\alpha)}{(1 - P_B)(1-\alpha) + P_B \alpha} \tag{11}$$

and

$$P_2(NB|nb) = \frac{(1 - P_B)\alpha}{(1 - P_B)\alpha + P_B(1-\alpha)}. \tag{12}$$

Note that for $\alpha \in (\frac{1}{2}, 1)$,

$$P_2(NB|b) < P_2(NB|nb). \tag{13}$$

Denote $A \succeq_s NA$ the case where 2 weakly prefers A on NA when observing the signal s. Note, that

$$A \succeq_s NA \text{ iff } P_2(NB|s) r_2 + (1 - P_2(NB|s)) w_2 \geq P_2(NB|s).$$

Equivalently

$$A \succeq_s NA \text{ iff } P_2(NB|s) \leq \frac{w_2}{1 - r_2 + w_2},\ s \in \{b, nb\} \tag{14}$$

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

Lemma 2A. *In every equilibrium either $y(A|b) = 1$ or $y(A|nb) = 0$.*

Proof. Claim 1: There is no equilibrium where $0 < y(A|s) < 1$ for both $s = b$ and $s = nb$.

Suppose to the contrary that $0 < y(A|s) < 1$ for both $s = b$ and $s = nb$. Then (14) holds as equality for both b and nb, implying that $P_2(NB|b) = P_2(NB|nb)$. By (11) and (12), this can not be true for any $\alpha \in (\frac{1}{2}, 1)$.

Claim 2: In every equilibrium $y(A|b) > 0$ and $y(A|nb) < 1$.

Suppose to the contrary that $y(A|b)=0$, namely, $NA \succeq_b A$. If in addition $y(A|nb)=0$ then 2 does not attack 1 irrespectively of the signal she receives. In this case B is the best reply of both L and H implying that 2 is better off deviating to A, a contradiction. Thus $y(A|b)=0$ and $y(A|nb)>0$ must hold. This implies $A \succeq_{nb} NA$ and $NA \succeq_b A$ and by (14) we have

$$P_2(NB|nb) \leq P_2(NB|b),$$

contradicting (13). Consequently $y(A|b)>0$. Similar arguments show that $y(A|nb)<1$. The proof of the lemma now follows by Claim 1. □

By Lemma 2A only the following three cases are possible in an equilibrium.

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

Since (10) holds as equality,

$$y(A|nb)=\frac{1-r_1-\alpha(1+w_1-r_1)}{1-\alpha(1+w_1-r_1)}$$

and $0<y(A|nb)<1$ if $\alpha<\bar{\alpha}$. Since for $1-w_1<r_1$, $\bar{\alpha}<\frac{1}{2}$, the region $\frac{1}{2}<\alpha<\bar{\alpha}$ is non-empty only if $1-w_1>r_1$.

For $\alpha\in(\frac{1}{2},\bar{\alpha})$ denote

$$\hat{c}(\alpha)=\frac{(w_1-r_1)(1-\alpha-\alpha w_1)}{1-\alpha-\alpha w_1+\alpha r_1}.$$

Note that $0<\hat{c}(\alpha)<w_1-r_1$ in $(\frac{1}{2},\bar{\alpha})$. By (9), L chooses NBO if $c_L<\hat{c}(\alpha)$, and chooses B if $c_L>\hat{c}(\alpha)$.

Let

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

By (12) and (14), for $c_L<\hat{c}(\alpha)$, $x_H=x$, and for $c_L>\hat{c}(\alpha)$, $P_B=x$. It is straightforward to verify that the expected payoffs of all players are weakly increasing in α.

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

Since (10) holds as equality,

$$y(A|b)=\frac{1-w_1}{\alpha-w_1+\alpha w_1-\alpha r_1+r_1} \tag{15}$$

and $0<y(A|b)<1$ for $\alpha\in(\bar{\alpha},1)$. For $\alpha\in(\bar{\alpha},1)$ denote

$$\hat{c}(\alpha)=\frac{(w_1-r_1)(1-\alpha)(1-w_1)}{\alpha(1+w_1-r_1)-(w_1-r_1)}.$$

It is easy to verify that $0<\hat{c}(\alpha)<w_1-r_1$ for $\alpha\in(\bar{\alpha},1)$. By substitution of (15) into (9), L chooses NBO if $c_L<\hat{c}(\alpha)$, and chooses B if $c_L>\hat{c}(\alpha)$.

Let

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

By (11) and (14), $x_H=y$ for $c_L<\hat{c}(\alpha)$, and $P_B=y$ for $c_L>\hat{c}(\alpha)$. It is straightforward to verify that expected payoffs of L, H and 2 are weakly increasing in α.

Case 3 $y(A|nb)=0$, $y(A|b)=1$.

Since (10) holds as equality, this case applies only to the case where $\alpha=\bar{\alpha}$. □

References

Arena, P., Wolford, S., 2012. Arms, intelligence, and war. Int. Stud. Q. 56 (2), 351–365.

Avenhaus, R., et al., 2002. Inspection games. In: Aumann, R.J., Hart, S. (Eds.), Handbook of Game Theory with Economic Applications, vol. 3. North–Holland, Amsterdam, pp. 1947–1987.

Baliga, S., Sjöström, T., 2008. Strategic ambiguity and arms proliferation. J. Polit. Economy 116 (6), 1023–1057.

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

Biran, D., Tauman, Y., 2009. The Decision to attack a nuclear facility: the role of intelligence. Unpublished manuscript.

Debs, A., Monteiro, N.P., 2014. Known unknowns: power shifts, uncertainty, and war. Int. Organ. 68 (01), 1–31.

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

Dresher, M., 1961. Some military applications of the Theory of Games. RAND Corporation.

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

Fearon, J.D., 1994. Domestic political audiences and the escalation of international disputes. Amer. Polit. Sci. Rev. 88 (03), 577–592.

Fearon, J.D., 1995. Rationalist explanations for war. Int. Organ. 49 (03), 379–414.
Finn, M.V., Kent, G.A., 1985. Simple Analytic Solutions to Complex Military Problems. Tech. rep., DTIC document.
Huth, P.K., Allee, T.L., 2002. The Democratic Peace and Territorial Conflict in the Twentieth Century. Cambridge University Press.
Jervis, R., 1976. Perception and Misperception in International Politics. Princeton University Press.
Jervis, R., 2014. The United States and Iran: perceptions and policy traps. In: Maleki, A., Tirman, J. (Eds.), U.S.–Iran Misperceptions. Bloomsbury Academic, New York, pp. 15–36.
Kamien, M.I., et al., 1990. On the value of information in a strategic conflict. Games Econ. Behav. 2 (2), 129–153.
Karr, A.F., 1981. Nationwide Defense against Nuclear Weapons: Properties of Prim-Read Deployments. Tech. rep., DTIC document.
Kirstein, R., 2014. Doping, the inspection game, and Bayesian enforcement. J. Sports Econ. 15 (4), 385–409.
Meirowitz, A., Sartori, A.E., 2008. Strategic uncertainty as a cause of war. Q. J. Polit. Sci. 3 (4), 327–352.
Miller, S., Bunn, M., 2014. Interpreting the implacable foe: American perceptions of Iran. In: Maleki, A., Tirman, J. (Eds.), U.S.–Iran Misperceptions. Bloomsbury Academic, New York, pp. 57–88.
Moon, C., Souva, M., 2016. Audience costs, information, and credible commitment problems. J. Conflict Resolution 60 (3), 434–458.
O'Neill, B., 1991. Why a Good Verification System Can Give Ambiguous Evidence. YCISS working paper.
O'Neill, B., 1993. Operations Research and Strategic Nuclear War. International Military Defense Encyclopedia. Pergammon-Brassey.
O'Neill, B., 1994. Game theory models of peace and war. In: Aumann, R.J., Hart, S. (Eds.), Handbook of Game Theory with Economic Applications, vol. 2. North–Holland, Amsterdam, pp. 995–1053.
Powell, R., 1996. Bargaining in the shadow of power. Games Econ. Behav. 15 (2), 255–289.
Powell, R., 2004. The inefficient use of power: costly conflict with complete information. Amer. Polit. Sci. Rev. 98 (02), 231–241.
Powell, R., 2006. War as a commitment problem. Int. Organ. 60 (01), 169–203.
Read Jr., W., 1958. Tactics and Deployment for Anti-Missile Defense. Bell Telephone Laboratories, Whippany, NJ.
Read Jr., W., 1961. Strategy for active defense. Amer. Econ. Rev., 465–471.
Schelling, T.C., 1960. The Strategy of Conflict. Harvard University Press.
Schultz, K.A., 1998. Domestic opposition and signaling in international crises. Amer. Polit. Sci. Rev. 92 (04), 829–844.
Schultz, K.A., 1999. Do democratic institutions constrain or inform? Contrasting two institutional perspectives on democracy and war. Int. Organ. 53 (02), 233–266.
Shubik, M., 1983. Game theory, the language of strategy. In: Shubik, M. (Ed.), Mathematics of Conflict. North-Holland, Amsterdam, pp. 1–28.
Shubik, M., 1985. The Uses, Value and Limitations of Game Theoretic Methods in Defense Analysis. Defense Technical Information Center.
Thomas, C., 1964. Some past applications of game theory to problems of the United States Air Force. In: Proceedings of a Conference Under the Aegis of the NATO Scientific Affairs Committee. Toulon, France, pp. 250–267.
Wittman, D., 1989. Arms control verification and other games involving imperfect detection. Amer. Polit. Sci. Rev. 83 (03), 923–945.