Balsgaard, Allan Anders

Doctoral Thesis

Arms races and war: A game theoretic analysis

PhD Series, No. 182

Provided in Cooperation with:
University of Copenhagen, Department of Economics

Suggested Citation: Balsgaard, Allan Anders (2016) : Arms races and war: A game theoretic analysis, PhD Series, No. 182, ISBN 978-87-93428-05-8, University of Copenhagen, Department of Economics, Copenhagen

This Version is available at:
https://hdl.handle.net/10419/233446

UNIVERSITY OF COPENHAGEN
FACULTY OF SOCIAL SCIENCES

Ph.D. thesis

Allan Anders Balsgaard

Arms Races and War

A game theoretic analysis

Supervisor: Thomas Jensen

Date of submission: 29 February 2016

Indhold

Introduction and summary

This thesis studies the behavior of states and the way they make some of their most important decisions, including the decision to acquire nuclear weapons or go to war. While these topics are not commonly dealt with in a Ph.D. thesis in economics, the methodology employed to analyzed them is. We use formal models and game theory to describe and analyze the behavior of states. This allows us to achieve a higher level of precision than is possible through verbal reasoning. The thesis consists of three self-contained parts all united under the common theme of arms races and war.

In the *first* part of the thesis, titled *The Spiral Model and the Shadow of the Future*, we study a repeated game with asymmetric information where two states decide whether to acquire a nuclear weapon or not in each period of the game. Players wish to avoid an arms race but are afraid that their adversary is an aggressive type that secretly tries to acquire nuclear weapons. Baliga & Sjöström (2004, 2012) show how *the fear of being left behind* in the arms race drives peaceful states into an arms race through an escalating cycle of pessimistic expectation. Using Baliga & Sjöström's (2004, 2012) static model of incomplete information as a stage game in a repeated game, we show that the destabilizing effect of the fear of being left behind is countered by a *fear of setting off an arms race*. If the adversary is a peaceful type with no intention of arming, acquiring a nuclear weapon could provoke an arms race that neither of the states wants. Thus, the fear of setting off an arms race represents a stabilizing force that discourages states from arming. We show that there is a perfect Bayesian equilibrium and provide the conditions for peaceful cooperation. In the equilibrium, players use a conditional trigger strategy similar to the grim-trigger strategy in the repeated prisoners' dilemma. Our dynamic arms race model is a merge of Baliga & Sjöström's (2004, 2012) static model and the repeated prisoners' dilemma. It has interesting applications and provides some surprising results. *First* of all, our model provides an explanation for the remarkable absence of widespread nuclear proliferation in the post-war period. *Secondly*, through comparative statics, we show that increasing the share of pacifist states makes peaceful cooperation less likely.

The *second* part of the thesis, titled *Military Alliances and the Power Curse*, studies the strategic interaction between a weak state and its stronger ally and the way that they respond to a threat to their security. Both states prefer a coordinated response to the threat, but mutual mistrust makes it difficult for the weaker state to rely on the stronger ally for protection. Therefore, the weaker state may decide to engage the threat unilaterally. Even though the stronger state prefers to delay any action, it may decide to engage the threat preemptively in order to forestall the adverse effects of an uncoordinated attack. We use a simple dynamic model with asymmetric information to represent these strategic dilemmas. Solving the model for perfect Bayesian equilibria, we show when a coordinated response is possible and when the weaker ally manages to force the stronger ally to take military action. Thus, our model provides a theoretical mechanism for the *power curse*. The power curse is a stylized fact in international relations according to which great powers tend to get involved in military operations against their will and therefore can seem powerless. Among other things, this model can be used to describe the strategic contradictions between the U.S. and Israel over Iran's nuclear program.

The *third* part of the thesis, titled *Does Better Information Make War Less Likely?*, has been written together with Thomas Jensen. It examines whether improved information lowers the probability of war. Researchers (Fearon, 1995) have long argued that asymmetric information is among the fundamental explanations for war. The inability of states to accurately assess whether their adversary will give in to a given demand may lead to war if the demand is too high. Under complete information, states' assessments of their adversaries are fully accurate and states therefore do not make unreasonable demands. Thus, complete information makes it possible to achieve a peacefully settled outcome. Given that asymmetric information leads to war and complete information eliminates war one may presume that reducing asymmetric information will decrease the probability of war. We show that this presumption is wrong. We employ a 3 type ultimatum bargaining model where state B's willingness to go to war is unknown to state A and where state A receives a noisy signal about state B's type. We show that improving the quality of the signal does not make war less likely. Though complete information

implies peace, better information does not necessarily decrease the probability of war. Sometimes, better information can make war more likely.

Resume på Dansk

Denne afhandling handler om staters adfærd og hvordan de træffer nogen af deres allervigtigste beslutninger som at anskaffe sig kernevåben eller gå i krig. Selvom disse temaer ikke er noget man typisk beskæftiger sig med i en Ph.d. afhandlinger i økonomi, er de analyseredskaber der anvendes i afhandlingen velkendte af økonomer. Vi anvender formelle matematiske modeller og spilteori til at beskrive og analysere staters adfærd med. Disse metoder gør det muligt at opnå en højere grad af præcision end det er muligt at opnå gennem verbale ræsonnementer.

Afhandlingens første del, der bærer titlen *The Spiral Model and the Shadow of the Future*, indeholder et gentaget spil med asymmetrisk information, hvor to stater skal beslutte, hvorvidt de skal anskaffe sig kernevåben i hvert af spillets perioder eller ej. Spillerne ønsker at undgå et våbenkapløb, men frygter samtidigt at deres modstander er en aggressive type, der i hemmelighed forsøger at anskaffe sig kernevåben. Baliga & Sjöström (2004, 2012) viser at *frygten for at kommer bagud* i våbenkapløbet kan føre til, at fredelige stater ender med at opruste gennem en eskalerede cyklus af pessimistiske forventninger. Vi bruger Baliga & Sjöströms (2004, 2012) statiske model med asymmetrisk information som grundspil i et gentaget spil og viser at den destabiliserende effekt af frygten for at kommer bagud modvirkes af *frygten for at sætte et våbenkapløb i gang*. Hvis modstanderen er fredelig og ikke har noget ønske om at opruste, kan anskaffelse af kernevåben fremprovokere et våbenkapløb, som ingen af parterne ønsker sig. Denne frygt for at sætte et våbenkapløb i gang udgør en stabiliserende faktor, der bidrager til at afholde stater fra at opruste. Vi viser, at der findes en perfekt Bayesiansk ligevægt og udleder betingelserne for at samarbejde mellem staterne er mulig. I ligevægten bruger spillerne en betinget aftrækker strategi, som ligner aftrækker strategien fra et gentaget fangernes dilemma. Modellen har interessante anvendelser og giver os samtidigt nogle overraskende resultater. For det *første* giver modellen os en forklaring på det bemærkelsesværdige fraværd af spredning af kernevåben i efterkrigstiden. For det *andet* viser vi vha. komparativ statik, at en stigning i andelen af pacifistiske stater gør sandsynligheden for et våbenkapløb større.

Afhandlingens anden del, der bærer titlen *Military Alliances and the Power*

Curse, handler om de strategiske interaktioner mellem en svag stat og dens stærke allierede og om, hvordan staterne reagerer på en trussel mod deres fælles sikkerhed. Staterne foretrækker, at truslen imødegås med et koordineret svar, men dette vanskeliggøres af gensidig mistillid, som gør den svage stat modvillig mod at lægge ansvaret for sin egen sikkerhed i hænderne på den stærke stat. Det skaber risiko for, at den svage stat vælger at imødegå truslen på egen hånd. Til trods for at den stærke stat foretrækker at imødegå truslen på et senere tidspunkt, kan det blive nødvendigt for den at handle hurtigt og i utide for at komme den svage stat i forkøbet og derved undgå de negative følger af et ukoordineret svar. Vi bruger en simpel dynamisk model med asymmetrisk information til at repræsentere dette strategiske dilemma med. Vi løser modellen for perfekte Bayesianske ligevægte og viser hvornår et koordineret svar er muligt og hvornår den svage stat kan tvinge sin stærke allierede til at gennemfører en utidig militær aktion. Vores model giver således en teoretisk funderet forklaring på *magtens forbandelse*. Magtens forbandelse er den stiliserede kendsgerning (*stylized fact*) i internationale relationer, at stormagter ofte bliver inddraget i militære konflikter mod deres vilje og derved kommer til at fremstå magtesløse. Denne model kan bl.a. bruges til at beskrive de strategiske modsætninger mellem USA og Israel i forbindelse med krisen om Irans atomprogram.

Afhandlingens tredje del med titlen *Does Better Information Make War Less Likely?* er skrevet sammen Thomas Jensen og undersøger om bedre information mindsker sandsynligheden for krig. Asymmetrisk information er længe blevet anset for at udgøre en grundlæggende forklaringer på krig (Fearon, 1995). Staters vanskelighed ved at foretage præcise vurderinger af modstanderens kampvilje kan føre til krig, hvis der stilles for høje krav til modstanderen. Under fuldstændig information er staters vurdering af deres modstanderes kampvilje altid korrekt, hvilket forhindre staterne i at fremsætte urimelige krav. Dette gør det lettere at nå til enighed gennem forhandlinger. Eftersom asymmetrisk information kan føre krig, mens fuldstændig information omvendt umuliggør krig, virker det rimeligt at antage at sandsynligheden for krig mindskes, når graden af asymmetrisk information aftager. Denne formodning viser sig dog at være forkert. Vi anvender et 3 typers

ultimatumspil, hvor stat A giver et tilbud til stat B, som stat B kan acceptere eller afslå. Stat B's kampvilje er ukendt, men stat A modtager et støjfyldt signal om stat B's kampvilje. Vi viser at sandsynligheden for krig ikke bliver mindre af at signalets kvalitet forbedres. Til trods for at fuldstændig information medfører fred, vil bedre information ikke nødvendigvis gøre krig mindre sandsynligt. I visse tilfælde kan bedre information gøre krig mere sandsynligt.

Acknowledgments

This Ph.D. thesis was written as part of the Ph.D. program at the Department of Economics, University of Copenhagen, 2013-2016.

I would like to express my gratitude to a number of people for their support and help during the time when I wrote this thesis. First and foremost, I would like to thank my supervisor Thomas Jensen for his persistent support and willingness to read earlier drafts. I am also grateful to Alexander Sebald, Edward Webb, Peter Norman Sørensen, Ole Jann and the participants of the 2015 APSA conference for their comments and suggestions.

During my Ph.D. studies, I had the opportunity to benefit from the inspiring research environment of Columbia University and Kellogg School of Management during a trip to the U.S. I owe a debt of gratitude to the Denmark-America Foundation, Augustinus Fonden, Knud Højgaards Fond for their financial support. The trip was a valuable experience and I met a lot of interesting people and researchers, including Tomas Sjöström and Massimo Morelli, who hosted me. I am also grateful to Christian Schultz for helping with organizing the trip and for creating a pleasant atmosphere for research at the Department of Economics.

Finally, I would like to thank my wife and two, often-neglected boys Viktor and Leon for their patience during my long days at the office.

Allan Anders Balsgaard
Copenhagen, 29 February 2016

The Spiral Model and the Shadow of the Future

Allan Anders Balsgaard*

February 2016

Abstract

We analyze a dynamic arms race model where states decide whether to acquire arms or cooperate in each period. States prefer to cooperate when the adversary cooperates but fear that the adversary is an aggressive type who prefers to arm. Uncertainty about intentions gives rise to a *fear of being left behind* in the arms race. Baliga and Sjöström (2004, 2012) show that this fear can lead to an arms race through a negative spiral of mutually reinforcing, pessimistic expectations. We show how the destabilizing effect of the fear of being left behind is countered by a *fear of setting off an arms race*. Our model provides a competing explanation for the remarkable absence of widespread nuclear proliferation during the post-war period and offers a warning against adopting excessively pacifistic foreign policy doctrines.

*Email: allanbalsgaard@gmail.com

Introduction

In a nuclear arms race, three strategic considerations are of great importance when states decide whether to acquire a nuclear weapon:

1. *The fear of being left behind* or the preemptive motive. If the adversary is an aggressive type, who attempts to gain an advantage by acquiring nuclear weapons unilaterally, abstaining from nuclear weapons could be dangerous. Being without nuclear weapons in a nuclear world may lead to worse bargaining outcomes due to lack of deterrence.

2. *The predatory motive.* Building nuclear weapons provides aggressive states with an opportunity to gain a temporary (or permanent) advantage over other states through enhanced coercive capabilities.

3. *The fear of setting off an arms race.* The adversary may not have any plans to acquire a nuclear weapon. Thus, building a nuclear weapon may be counterproductive and inadvertently provoke a nuclear arms race if acquisition of a weapon is reciprocated by the adversary.

The first and second strategic considerations are part of the underlying structure in the spiral model (Schelling, 1960; Jervis, 1976, 1978). They are also key elements in Baliga and Sjöström's (2004, 2012) formalization of the spiral model. From now on we will refer to Baliga and Sjöström's model (2004, 2012) as SARM (Static Arms Race Model). Using a static model with asymmetric information, they show how the fear of facing an aggressive adversary can spur an arms race among peaceful states through an escalating cycle of pessimistic expectations.

The third strategic consideration is a key element in repeated games where the threat of future punishment is used to sustain cooperation. States decide to cooperate rather than acquire a nuclear weapon in order to avoid a nuclear arms race which would make them both worse off. In repeated games, the *Folk Theorems* ensure that players do not defect (Friedman 1971; Axelrod, 1984).

The spiral model and repeated games of cooperation represent two distinct ways of modeling an arms race and each approach captures some relevant features

hereof. However, to our knowledge, no one has merged these approaches into a single framework. We propose a unified framework, which combines the SARM with a repeated game by having the SARM be the stage game in a repeated game.

The SARM is a coordination game, in which states only wish to arm if the adversary arms. The wish of both players to stay out of an arms race if the adversary stays out would seem to make it easy for them to reach a peaceful equilibrium. However, uncertainty about the intentions of the adversary makes it difficult to obtain peace. If there is a small probability that the adversary is an aggressive type that prefers to arm unilaterally, then the fear of facing an aggressive type may lead to an arms race through a negative spiral of mutually reinforcing, pessimistic expectations. Thus, the first and second strategic considerations together are destabilizing and may lead to an arms race between peaceful states.

Repeating the game forces players to take the future into account when they make the decision of whether or not to arm. In this way, a stabilizing element is added, which provides a counterweight to the destabilizing effect of the preemptive and predatory motive.

The first and second strategic considerations alone imply a rather pessimistic view of international relations, in which the fear of being left behind plays a major role and peaceful states are forced to arm. The third strategic consideration, in contrast, implies a much more optimistic view of international relations, in which even aggressive states are willing to cooperate. Thus, when the fear of setting off an arms race is added to the theoretical framework, a higher degree of stability is achieved. Despite the potential gains from arming unilaterally, states abstain from building nuclear weapons fearing that such actions may provoke the adversary to do the same.

Combining the three strategic considerations in a single theoretical framework creates a framework with a high degree of explanatory power. A theoretical framework that only relies on the first and second strategic considerations predicts a rather high level of instability and occurrences of arms races whereas the historical record of nuclear proliferation shows the opposite. After World War II, only a limited number of countries acquired nuclear weapons. Far fewer than previously

predicted. By including the fear of setting of an arms race, states become more cautious and less willing to acquire nuclear weapons. Thus, our unified framework provides a competing explanation for the absence of nuclear weapons in the post war period.

Our unified approach yields some interesting insights that are not revealed through partial analysis. In contrast to the SARM, we show that increasing the proportion of pacifist states (dominant strategy doves) *decreases* the likelihood of cooperation. In the SARM where only the first and second considerations are taken into account, increasing the share of pacifist states leads to more cooperation. When the share of states that abstain from nuclear weapons increases, the level of mutual distrust decreases and so does decreases the likelihood of an arms race. However, when the *fear of setting off an arms race* is included in the theoretical framework, increasing the share of pacifist states makes it more attractive for aggressive states to acquire nuclear weapons. When the likelihood of facing a pacifist state increases, aggressive actions are less likely to trigger punishment. Aggressive actions are more likely to pay off and nuclear proliferation gets more likely. In other words, increasing the share of pacifist states decreases the fear of setting off an arms race and therefore leads to more nuclear proliferation. Therefore, our model warns of adopting foreign policy doctrines that exclude the possibility of acquiring nuclear weapons regardless of the choices made by other states.

The following section contains a review of related literature. Next, the SARM is presented. Subsequently, we present our dynamic model, identify an equilibrium and use comparative statics to show how pacifism can make an arms race more likely. Finally, we discuss the empirical implications of our model and its ability to explain the empirical record of proliferation of nuclear weapons.

Literature

Our model is related to at least two different strands of literature. *Firstly*, as discussed above, our model is related to the literature on repeated games where cooperation is sustained through the possibility of punishment in later stages of the

game. Analysis of these types of games is due to Friedman (1971), Axelrod (1984), Downs, Rocke and Siverson (1986) and others. Using *Folk theorems*, cooperation can be maintained through the prospect of setting off an arms race. The topic is developed by Calvert (1995) and Fearon (1998). Like these models, our model is also an infinitely repeated game. The major difference is that in our model, there is two-sided asymmetric information, which adds more realism but also provides additional computational challenges.

Secondly, our model is a spiral model in which mutual fears play a central role. Uncertainty about the intentions of the adversary is a central feature of these models. Originally due to Schelling (1960) and Jervis (1976, 1978), the spiral models describe how the fear of an attack induces states to take defensive actions, which may be misperceived by the adversary as offensive actions. The adversary responds in kind, hence creating a negative spiral of armed escalation that may lead to war. Thus, mutual fear of becoming a victim of aggression may lead to an inefficient outcome such as an arms race or war. The global games literature by Carlsson and van Damme (1993) and Morris and Shin (2003) constitutes yet another theoretical contribution to our understanding of spiral models. Morris and Shin (2003) show how the multiplicity of equilibria in a coordination game can be eliminated through a common signal. States are uncertain about the intentions of their adversary, but are able to coordinate on a single equilibrium because of a common signal. Using independent types, Baliga and Sjöström (2004, 2012) show how even a small risk of encountering an aggressive adversary can cause an arms race through an escalating cycle of mutually pessimistic expectations.

Our model draws on both of these literatures as we use the SARM as a stage game in an infinitely repeated game.

Our model also shares common traits with Chassang and Padro i Miquel (2010) and Kydd (1997). Chassang and Padro i Miquel (2010) use an infinitely repeated global game where states accumulate arms and decide whether or not to attack. They demonstrate that parity of military power is destabilizing. Fear of being targeted by a first strike inadvertently leads to war through a negative spiral of pessimistic expectations. On the other hand, when there is disparity in military

power, none of the states can get a decisive advantage through a first strike. Therefore, a destabilizing spiral of fear does not take place.

Kydd (1997) builds a simple three period spiral model in which states decide whether to arm and to attack. Whether the adversary arms or not signals what type he is - a *security seeker* or *greedy* (i.e. a potential attacker) - and therefore influences the decision of whether to attack or not. Our model is more general in that we allow for a continuum of types. However, the strategic environment in our model is simpler than in both Chassang and Padro i Miquel (2010) and Kydd (1997). In our model, states only decide whether or not to arm.

A series of recent models deals more specifically with nuclear proliferation. In these models, an established nuclear power tries to prevent a potential proliferant from acquiring nuclear weapons. Debs and Monteiro (2014) show that the cost of preventive war and the potential change in the balance of power are decisive factors for determining whether nuclear proliferation takes place. If the cost of preventive war is sufficiently low and the effect on the balance of power is sufficiently large, then the proliferant is deterred from building nuclear weapons. Furthermore, uncertainty about the nature of a nuclear program sometimes leads to preventive wars against states who do not intend to acquire nuclear weapons.

Bas and Coe (2015) adds additional realism to this framework by assuming that the acquisition process is subject to stochastic shocks. Thus, elements of chance in the progress of the nuclear program become crucial for explaining whether proliferation takes place. Spaniel (2015) explains how established nuclear powers use bargaining to buy off potential proliferants.

Whereas these models are well suited for explaining the relationship between the US and a potential proliferant (e.g. Iran), they are ill-suited for describing an arms race between peer competitors. In our model, there is complete symmetry between the states. Neither of the states has an inherent advantage or possesses a nuclear weapon from the onset. Our model is better at explaining nuclear proliferation as a phenomenon related to peer competition, than as one related to the subordination of a weaker state by a hegemon. Thus, our model is better suited for explaining the relationship between Iran and Saudi Arabia or the US and the

Soviet Union in the immediate aftermath of WWII.

Arms Race Models

Our model is an infinitely repeated game where the stage game is identical to the SARM. In each period, players simultaneously decide whether to build a nuclear weapon (B) or not (N). There is uncertainty about the adversary's type.

In the SARM, an arms race is less likely when the share of pacifistic states increases. In the dynamic arms race model where the game is repeated infinitely many times, the opposite holds, i.e. we show that pacifist states are detrimental for cooperation.

The Static Arms Race Model

Figure 1 illustrates the stage game.

		State 2	
		B	N
State 1	B	$-c_1, -c_2$	$\mu - c_1, -d$
	N	$-d, \mu - c_2$	$0, 0$

Figure 1: **Stage game**

c_1 is the cost of building nuclear weapons for state 1. μ is the advantage of acquiring nuclear weapons while the adversary remains unarmed. d is the disadvantage of being left behind in the arms race.

If both players choose N, an arms races is avoided and both players receive the normalized payoff 0. If player 1 chooses B and player 2 chooses N, player 1 arms unilaterally, pays the cost of arming $-c_1$ and receives the gain μ. Player 2 suffers the loss of being left behind $-d$. If both players choose B, there is an arms race and players suffer $(-c_1, -c_2)$ and neither of them gains any advantage.

c_i is a stochastic variable with distribution function $F(c_i)$ where $\text{supp}\,(F) = [0,1]$, $F(0) = 0$, $F'(c) > 0$ for $i \in [0,1]$ and $F(1) = 1$. c_i is private information and only known to player i where $i \in \{1,2\}$.

Assumption 1. $0 < \mu < d < 1$

According to assumption 1, the disadvantage of being left behind d is larger than the gain of getting ahead in the arms race μ. Thus, state 1 does not gain what state 2 loses if state 2 is left behind in the arms race[1]. Everything - except the true values of c_1 and c_2 - is common knowledge. We can categorize players in 3 different ways:

$c_i \in [0, \mu]$: Dominant strategy hawks (aggressive players). These players prefer mutual cooperation (N, N) over an arms race (B, B), but would like to acquire nuclear weapons unilaterally if they can get away with it. When both types are in this interval, the payoff-structure of the game is identical to that of the prisoners' dilemma.

$c_i \in]\mu, d]$: Coordinating types. These players also prefer the cooperative outcome (N, N), but unlike the aggressive players, they have no gain of defecting from the cooperative outcome unilaterally. Only if the adversary is certain to defect does it pay for this type of player to choose B.

$c_i \in]d, 1]$: Dominant strategy doves (pacifist players). These players choose N regardless of the adversary's strategy.

Coordinating types that are close to μ are almost dominant strategy hawks. For these types the gain of cooperation is only marginally better than arming unilaterally. Therefore, these types only choose N if they are certain that the adversary

[1]Note that the support of the distribution function F is limited to the unit-interval in this version of the model. In Baliga and Sjöström (2004, 2012) a more general interval $[\underline{c}, \overline{c}]$ is used. Moreover, we limit ourselves to the case where $\mu < d$ whereas Baliga and Sjöström (2004, 2012) also consider $d < \mu$. When $\mu < d$, strategies are strategic complements, meaning that player 1 is more likely to choose B, the more likely player 2 is to choose B. In contrast, when $\mu < d$ strategies are strategic substitutes and player 1 is less likely to choose B, the more likely player 2 is to choose B.

also chooses N. Similarly, coordinating types that are very close to d are almost dominant strategy doves. For these types, the fear of being left behind is only marginally worse than an arms race (B, B). Therefore, these types only choose B if they are certain that the adversary also chooses B.

In the equilibrium, players use a cutoff strategy, which consists of a cutoff point c^* such that B is chosen if and only if $c_i \leq c^*$.

Proposition 1 (Existence and uniqueness). *There is a Bayesian Nash equilibrium with cutoff point $c^* \in (\mu, d)$. The equilibrium is unique if $F'(c) < \frac{1}{d-\mu}$ for all $c \in (\mu, d)$.*

Proof. Suppose that player 2 chooses B with probability p_2. Player 1's expected utility from choosing B is $-p_2 c_1 + (1 - p_2)(\mu - c_1)$. If player 1 chooses N, his expected utility is $-p_2 d$. Player 1's net gain from choosing B is

$$-c_1 + \mu + (d - \mu)p_2$$

Player 1 chooses B if the net gain is positive and N if the net gain is negative. Player 1 is indifferent between B and N if the net gain is zero. However, for convenience, we assume that B is chosen when the net gain equals zero.

The monotonicity of the net gain in p_2 implies that all BNE must be in cutoff strategies. Notice that players' strategies are functions $[0, 1] \to \{B, N\}$. But since players only use cutoff strategies, each strategy that players use can be identified through its cutoff point. If player 2's cutoff point is $\tilde{c}_2$, player 2's probability of choosing B is $p_2 = F(\tilde{c}_2)$ and player 1's best response is to use the cutoff point $\tilde{c}_1 = \Gamma(\tilde{c}_2)$ where

$$\Gamma(c) = \mu + (d - \mu)F(c)$$

$\Gamma(c)$ is the best response function for cutoff strategies. Since dominant strategy hawks always choose B and dominant strategy doves always choose N, the cutoff point $c^* \in [\mu, d] \subsetneq [0, 1]$ by assumption 1. Thus, B and N are chosen with a strictly positive probability in any BNE.

The continuity of $\Gamma(c)$, $\Gamma(\mu) = \mu + (d - \mu)F(\mu) > \mu$ and $\Gamma(d) = dF(d) + \mu(1 - F(d)) < d$ ensure that a fixed-point $c^* \in (\mu, d)$ exists. Using symmetry, we infer that there is a BNE where both players use the cutoff point c^*.

Next, we prove uniqueness. The derivative of $\Gamma(c)$ is $\Gamma'(c) = (d-\mu)F'(c)$. Now we see that $F'(c) < \frac{1}{d-\mu}$ implies $\Gamma'(c) = (d-\mu)F'(c) < 1$. Moreover, $\mu < d$ (from Assumption 1) implies that $\Gamma'(c) = (d-\mu)F'(c) > 0$. Therefore, $0 < \Gamma'(c) < 1$, which is a well-known condition for uniqueness. Hence, c^* is the unique cutoff point. □

Note that the assumption $F'(c) < \frac{1}{d-\mu}$ ensures that the probability is not too concentrated in certain regions of the interval $c \in (\mu, d)$. If this condition is not satisfied, the best response function will become too steep in certain regions, which may lead to multiple cutoff points. The increasing best response functions of both players and the cutoff point have been illustrated in figure 2.

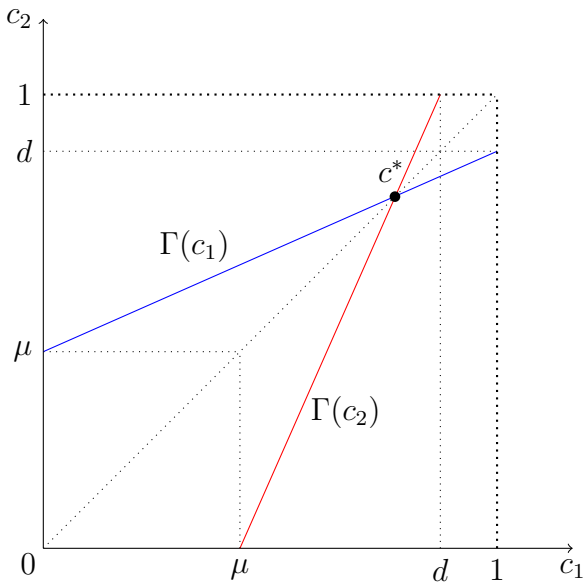

Figure 2: Best response functions and fixed point when types are uniformly distributed, i.e. $F(x) = 1$.

To demonstrate how the fear of being left behind can compel coordinating types to arm through a negative spiral of pessimistic expectations, suppose that player 1 uses a cutoff strategy with cutoff point $\tilde{c}_1^1 = \mu$ such that only dominant strategy hawks arm. Thus, the probability that player 1 arms is $F(\tilde{c}_1^1) = F(\mu)$. Since $\tilde{c}_1^1$ is below the cutoff point c^*, player 2's best response is to use a *higher* cutoff point, i.e.

$\tilde{c}_2^2 = \Gamma(\tilde{c}_1^1) > \tilde{c}_1^1$. Next, if player 2 arms with probability $F(\tilde{c}_2^2)$, the fear of setting off an arms race is the dominant concern again. Thus, player 1's best response is to use a higher cutoff point, i.e. $\tilde{c}_1^3 = \Gamma(\tilde{c}_2^2) > \tilde{c}_2^2$. Hence, the equilibrium can be reached through iterated deletion of dominated strategies. During this cascade, coordinating types who are almost dominant strategy hawks choose to arm, which in turn causes "almost-almost dominant strategy hawks" to arm, etc. Continuing this way, the cutoff points eventually converge up towards c^*.

Comparative Statics

In the following we derive some comparative statics results that are not included in Baliga and Sjöström (2012). We do this first and foremost in order to obtain a reference point for results derived later in the section on the dynamic arms race model. The first result shows that dominant strategy hawks are destabilizing and the second that dominant strategy doves are stabilizing.

Destabilizing Hawks

Increasing the share of dominant strategy hawks makes cooperation less likely. The share of dominant strategy hawks is μ. Thus, increasing μ makes cooperation less likely.

Proposition 2. *Increasing the proportion of dominant strategy hawks leads to lower levels of cooperation if* $F'(c) < \frac{1}{d-\mu}$ *for all* $c \in (\mu, d)$.

$$\frac{\partial c^*}{\partial \mu} > 0$$

Proof. Using implicit differentiation, we find the partial derivative of $\Gamma(c)$ with respect to μ.

$$\frac{\partial c^*}{\partial \mu} = 1 - F(c^*) + (d-\mu)F'(c^*)\frac{\partial c^*}{\partial \mu}$$

Rearranging,

$$\frac{\partial c^*}{\partial \mu} = \frac{1 - F(c^*)}{1 - (d-\mu)F'(c^*)}$$

The numerator is strictly positive since $c^* \in (\mu, d)$, $F(1) = 1$ and $F'(c) > 0$ for all $c \in [0, 1]$. Moreover, assumption $F'(c) < \frac{1}{d-\mu}$ implies that the denominator is strictly positive. Hence, $\frac{\partial c^*}{\partial \mu} > 0$. □

When μ is increased, the share of dominant strategy hawks goes up at the expense of the share of coordinating types $(d - \mu)$. Therefore, the probability of encountering an aggressive player increases, which makes it more risky for coordinating types to cooperate. Fear of being left behind increases and the level cooperation therefore decreases (c^* increases).

Stabilizing Doves

Increasing the share of dominant strategy doves makes cooperation more likely. The share of dominant strategy doves is $1 - d$. Thus, the share of dominant strategy doves is large if d is low.

Proposition 3. *Increasing the proportion of dominant strategy doves leads to less conflict if* $F'(c) < \frac{1}{d-\mu}$ *for all* $c \in (\mu, d)$.

$$\frac{\partial c^*}{\partial d} > 0$$

Proof. Using implicit differentiation, we find the partial derivative of $\Gamma(c)$ with respect to μ.

$$\frac{\partial c^*}{\partial d} = F(c^*) + (d - \mu)F'(c^*)\frac{\partial c^*}{\partial d}$$

Rearranging and using assumption 1,

$$\frac{\partial c^*}{\partial d} = \frac{F(c^*)}{1 - (d - \mu)F'(c^*)}$$

The numerator is strictly positive since $c^* \in (\mu, d)$, $F(0) = 0$ and $F'(c) > 0$ for all $c \in [0, 1]$. Moreover, assumption $F'(c) < \frac{1}{d-\mu}$ implies that the denominator is strictly positive. Hence, $\frac{\partial c^*}{\partial d} > 0$. □

Increasing the share of dominant strategy doves at the expense of coordinating types leads to more cooperation, i.e. a lower d leads to fewer types choosing B.

Increasing the share of pacifist players that abstain from arming – regardless of the action taken by the adversary – inhibits the spiral of fear by reducing the proportion of players that potentially could arm. As more players shift towards the pacifistic category, cooperation increases. Note that this result is only true in the static arms race model. In the repeated game, which we consider in the next section, the result is reversed.

Uniform distribution

Here we consider the SARM in the special case where c_1 and c_2 are uniformly distributed on $[0, 1]$. Analyzing the SARM for the uniform distribution is a useful precursor to studying the dynamic model where the issue of tractability restricts us from using a general distribution function on $[0, 1]$. The uniform distribution is $F(c) = c$ and the best response function is

$$\Gamma(c) = \mu + (d - \mu)c$$

The fixed-point c^* where $\Gamma(c^*) = c^*$ becomes

$$c^* = \frac{\mu}{1 - (d - \mu)}$$

It follows straightforwardly that $\frac{\partial c^*}{\partial \mu} = \frac{1-d}{(1-(d-\mu))^2} > 0$ and $\frac{\partial c^*}{\partial d} = \frac{\mu}{(1-(d-\mu))^2} > 0$.

A Dynamic Arms Race Model

In the infinitely repeated game, players decide between B and N in each period in the same way as in SARM. Hence, the fear of being left behind still guides players' behavior. The major difference from the static model is that players are now forced to take the future impact of their decisions into account in each period. Hence, aggressive players need to consider the possibility that defection may lead to a breakdown of cooperation. Notice that unlike the repeated prisoner's dilemma, our model incorporates uncertainty. Therefore, a player will not be punished for opportunistic behavior if the adversary turns out to be a dominant strategy dove.

Payoffs are discounted in each period t with a discount factor δ.

Assumption 2. $\frac{1}{2} < \delta < 1$

Types are uniformly distributed on the unit-interval, $F(c) = c$ for all $c \in [0, 1]$. Everything – except the true value of c_1 and c_2 – is common knowledge.

In the following, we argue that there is a perfect Bayesian equilibrium in which players use a *conditional trigger strategy*. Players cooperate as long as the adversary cooperates. If either of the players defects by choosing B in any of the periods, a punishment phase is initiated in which both players choose B for the remainder of the game if they are not dominant strategy doves.

In the equilibrium, players use cutoff strategies with a cutoff point c^* in period 1 so that player 1 chooses B if $c_1 \leq c^*$ and N otherwise. As we show below, assumption 2 ensures that $c^* < \mu$. Hence, only dominant strategy hawks choose B in period 1.

We say that (B, B) is chosen if both players choose B. Similarly, we say that (N, N) is chosen if both players choose N. Finally, we say that (B, N) is chosen if player 1 chooses B and player 2 chooses N. A perfect Bayesian equilibrium consists of a strategy profile and a belief system where the strategy profile is sequentially rational and the belief system is consistent. Also, the restriction of the strategy profile and belief system in every continuation game must be a Bayesian Nash equilibrium. Hence, we need to define the beliefs and strategies in a large number of continuation games. This makes the proof a particularly comprehensive undertaking due to the large number of continuation games. The parts of the proof that are related to these continuation games have therefore been moved to the appendix.

Proposition 4 (Conditional trigger strategy).
There is a PBE where the following strategies are used[2]*:*

Period 1*: Players use cutoff strategies with the cutoff point*

$$c^* = \frac{1-(1-\delta)(d-\mu)-\sqrt{(1-(1-\delta)(d-\mu))^2-4\delta(1-\delta d)\mu}}{2\delta}$$

and $c^* \in (0, \mu)$*.*

Periods $t > 1$*:*

(i) If (N, N) *was chosen in all previous periods* $1, \ldots, t-1$*, players also choose* N *in all subsequent periods* $t, t+1, \ldots$*.*

(ii) If anything else was chosen in any of the previous periods $1, \ldots, t-1$*, player 1 chooses* B *in all subsequent periods* $t, t+1, \ldots$ *if* $c_1 \leq d$ *and* N *if* $c_1 > d$*. Player 2 uses the same strategy.*

Proof. In the following, we show that c^* is the optimal cutoff point for both players in period 1 given that the conditional trigger strategies are used in the subsequent periods. In the appendix, we show that the conditional trigger strategy is the optimal strategy in the continuation games in period $t > 1$.

Period 1 Players use a cutoff strategy with a cutoff point c^*. c^* is the type that is indifferent between choosing B and N in period 1. We consider the strategic trade-off of player 1 with a type $c_i < \mu$. Since $c^* \in [0, \mu]$, it is only relevant to consider the strategic dilemma of a dominant strategy hawk. In equilibrium, coordinating types always cooperate and their trade-off is therefore irrelevant to consider. Later, we verify that $c^* < \mu$.

Suppose that player 2 uses a cutoff strategy with cutoff point $c_2^* < \mu$. Player 1's expected payoff from choosing B in period 1 is

$$c_2^*\left(\sum_{t=0}^{\infty}\delta^t(-c_1)\right) + \left(d - c_2^*\right)\left(\mu + \sum_{t=0}^{\infty}\delta^t(-c_1)\right) + \left(1-d\right)\left(\sum_{t=0}^{\infty}\delta^t(\mu - c_1)\right) \quad (1)$$

[2]Strictly speaking, only the equilibrium strategies are defined here. More about this in the proof.

where the probability that player 2 chooses B is $P(c_2 < c_2^*) = F(c_2^*) = c_2^*$. We interpret the expression:

- The *1. part* is player 1's payoff if player 2 also chooses B in period 1. Nobody cooperates in period 1 and B is chosen by both players in all subsequent periods.

- The *2. part* is player 1's payoff if player 2 chooses N in period 1 and player 2 is not a dominant strategy dove. Cooperation also fails and B is chosen by both players in all subsequent periods. $d - c_2^*$ is the probability that $c_2 \in [c_2^*, d]$.

- The *3. part* is player 1's payoff if player 2 chooses N in period 1 and player 2 is a dominant strategy dove. In this case, cooperation also fails and B is chosen by player 1, but player 2 is unwilling to retaliate and (B, N) is therefore chosen in all subsequent periods. $1 - d$ is the probability that player 1 is a dominant strategy dove.

Player 1's expected payoff from choosing N in period 1 is

$$c_2^* \left(-d + \sum_{t=1}^{\infty} \delta^t (-c_1) \right) + (1 - c_2^*) \left(\sum_{t=0}^{\infty} \delta^t \cdot 0 \right) \tag{2}$$

- The *1. part* is player 1's payoff if player 2 chooses B in period 1. Player 1 receives the payoff $-d$ in period 1 for being left behind. In the rest of the game $t > 1$, cooperation fails and B is chosen by both players.

- The *2. part* represents the cooperative outcome. Both players choose N in period 1 and in the remaining periods $t > 1$ as well.

Subtracting (2) from (1) and rearranging, player 1's net gain from choosing B is

$$\frac{1}{1-\delta}\mu\big(1 - \delta d\big) + (d - \mu)c_2^* - \frac{1}{1-\delta}\big(1 - \delta c_2^*\big)c_1$$

where the formula $\sum_{t=0}^{\infty} \delta^t = \frac{1}{1-\delta}$ has been used.

If the net gain is positive, choosing B is optimal for player 1. If the net gain is negative, choosing N is the optimal strategy for player 1. If the net gain is zero, player 1 is indifferent between choosing B and N. However, for convenience we assume that player 1 chooses B.

Observe that the net gain is decreasing in c_1 due to Assumption 2. Hence, it is the best response for player 1 to use a cutoff strategy when player 2 uses a cutoff strategy. Setting the net gain equal to zero, we find the optimal cutoff point $\tilde{c}_1$ for player 1. Thus, if player 2's cutoff point is $\tilde{c}_2$, player 1's best response is to use a cutoff strategy with cutoff point $\tilde{c}_1 = \Gamma(\tilde{c}_2)$ where

$$\Gamma(c) = \frac{\mu\big(1-\delta d\big) + (1-\delta)(d-\mu)c}{1-\delta c}$$

$\Gamma(c)$ is the best response function for cutoff strategies in period 1.

Now we show that $c^* \in (0,\mu)$. First, note that $\Gamma(0) = \mu(1-\delta d) > 0$. Moreover, $\Gamma(\mu) < \mu$ is implied by $\frac{1}{2} < \delta$ (Assumption 2). Finally, since the best response function is continuous, a fixed point $c^* \in (0,\mu)$ exists. Due to symmetry, there is a PBE where c^* is used as a cutoff point by both players. Hence, we know that there is *at least one* cutoff point in the interval $(0,\mu)$.

Next, we derive the closed-form expression of c^*. Setting $\Gamma(c) = c$ and rearranging, we derive the fixed points.

$$\delta c^2 - (1-(1-\delta)(d-\mu))\,c + (1-\delta d)\mu = 0$$

Solving this 2. order equation, we find two fixed points.

$$c^* = \frac{1-(1-\delta)(d-\mu) - \sqrt{(1-(1-\delta)(d-\mu))^2 - 4\delta(1-\delta d)\mu}}{2\delta}$$

$$\check{c} = \frac{1-(1-\delta)(d-\mu) + \sqrt{(1-(1-\delta)(d-\mu))^2 - 4\delta(1-\delta d)\mu}}{2\delta}$$

The largest of the fixed points $\check{c}$ is above μ (See Lemma 2 in the appendix for a proof) and therefore outside of the interval of interest $(0,\mu)$. Hence, we conclude that $c^* \in (0,\mu)$. □

Observe that $\Gamma(c)$ is strictly increasing in c.

$$\frac{\partial\Gamma(c)}{\partial c} = \frac{(1+\delta)\,(d-\mu)}{1-\delta\,c} + \frac{(\mu\,(1-d\delta) + (1+\delta)\,(d-\mu)\,c)\,\delta}{(1-\delta\,c)^2} > 0$$

Reducing,

$$(1-\delta\,c)\,(d\delta\,(1-\delta\mu) + d - \mu) > 0$$

$0 < c^* < \mu < d < 1$ and $\frac{1}{2} < \delta < 1$ due to Assumption 1 and 2 and therefore $1 - \delta c > 0$, $d\delta\,(1 - \delta\mu) > 0$ and $d - \mu > 0$. Hence, $\Gamma(c)$ is strictly increasing.

The increasing best response functions and the cutoff point have been illustrated in figure 3.

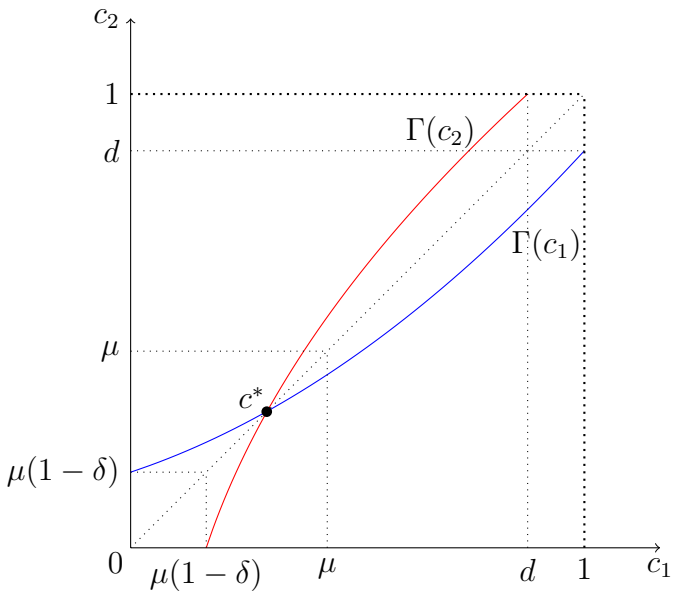

Figure 3: Best response functions and fixed point

We now demonstrate how convergence towards equilibrium takes place. Suppose that player 1 uses a cutoff strategy with cutoff point $\tilde{c}_1^1 = \mu$, i.e. all dominant strategy hawks arm. Thus, the probability that player 1 decides to arm is $\tilde{c}_1^1 = \mu$. For this probability, the fear of setting off an arms race dominates the first and second strategic considerations. Therefore, player 2's best response is to use a *lower* cutoff point, i.e. $\tilde{c}_2^2 = \Gamma(\tilde{c}_1^1) < \tilde{c}_1^1$ (higher level of cooperation). Next, if player 2 arms with probability $\tilde{c}_2^2$, the fear of setting off an arms race again dominates. Thus, player 1's best response is to use a lower cutoff point, i.e. $\tilde{c}_1^3 = \Gamma(\tilde{c}_2^2) < \tilde{c}_2^2$. Hence, the equilibrium can be reached through iterated deletion of dominated strategies. During this cascade, dominant strategy hawks who are almost coordinating types choose to cooperate, which in turn causes "almost-almost coordinating types" to cooperate, etc. Continuing this way, the cutoff points eventually converge down towards c^*.

In period 1, a dominant strategy hawk makes the following strategic calculation:

> If I build a nuclear weapon now, I may get a temporary nuclear monopoly, which could be permanent if my adversary is sufficiently pacifistic. Moreover, my adversary may be building a nuclear weapon himself and I do not want to be left behind in the event of an arms race without a weapon myself. On the other hand, my adversary may not have any plans to build a nuclear bomb and if I build one, he could be forced to respond in kind, thus provoking a nuclear arms race.

In the above we see how the third strategic consideration acts as a counterweight to the destabilizing effect of the first and second strategic considerations. The fear of being left behind and the prospect of getting ahead are counted by the fear of setting off an arms race. Which of the strategic considerations are strongest? Proposition 4 provides an answer. For the share $[0, c^*]$, the first and second strategic considerations dominate the third and consequently this share of the dominant strategy hawks decides to arm. For the remaining share $(c^*, 1]$, the fear of setting off an arms race is the dominant strategic consideration, and this share therefore does not arm.

Three different outcomes are possible in the long run. Figure 4 illustrates these outcomes in the different periods of the game. c^* and d are key determinants of the outcome. Only if both types are above c^*, is peaceful cooperation possible. If player 1 chooses B in period 1 and player 2 chooses N, then player 2 retaliates in period 2 and onwards by also choosing B if he is not a dominant strategy dove.

Notice that $c^* > \mu$ in the SARM, whereas $c^* < \mu$ in the dynamic model. In our model, the fear of setting off an arms race leads to higher levels of cooperation (i.e. a lower c^*). In the SARM, dominant strategy hawks always arm and some of the coordinating types also arm out of fear of being left behind. Thus, when the shadow of the future is taken into account, coordinating types never arm and even some dominant strategy hawks choose to cooperate.

The lower bound on δ ($\delta > \frac{1}{2}$ by Assumption 2) ensures that players care sufficiently much about the future so that $c^* < \mu$. Dominant strategy doves never engage in the punishment phase by arming if the adversary arms. Only dominant

Period 1

	$c_2 \in [0, c^*]$	$c_2 \in (c^*, d]$	$c_2 \in (d, 1]$
$c_1 \in [0, c^*]$	(B, B)	(B, N)	(B, N)
$c_1 \in (c^*, d]$	(N, B)	(N, N)	(N, N)
$c_1 \in (d, 1]$	(N, B)	(N, N)	(N, N)

Period $t > 1$

	$c_2 \in [0, c^*]$	$c_2 \in (c^*, d]$	$c_2 \in (d, 1]$
$c_1 \in [0, c^*]$	(B, B)	(B, B)	(B, N)
$c_1 \in (c^*, d]$	(B, B)	(N, N)	(N, N)
$c_1 \in (d, 1]$	(N, B)	(N, N)	(N, N)

Figure 4: Short-term and long-term outcomes

strategy hawks would want to arm against an unarmed adversary. Now imagine that δ falls below $\frac{1}{2}$ such that $c^* > \mu$. If player 1 is a coordinating type ($c_1 \in (\mu, c^*)$) and chooses B in period 1 and player 2 is a dominant strategy dove that always chooses N, then player 1 will not be willing to act in accordance with the conditional trigger strategy and choose B for all subsequent periods. Remember that coordinating types only prefer to arm if the adversary arms. Hence, the equilibrium described in Proposition 4 can only be a PBE if $\delta > \frac{1}{2}$.

Interestingly, we observe that asymmetric information improves the likelihood of peaceful cooperation for certain types c_1 and c_2. If player 1 is a dominant strategy dove $d \in (d, 1]$ and player 2 is a dominant strategy hawk $c_2 \in (c^*, \mu]$, players cooperate, i.e. (N, N) is chosen in each period. Player 2 does not realize that player 1 is unwilling to engage in punishment. Therefore, peaceful cooperation ensues. Under complete information however if player 2 realized that player 1 was merely a dominant strategy dove, he would not fear that acquiring nuclear weapons would provoke an arms race and he would therefore arm unilaterally. Thus, when types are in this interval, asymmetric information facilitates mutual cooperation.

The Shadow of the Future

In our dynamic arms race model, the shadow of the future induces a high level of cooperation. In the following, we show that the level of cooperation increases when the shadow of the future is increased. The magnitude of the shadow of the future is defined as δ. A higher δ means that players discount the payoff from future events at a lower rate. Consequently, players are more concerned about the future and thus more likely to cooperate. The stabilizing effect of increasing the shadow of the future is formally shown in Proposition 5.

Proposition 5. *Increasing the shadow of the future leads to higher levels of cooperation. In particular,*

$$\frac{\partial c^*}{\partial \delta} < 0$$

When states are highly concerned about the prospect of a future nuclear arms race, they are less likely to be tempted by the possible gains of getting ahead and less likely to let their fear of being left behind drive them into an arms race.

Destabilizing Hawks

Increasing the share of dominant strategy hawks leads to lower levels of cooperation. Thus, dominant strategy hawks are destabilizing regardless of whether we are looking at the SARM or our dynamic model. μ represents the share of dominant strategy hawks. A higher μ leads to more instability.

Proposition 6. *Increasing the proportion of dominant strategy hawks makes cooperation less likely. In particular,*

$$\frac{\partial c^*}{\partial \mu} > 0$$

Increasing μ means there is a higher risk of facing an aggressive player. Therefore, it will be less attractive for the coordinating types to stay out of the arms race. When the fear of being left behind increases, cooperation becomes unattractive for a larger share of types. Hence, c^* increases.

Destabilizing Doves

The effect of increasing the share of dominant strategy doves on the level of cooperation is not immediately clear. On the one hand, increasing the proportion of doves generates more stability by reducing the fear of being left behind in the arms race. Dominant strategy doves never arm and therefore the preemptive motive is less significant. On the other hand, a larger share of dominant strategy doves increases the predatory motive of dominant strategy hawks since dominant strategy hawks are more likely to get away with unilaterally acquiring nuclear weapons. If the risk of provoking an arms race is sufficiently low, there is less stimulus to cooperate for the predatory types. Thus, from the unset, it is unclear whether dominant strategy doves are stabilizing or destabilizing.

Proposition 7 shows that the destabilizing effect dominates and therefore a larger share of dominant strategy doves leads to a lower level of cooperation.

Proposition 7. *Increasing the proportion of dominant strategy doves makes cooperation less likely. In particular,*

$$\frac{\partial c^*}{\partial d} < 0$$

Empirics and Policy Implications

The strategic dilemma depicted by our dynamic model provides an explanation for the remarkable absence of nuclear proliferation in the post-war period. In the 1960's, it was predicted that there would be 15-25 nuclear powers at the end of the 1970's. Nuclear weapons were seen as the ultimate means of defending one's national interests and since no states wanted to be left behind in a nuclear arms race, predictions in the immediate aftermath of World War II were that nuclear proliferation would be very widespread[3].

These pessimistic predictions are consistent with the pessimism implied by the SARM where only the first and second strategic considerations are included. In the

[3]See Potter and Mukhatzhanova (2010) for a lengthy discussion of this.

SARM, even peaceful states can be drawn into an arms race due to an escalating cycle of pessimistic expectations.

Despite the predictions made earlier, the extent of nuclear proliferation turned out to be rather limited. In 2015, there were only 9 nuclear powers – far fewer than what had previously been envisioned.

Numerous international relations theories have attempted to provide an explanation for this phenomenon. Neoliberalists have pointed to the Nuclear Non-Proliferation Treaty (NPT) of 1968 and similar international institutions as having played an important role in preventing unrestricted nuclear proliferation (Keohane, 1984). Liberalists have stressed the importance of domestic politics (Solingen, 2010) and constructivists – the psychology of the leader (Hymans, 2006).

However, according to our model, these higher order explanations are not necessary in order to account for the phenomenon. The absence of nuclear proliferation can be attributed to the simple fact that states are concerned about the future and do not wish to set off an arms race. If the risk of triggering an arms race is sufficiently high, most states – even predatory ones – prefer to cooperate.

Thus, our model provides a plausible explanation behind the surprising absence of nuclear proliferation in the post-war period.

On pacifistic military doctrines As we have seen in the previous section, our dynamic arms race model reveals some surprising results about the role of pacifistic states and the likelihood of peaceful cooperation. On the one hand, the presence of pacifistic states alleviates fears of being left behind, thus making an arms race less likely. On the other hand, pacifistic states induce predatory states to arm unilaterally, thus making an arms race more likely. From the unset, it is not clear which of these strategic considerations is dominant.

Proposition 7 shows that pacifistic states are destabilizing. Their refusal to respond in kind to the actions of aggressive states undermines the fear of setting off an arms race and thus the incentive to cooperate.

That pacifistic states are destabilizing should not however be interpreted as an encouragement to pursue an aggressive foreign policy. Rather, it suggests that

any foreign policy doctrine of foregoing nuclear weapons should not be absolute and unconditional. In foreign policy doctrines, states should emphasize that their decision to abstain from developing nuclear weapons is *conditional* and can be reversed if others states start to acquire nuclear weapons.

Also this should not be taken as evidence for avoiding friendly relationships with potential adversaries. On the contrary, confidence building measures that alleviate suspicion and fear among states should be encouraged. Remember that a lower share of dominant strategy hawks or *the belief hereof* is stabilizing. However, blind pacifism and strategies of foregoing nuclear weapons regardless of threat perception should be avoided.

Conclusion

We develop an arms race model using a repeated game where states simultaneously decide whether to build nuclear weapons or to abstain in each period. When determining the optimal level of cooperation, states take into account three strategic considerations. The *fear of being left behind* and the *opportunity to get ahead* make states more likely to acquire nuclear weapons. However, the *fear of setting off an arms race* makes states less likely to acquire nuclear weapons. The destabilizing effect of the first and the second strategic considerations are countered by the stabilizing effect of the third. In equilibrium, the level of cooperation is determined as a balance between these three strategic considerations.

The contribution of this paper is threefold:

(i) We merge the static arms race model by Baliga and Sjöström (2004, 2012) with a game of repeated interaction creating a unified theoretical framework, in which all three strategic considerations are included.

(ii) Our spiral model reverses some of the results in existing arms races models. In particular, we show that increasing the share of pacifistic states makes cooperation *less likely*.

(iii) Our model provides an explanation for the remarkable absence of nuclear proliferation in the post-war period.

Appendix

Proof (for period $t > 1$ in Proposition 4). There are four possible outcomes in period 1: (B, B), (N, B), (B, N) and (N, N). These outcomes give rise to four different continuation games. We analyze each of the four continuation games consecutively and show that a conditional trigger strategy, in which player 1 punishes defection from the cooperation (N, N) by choosing B in all subsequent periods if and only if $c_1 \leq d$, is an equilibrium strategy. In a perfect Bayesian equilibrium, the restriction of strategy and belief-system in every continuation game must be a Bayesian Nash equilibrium.

(B,B): We consider the continuation games in period $t > 1$ where B was chosen by both players in period 1. We can identify two different types of continuation games belonging to this category. The *first* category is the continuation games in period $t > 1$ where (B, B) was chosen in all previous periods $1, \ldots, t-1$, i.e. the continuation game on the equilibrium path. The *second* category is the rest, i.e. all the continuation games off the equilibrium path, in which (B, B) was not chosen in at least one of the previous periods $2, \ldots, t-1$. We define the beliefs and strategies of both categories of continuation games. Then, we show that these beliefs and strategies must be part of an equilibrium.

Strategies: Players choose (B, B) in period $t, t+1, \ldots$ in both types of continuation games, i.e. regardless of whether players are on the equilibrium path or not.

Beliefs: In the *first* category of continuation games located on the equilibrium path, Bayes' rule apply. Since (B, B) was chosen in period 1, players infer that types are below c^*. Thus, in period 2, beliefs become the uniform distribution $G(c) = \frac{c}{c^*}$ restricted to the interval $[0, c^*]$. Since B is chosen by all remaining types along the equilibrium path, beliefs are unchanged for the remaining of the game. Hence, the updated beliefs are the uniform distribution $G(c) = \frac{c}{c^*}$ restricted to the interval $[0, c^*]$ for all periods $t > 1$.

In the *second* category of continuation games located off the equilibrium path where N was chosen by either or both of the players in any of the periods $2, \ldots, t-$

1, beliefs remain unchanged due to the Pessimistic Bias Assumption, i.e. beliefs follow the uniform distribution $G(c) = \frac{c}{c^*}$ restricted to the interval $[0, c^*]$ in period t. In effect, this implies that only the actions in period 1 matter for the beliefs in the remaining of the game. The beliefs formed in period 2 remain unchanged throughout the game regardless of the actions taken by players in period 2 and onwards.

Payoff and equilibrium: In the *first* category of continuation games, we show that choosing B in period $t, t+1, \ldots$ is the best response for player 1 given that player 2 chooses B in period $t, t+1, \ldots$ and given that player 2 will choose B in all periods in any subsequent, off equilibrium continuation game. If player 1 chooses B in period $t, t+1, \ldots$, he receives the following payoff: $-c_1 + \sum_{s=1}^{\infty} \delta^s(-c_1)$. If instead player 1 deviates and chooses B in period t and N in period $t+1, t+2, \ldots$, he receives following payoff: $-d + \sum_{s=1}^{\infty} \delta^s(-c_1)$. This is clearly lower since $c_1 < c^* < \mu < d$. Hence, deviating yields a lower payoff. Hence, choosing B in period $t, t+1, \ldots$ is the best response for player 1 when player 2 chooses B in period $t, t+1, \ldots$.

In the *second* category of continuation games, beliefs are the same and an analogous argument applies. Thus, choosing B in period $t, t+1, \ldots$ is the best the response for player 1 when player 2 is choosing B in period $t, t+1, \ldots$ and will choose B in all periods in any subsequent, off equilibrium continuation game.

> **Pessimistic Bias Assumption** *In the off-equilibrium information sets, if player 2's actions are consistent with both $c_2 \leq c^*$ and $c_2 > c^*$, then player 1 believes that $c_2 \leq c^*$.*

The Pessimistic Bias Assumption states that player 1 is always negatively biased towards believing that player 2 is of a low type if player 2's behavior is consistent with the actions of both a low type and a high type. Only in the off-equilibrium information sets does this assumption become relevant. The Pessimistic Bias Assumption imposes a high degree of simplicity on the off-equilibrium beliefs, which makes the analysis a lot easier[4].

[4]Notice that other assumptions that impose a more even balance between positive and nega-

Notice that the Pessimistic Bias Assumption has been applied to the beliefs of the players above. If player 1 chose B in periods $1, \ldots, t-1$, then player 2's beliefs remain unchanged if player 1 suddenly chooses N in period t, i.e. beliefs remain $G(c) = \frac{c}{c^*}$ restricted to the interval $[0, c^*]$ in period $t+1$. In contrast, assume that player 1 had chosen N in periods $1, \ldots, t-1$, then player 2's beliefs would change from the uniform distribution $H(c) = \frac{c-c^*}{1-c^*}$ restricted to the interval $[c^*, 1]$ to $G(c) = \frac{c}{c^*}$ restricted to the interval $[0, c^*]$ in period $t+1$ if player 1 chooses B in period t.

(B,N): We consider the continuation games in period $t > 1$ in which (B, N) was chosen in period 1. This continuation game gives rise to different continuation games. The *first* category is the continuation games on the equilibrium path in period $t > 1$, in which (B, B) was chosen in all periods $2, \ldots, t-1$ if $c_2 \leq d$ and (B, N) was chosen in all periods $2, \ldots, t-1$ if $c_2 > d$.

The *second* category is the rest, i.e. all the continuation games off the equilibrium path in period $t > 1$, in which either (B, N), (N, B) or (N, N) was chosen at least once during the periods $2, \ldots, t-1$ if $c_2 \leq d$.

First we define the beliefs and strategies of the continuation games. Then, we show that these beliefs and strategies are part of an equilibrium.

Strategies: After (B, N) was chosen in period 1, a punishment phase is triggered, in which player 2 chooses B in all subsequent periods $t+1, t+2, \ldots$ if $c_2 \leq d$ and N in all subsequent periods $t+1, t+2, \ldots$ if $c_2 > d$. Player 1 continues to choose B in all subsequent periods regardless of player 2's choice. These strategies are used in continuation games of both *first* and *second* category.

Player 1's beliefs in continuation game of first category: Since (B, N) was chosen in period 1, player 1 infers that player 2's type must be above c^*. Using Bayes' rule, player 1's updated beliefs in period 2 are the uniform distribution $H(c) = \frac{c-c^*}{1-c^*}$ restricted to the interval $[c^*, 1]$. In the *first* category of the continuation games on the equilibrium path, player 1's beliefs in period 3 and onwards depend on the

tive bias when inconclusive evidence of types arise are possible. Assumptions giving more favor to the positive bias would also be able to support the equilibrium strategies on the equilibrium path. However, using these more sophisticated assumptions would complicate the analysis unduly.

actions taken by player 2 in period 2.

- If player 2 chooses B in period 2, player 1 infers that $c^* < c_2 \leq d$ and therefore player 1's updated beliefs in period 3 are the uniform distribution $H_l(c) = \frac{c-c^*}{d-c^*}$ restricted to the interval $(c^*, d]$. Since all of player 2's types continue to choose B, no further updating takes place and player 1's beliefs are the uniform distribution $H_l(c)$ restricted to $(c^*, d]$ for all periods $t > 2$.

- If player 2 chooses N in period 2, player 1 infers that $c_2 > d$ and thus player 1's updated beliefs in period 3 are the uniform distribution $H_h(c) = \frac{c-d}{1-d}$ restricted to the interval $(d, 1]$. Since all of player 2's types continue to choose N, player 1's beliefs remain the uniform distribution $H_h(c)$ restricted to $(d, 1]$ for all periods $t > 2$.

Player 2's beliefs in continuation game of first category: Since (B, N) was chosen in period 1, player 2 infers that player 1's type must be below c^*. Thus, player 2's updated beliefs in period 2 are the uniform distribution $G(c) = \frac{c}{c^*}$ restricted to the interval $[0, c^*]$. Since all of player 1's remaining types choose B for the rest of the game, no further updating takes place and player 2's beliefs are the uniform distribution $G(c)$ restricted to $[0, c^*]$ for all periods $t > 1$. Note, that if for some reason player 1 defects from this strategy by choosing N instead of B, player 2's beliefs are assumed to be identical to the beliefs on the equilibrium path, i.e. beliefs are the uniform distribution $G(c)$ restricted to $[0, c^*]$.

In the *second* category of the continuation games, if either of the players chooses N in any of the periods $3, \ldots, t-1$ after choosing B in period 2, beliefs are unchanged according to the Pessimistic Bias Assumption, i.e. player 1's beliefs remain the uniform distribution $H_l(c)$ restricted to $(c^*, d]$ and player 2's beliefs – the uniform distribution $G(c) = \frac{c}{c^*}$ restricted to the interval $[0, c^*]$. In contrast, if player 2 chooses B in any of the periods $3, \ldots, t-1$ while choosing N in all the previous periods, player 1's beliefs immediately change to $H_l(c)$ restricted to $(c^*, d]$.

Payoff and equilibrium in a continuation game of the first category:

Player 1: Given that player 2 chooses B in all periods $t > 1$ if $c_2 \leq d$ and N in all periods $t > 1$ if $c_2 > d$, we show that choosing B in all periods $t > 1$ is the optimal strategy for player 1. player 1's payoff from choosing B is

$$H(d)\left(\sum_{s=0}^{\infty}\delta^s(-c_1)\right) + (1 - H(d))\left(\sum_{s=0}^{\infty}\delta^s(\mu - c_1)\right)$$

The first part of the expression is the player 1's payoff if player 2 is a type $c_2 \in (c^*, d]$ that always chooses B. The second part of the expression is player 1's payoff if player 2 is a dominant strategy dove $c_2 \in (d, 1]$ that always chooses N.

If player 1 deviates and chooses N in period t and B in period $t+1, t+2, \ldots$ the payoff is

$$H(d)\left(-d + \sum_{s=1}^{\infty}\delta^s(-c_1)\right) + (1 - H(d))\left(\sum_{s=1}^{\infty}\delta^s(\mu - c_1)\right)$$

In this expression, we have substituted $-c_1$ with $-d$ and $\mu - c_1$ with zero. Since, $c_1 < c^* < \mu < d$ the payoff from choosing B is strictly lower than from choosing N. Note that the payoff assumes that players return to their equilibrium strategies if a deviation is ever made. Remember that the strategies in the continuation games of the *second* category are identical to the strategies chosen in continuation games of the *first* category.

Player 2: Given that player 1 chooses B in all periods $t > 1$, we need to show that choosing B in all periods $t > 1$ is the optimal strategy for player 2 if $c_2 \leq d$. Choosing B in all periods $t > 1$ yields the payoff $\sum_{s=0}^{\infty}\delta^s(-c_2)$ for player 2. Choosing N in one or several periods irrevocably lowers the payoff since $c_2 < d$, e.g. if player 2 deviates in period t by choosing N the payoff is $-d + \sum_{s=1}^{\infty}\delta^s(-c_2)$ which is clearly lower than $\sum_{s=0}^{\infty}\delta^s(-c_2)$. If $c_2 > d$, player 2 is a dominant strategy dove and choosing N is therefore optimal by default. Note, player 2's beliefs are unchanged off the equilibrium path, choosing B in all periods $t, t+1, \ldots$ is a best response for player 2 regardless of whether we are on the equilibrium path or not.

Payoff and equilibrium in a continuation game of the second category: In the continuation games of the *second* category where a deviation from the equilibrium strategies has been made in period $2, \ldots, t-1$ beliefs remain unchanged except

that player 1's beliefs change from the uniform distribution $H_h(c) = \frac{c-d}{1-d}$ restricted to the interval $(d, 1]$ to the uniform distribution $H_l(c) = \frac{c-c^*}{d-c^*}$ restricted to the interval $(c^*, d]$ if (B, B) was chosen in any of the periods $2, \ldots, t-1$ when $c_2 > d$. Therefore, the arguments above also remain unchanged. Player 1 chooses B in all subsequent periods $t, t+1, \ldots$. Player 2 chooses B in all subsequent periods $t, t+1, \ldots$ if $c_2 \leq d$ and N in all subsequent periods $t, t+1, \ldots$ if $c_2 > d$. Note that player 1 chooses B in all subsequent periods $t, t+1, \ldots$ regardless of player 2's strategy. Thus, if (B, B) was chosen in any period $2, \ldots, t-1$ when $c_2 > d$ and player 1's beliefs subsequently changed, player 1 still prefers to choose B in all subsequent periods $t, t+1, \ldots$.

(N,N): We consider the continuation games in period $t > 1$, in which N was chosen by both players in period 1. We can identify two different types of continuation games belonging to this category.

The *first* category is the continuation games in period $t > 1$ where (N, N) was chosen in all previous periods $1, \ldots, t-1$, i.e. the continuation game on the equilibrium path.

The *second* category is the continuation games off the equilibrium path, in which (B, N) was chosen in a period r where $r < t$ and (N, N) was chosen in all previous periods $1, \ldots, r-1$.

We define the strategies and beliefs of both categories of continuation games and show that these strategies and beliefs are a part the conditional trigger equilibrium.

Notice that there are other continuation games which belong to neither the *first* or the *second* category, e.g. the continuation game where (B, B) was chosen in one of the previous periods $2, ..., t-1$. Fortunately, we do not need to define strategies and beliefs in this type of continuation game since neither of players can decide to reach this continuation game on their own. If player 1 decides to defect, he will compare the payoff in a continuation game of the *first* category with the payoff in a continuation game of the *second* category. The expected payoff from a continuation game where (N, N) was chosen will not figure in this calculation.

Strategies: In the *first* category of continuation games where (N, N) was chosen in all previous periods $1, \ldots, t-1$, players also choose (N, N) in all subsequent periods $t, t+1, \ldots$. In the *second* category of continuation games where (B, N) was chosen in some period r where $1 < r < t$, a punishment phase follows in which player 2 chooses B in all subsequent periods $r+1, r+2, \ldots$ if $c_2 \leq d$ and N in all subsequent periods $r+1, r+2, \ldots$ if $c_2 > d$. Given player 2's strategy, player 1 also chooses B in all subsequent periods $r+1, r+2, \ldots$.

Beliefs: The *first* category of continuation games is located on the equilibrium path and Bayes' rule can therefore be used to update beliefs. Since (N, N) was chosen in period 1, players infer that their adversary is above c^*. Hence, the updated beliefs in period 2 are the uniform distribution $H(c) = \frac{c-c^*}{1-c^*}$ restricted to the interval $[c^*, 1]$. Since N is chosen by all remaining types along the equilibrium path, beliefs are unchanged for the remaining of the game. Hence, the updated beliefs are the uniform distribution $H(c) = \frac{c-c^*}{1-c^*}$ restricted to the interval $[c^*, 1]$ for all periods $t > 1$.

In the *second* category of continuation games, (B, N) was chosen in some period r where $1 < r < t$. If player 1 chooses N in period r, player 2's beliefs are assumed to be the uniform distribution $G(c) = \frac{c}{c^*}$ restricted to the interval $[0, c^*]$ in period $r+1$. Due to the Pessimistic Bias Assumption, player 2's beliefs remain unchanged throughout the game regardless of the actions taken by player 1 in the subsequent periods, i.e. player 2's beliefs are the uniform distribution $G(c) = \frac{c}{c^*}$ restricted to the interval $[0, c^*]$ in periods $r+1, r+2, \ldots$ Thus, player 2's beliefs are identical to the beliefs in a situation where B was chosen by player 1 from the very beginning.

Player 1's beliefs remain unchanged in period $r+1$, i.e. the uniform distribution $H(c) = \frac{c-c^*}{1-c^*}$ restricted to the interval $[c^*, 1]$. Subsequently, if player 2 chooses B in period $r+1$, player 1's beliefs are changed to the uniform distribution $H_l(c) = \frac{c-c^*}{d-c^*}$ restricted to the interval $(c^*, d]$ and remain so throughout the game. Alternatively, if player 2 chooses N in period $t+1$, player 1's beliefs change to the uniform distribution $H_h(c) = \frac{c-d}{1-d}$ restricted to the interval $(d, 1]$ and remain this way throughout the game unless player 2 suddenly chooses B, in which case player 1's beliefs change to the uniform distribution $H_l(c) = \frac{c-c^*}{d-c^*}$ restricted to the interval

$(c^*, d]$ according to the Pessimistic Bias Assumption.

Payoff and equilibrium: In the *first* category of continuation games, we show that choosing N in period $t, t+1, \ldots$ is the best response for player 1 given that player 2 chooses N in all subsequent periods $t, t+1, \ldots$ and given that player 2 will start a punishment phase if player 1 defects.

We immediately see that choosing N in all subsequent periods $t, t+1, \ldots$ is strictly preferred by both dominant strategy doves and coordinating types. These types have neither a short-term nor a long-term benefit of choosing B regardless of the strategies chosen by the adversary given that (N, N) was chosen in previous periods $1, \ldots, t-1$. Thus, in the following we only consider the incentive to defect for player 1 if he is a dominant strategy hawk $c_1 \in (c^*, \mu]$.

We show that it is optimal for player 1 to use a cutoff strategy with a cutoff point $\hat{c}$ with $\tilde{c} < c^*$. $\tilde{c} < c^*$ implies that all of player 1's remaining types in the interval $(c^*, \mu]$ choose N. Hence, N must be player 1's best response.

Player 1's payoff from choosing N in periods $t+1, t+2, \ldots$ is

$$\sum_{s=0}^{\infty} \delta^s \cdot 0 \tag{3}$$

If player 1 defects by choosing B in any of the periods $t > 1$, players enter a *second* category continuation game, in which player 1 chooses B in all subsequent periods $t+1, t+2, \ldots$ and player 2 chooses B in all subsequent periods $t+1, t+2, \ldots$ if $c_2 \leq d$ and chooses N in all subsequent periods $t+1, t+2, \ldots$ if $c_2 > d$.

Thus, player 1's expected payoff from defecting by choosing B in periods $t+1, t+2, \ldots$ is

$$H(d)\left(\mu + \sum_{s=0}^{\infty} \delta^s(-c_1)\right) + (1 - H(d))\left(\sum_{s=0}^{\infty} \delta^s(\mu - c_1)\right) \tag{4}$$

The first part of the expression is player 1's payoff if player 2 is a type $c_2 \in (c^*, d]$ that always chooses B. The second part of the expression is player 1's payoff if player 2 is a dominant strategy dove $c_2 \in (d, 1]$ that always chooses N.

Subtracting (3) from (4), simplifying and using $\sum_{s=0}^{\infty} \delta^s = \frac{1}{1-\delta}$, player 1 receives

the following net gain from defecting:

$$\frac{(1-d)\,\mu}{(1-\delta)\,(1-c^*)} - \frac{1}{1-\delta}c_1$$

If the net gain is positive, choosing B is optimal for player 1. If the net gain is negative, choosing N is the optimal strategy for player 1. If the net gain is zero, player 1 is indifferent between choosing B and N. However, for convenience we assume that player 1 chooses B in this case. The net gain from defecting is clearly decreasing in player 1's type c_1. This proves that the best response for player 1 is to use a cutoff strategy. We find the cutoff point $\tilde{c}$ by setting the net gain from defecting equal to zero and simplifying. Thus,

$$\tilde{c} = \frac{(1-d)\,\mu}{1-c^*}$$

$\tilde{c}$ is the point at which player 1 is indifferent between choosing N in all subsequent periods $t+1, t+2, \ldots$ and choosing B in all subsequent periods $t+1, t+2, \ldots$. In Lemma 1 we show that $\tilde{c} \leq c^*$. This implies that all of player 1's remaining types in the interval $(c^*, \mu]$ choose N.

In the *second* category of continuation games, (B, N) was chosen in one of the previous periods r where $r < t$, i.e. player 1 deviated by choosing B in period r. Again, we only consider this deviation for a dominant strategy hawk $c_1 \in (c^*, \mu]$ since only dominant strategy hawks would ever consider to deviate from the cooperative outcome. The coordinating types and the dominant strategy doves have neither short-term nor long-term benefits of defecting regardless of the actions subsequently taken by player 2.

We show that choosing B in period $t, t+1, \ldots$ is an optimal strategy for player 1 given that player 2 chooses B in period $t, t+1, \ldots$ whenever $c_2 \leq d$ and N in period $t, t+1, \ldots$ if $c_2 \leq d$. Player 1's payoff in period t from choosing B in period $t, t+1, \ldots$ is

$$H(d)\left(\sum_{s=0}^{\infty}\delta^s(-c_1)\right) + (1-H(d))\left(\sum_{s=0}^{\infty}\delta^s(\mu-c_1)\right) \tag{5}$$

The first part of the expression is the player 1's payoff if player 2 is a type $c_2 \in (c^*, d]$ that always chooses B. The second part of the expression is player 1's payoff if player 2 is a dominant strategy dove $c_2 \in (d, 1]$ that always chooses N.

Player 1's payoff in period t from deviating by choosing N in period t and B in periods $t+1, t+2, ...$ is

$$H(d)\left(-d+\sum_{s=1}^{\infty}\delta^s(-c_1)\right)+(1-H(d))\left(\sum_{s=1}^{\infty}\delta^s(\mu-c_1)\right) \tag{6}$$

Clearly, (5) is larger than (6) since $c_1 \leq \mu < d$. Hence, choosing B in period $t, t+1, \ldots$ is an optimal strategy for player 1.

Next, we show that choosing B in all subsequent periods $t, t+1, \ldots$ is an optimal strategy for player 2 if $c_2 \leq d$ given that player 1 chooses B in all subsequent periods $t, t+1, \ldots$. If player 1 chooses B in period r, player 2's belief in periods $r+1, t+2, \ldots$ becomes $G(c) = \frac{c}{c^*}$ restricted to the interval $[0, c^*]$ using the Pessimistic Bias Assumption. Player 2's payoff in period t from choosing B in period $t, t+1, \ldots$ is $\sum_{s=0}^{\infty}\delta^s(-c_2)$. Player 2's payoff in period t from choosing N in period t and B in period $t+1, t+2, \ldots$ is $-d+\sum_{s=1}^{\infty}\delta^s(-c_2)$. Hence, choosing B in period $t, t+1, \ldots$ is an optimal strategy for player 2 whenever $c_2 \leq d$. □

Lemma 1.

$$\tilde{c} \leq c^*$$

Proof.

$$\tilde{c} = \frac{(1-d)\,\mu}{1-c^*}$$

Since $1-c^* < 0$, we have

$$c^*(1-c^*) \leq (1-d)\,\mu$$

Inserting

$$c^* = \frac{1-(1-\delta)(d-\mu)-\sqrt{(1-(1-\delta)(d-\mu))^2-4\delta(1-\delta d)\mu}}{2\delta}$$

we get

$$\frac{1-(1-\delta)(d-\mu)-\sqrt{(1-(1-\delta)(d-\mu))^2-4\delta(1-\delta d)\mu}}{2\delta}\cdot$$
$$\left(1-\frac{1-(1-\delta)(d-\mu)-\sqrt{(1-(1-\delta)(d-\mu))^2-4\delta(1-\delta d)\mu}}{2\delta}\right) \leq (1-d)\,\mu$$

Collecting the square root and reducing,

$$(1+\mu-d)\sqrt{(1-(1-\delta)(d-\mu))^2-4\delta(1-\delta d)\mu}$$
$$\geq -d^2\delta+2\,d\delta\,\mu-\delta\,\mu^2+d^2+d\delta-2\,d\mu-3\,\delta\,\mu+\mu^2-2\,d+2\,\mu+1$$

Assumption 1 implies that the left hand side is clearly positive. If the right hand side is negative, the expression is clearly true. If the right hand side is positive we can apply the function $f(x)=x^2$ which is increasing on $[0,\infty)$ on both sides.

$$(1+\mu-d)^2\left((1-(1-\delta)(d-\mu))^2-4\delta(1-\delta d)\mu\right)$$
$$\geq \left(-d^2\delta+2\,d\delta\,\mu-\delta\,\mu^2+d^2+d\delta-2\,d\mu-3\,\delta\,\mu+\mu^2-2\,d+2\,\mu+1\right)^2$$

Reducing,

$$4\,\mu\,\delta^2\,(1-\delta)^2\,(1-d)\,(d-\mu)\,(2+\mu-d)\geq 0$$

Using Assumptions 1 and 2, we see that this expression is clearly positive. Thus, $\tilde{c}\leq c^*$. □

Lemma 2.

$$\mu<\frac{1-(1-\delta)(d-\mu)+\sqrt{(1-(1-\delta)(d-\mu))^2-4\delta(1-\delta d)\mu}}{2\delta}$$

Proof. Isolating the square root,

$$-d\delta+3\,\delta\,\mu+d-\mu-1<\sqrt{(1-(1-\delta)(d-\mu))^2-4\delta(1-\delta d)\mu}$$

If the left hand side is negative, the expression is true because the right hand side is always positive. If the left hand side is positive, we can apply $f(x)=x^2$ which is increasing on $[0,\infty)$ on both sides.

$$(-d\delta+3\,\delta\,\mu+d-\mu-1)^2<(1-(1-\delta)(d-\mu))^2-4\delta(1-\delta d)\mu$$

Reducing,

$$4\,\delta\,\mu\,(2\,\delta-1)\,(d-\mu)>0$$

Assumptions 1 and 2 ensure that all the terms are positive. Thus, the expression is true. □

Proof (of proposition 5). Differentiating c^* with respect to δ,

$$\frac{\partial c^*}{\partial \delta} = \frac{1}{2\delta}\left(d-\mu-\frac{\left(1-\left(1-\delta\right)\left(d-\mu\right)\right)\left(d-\mu\right)-2\,\left(-d\delta+1\right)\mu+2\,\delta\,d\mu}{\sqrt{\left(1-\left(1-\delta\right)\left(d-\mu\right)\right)^2-4\,\delta\,\left(-d\delta+1\right)\mu}}\right)$$
$$-\frac{1-\left(1-\delta\right)\left(d-\mu\right)-\sqrt{\left(1-\left(1-\delta\right)\left(d-\mu\right)\right)^2-4\,\delta\,\left(-d\delta+1\right)\mu}}{2\delta^2}<0$$

Rearranging,

$$-\frac{\delta\,\left(\left(1-\left(1-\delta\right)\left(d-\mu\right)\right)\left(d-\mu\right)-2\,\left(-d\delta+1\right)\mu+2\,\delta\,d\mu\right)}{\sqrt{\left(1-\left(1-\delta\right)\left(d-\mu\right)\right)^2-4\,\delta\,\left(-d\delta+1\right)\mu}}$$
$$<-d\delta+\delta\,\mu+1-\left(1-\delta\right)\left(d-\mu\right)-\sqrt{\left(1-\left(1-\delta\right)\left(d-\mu\right)\right)^2-4\,\delta\,\left(-d\delta+1\right)\mu}$$

Multiplying with $\sqrt{\left(1-\left(1-\delta\right)\left(d-\mu\right)\right)^2-4\,\delta\,\left(-d\delta+1\right)\mu}$ on both sides and rearranging,

$$-\left(\delta-1\right)\mu^2-\left(\left(-2\,\delta+2\right)d+3\,\delta-2\right)\mu-\left(d-1\right)\left(\left(\delta-1\right)d+1\right)$$
$$<\left(-d+\mu+1\right)\sqrt{\left(d+\mu\right)^2\delta^2+\left(-2\,\mu^2+\left(4\,d-6\right)\mu-2\,d^2+2\,d\right)\delta+\left(d-\mu-1\right)^2}$$

If the left hand side is negative, the expression is clearly true since the right hand side is always positive. If the left hand side is positive, we can apply $f(x)=x^2$ on both sides of the inequality.

$$\left(-\left(\delta-1\right)\mu^2-\left(\left(-2\,\delta+2\right)d+3\,\delta-2\right)\mu-\left(d-1\right)\left(\left(\delta-1\right)d+1\right)\right)^2$$
$$<\left(-d+\mu+1\right)^2\left(\left(d+\mu\right)^2\delta^2+\left(-2\,\mu^2+\left(4\,d-6\right)\mu-2\,d^2+2\,d\right)\delta+\left(d-\mu-1\right)^2\right)$$

Reducing,

$$4\,\delta^2\mu\,\left(1-d\right)\left(d-\mu\right)\left(2-\left(d-\mu\right)\right)>0$$

Due to Assumptions 1 and 2, the expression is clearly true. □

Proof (of proposition 6). Differentiating c^* with respect to μ,

$$\frac{\partial c^*}{\partial \mu} = \frac{1}{2\delta}\left(1-\delta-\frac{(1-\delta)\left(1-(1-\delta)(d-\mu)\right)-2\delta(1-\delta d)}{\sqrt{\left(1-(1-\delta)(d-\mu)\right)^2+4\delta(1-\delta d)\mu}}\right) > 0$$

Rearranging,

$$\begin{aligned} z \equiv\ & (1-\delta)\left(1-(1-\delta)\left(d-\mu\right)\right)-2\,\delta\,\left(1-d\delta\right) \\ & < (1-\delta)\sqrt{\left(1-(1-\delta)\left(d-\mu\right)\right)^2+4\,\delta\,\left(d\delta-1\right)\mu} \end{aligned}$$

The right hand side is always positive due to Assumption 2. Therefore, the expression will be true if we can prove that $z < 0$. Rearranging $z < 0$,

$$0 > (d+\mu)\,\delta^2+(2\,d-2\,\mu-3)\,\delta-d+\mu+1$$

The right hand side is a 2. order polynomial. Solving for δ, we get

$$\delta_1 = \frac{-2\,d+2\,\mu+3-\sqrt{8\,d^2-8\,d\mu-16\,d+8\,\mu+9}}{2(d+\mu)}$$

$$\delta_2 = \frac{-2\,d+2\,\mu+3+\sqrt{8\,d^2-8\,d\mu-16\,d+8\,\mu+9}}{2(d+\mu)}$$

Since $0 < \mu + d$ by assumption 1, the coefficient of δ^2 in the polynomial is positive. Therefore, $z < 0$ must be true for $\frac{1}{2} < \delta < 1$ if $\delta_1 < \frac{1}{2}$ and $1 < \delta_2$. First, we show that $\delta_1 < \frac{1}{2}$.

$$\frac{-2\,d+2\,\mu+3-\sqrt{8\,d^2-8\,d\mu-16\,d+8\,\mu+9}}{2(d+\mu)} < \frac{1}{2}$$

Isolating the square root,

$$\sqrt{8\,d^2-8\,d\mu-16\,d+8\,\mu+9} < -d+3\,\mu+3$$

Since the left hand side is clearly positive, we can apply the function $f(x) = x^2$ which is increasing on $[0,\infty)$ on both sides.

$$0 < (-d+3\,\mu+3)^2-8\,d^2+8\,d\mu+16\,d-8\,\mu-9$$

Reducing, we get

$$0 < (d + \mu)\,(10 - 7\,d + 9\,\mu)$$

which is clearly true due to assumption 1.

Next, we show that $1 < \delta_2$.

$$1 < \frac{-2\,d + 2\,\mu + 3 + \sqrt{8\,d^2 - 8\,d\mu - 16\,d + 8\,\mu + 9}}{2(d + \mu)}$$

Rearranging,

$$4\,d - 3 < \sqrt{8\,d^2 - 8\,d\mu - 16\,d + 8\,\mu + 9}$$

If the left hand side is negative, the expression is clear true. The left hand side is positive if $\frac{3}{4} < d$. In this case, we can apply the function $f(x) = x^2$ which is increasing on $[0, \infty)$ on both sides.

$$0 < 8\,d^2 - 8\,d\mu - 16\,d + 8\,\mu + 9 - (4\,d - 3)^2$$

Reducing,

$$0 < 8\ (1 - d)\,(d + \mu)$$

This is clearly true for all $0 < d < 1$ and hence also when $\frac{3}{4} < d$. Therefore, $z < 0$ is true. Thus, $\frac{\partial c^*}{\partial \mu} > 0$. □

Proof (of proposition 7). The derivative of c^* with respect to N must be strictly negative, i.e.

$$\frac{\partial c^*}{\partial d} = \frac{1}{2\delta}\left(-(1 - \delta) + \frac{(1 - \delta)\,(1 - (1 - \delta)(d - \mu)) - 2\delta^2\mu}{\sqrt{(1 - (1 - \delta)(d - \mu))^2 - 4\delta(1 - \delta d)\mu}}\right) < 0$$

Rearranging,

$$z \equiv (1-\delta)\,(1 - (1 - \delta)(d - \mu)) - 2\delta^2\mu < (1-\delta)\sqrt{(1 - (1 - \delta)(d - \mu))^2 - 4\delta(1 - \delta d)\mu}$$

If $z < 0$, the expression is trivially satisfied since the right hand side is always positive. If $z \geq 0$, we can apply the function $f(x) = x^2$ which is increasing on $[0, \infty)$ on both sides.

$$\left((1-\delta)\left(1-(1-\delta)(d-\mu)\right)-2\delta^2\mu\right)^2 < (1-\delta)^2\left(\left(1-(1-\delta)(d-\mu)\right)^2-4\delta(1-\delta d)\mu\right)$$

Reducing,

$$0 < 4\,\delta\,\mu\,(2\,\delta-1)\,(1-\delta\,\mu-\delta)$$

Using $0 < \mu < d$ and $\frac{1}{2} < \delta < 1$ from Assumption 1 and 2, the expression reduces to

$$0 < 1-\delta\,\mu-\delta$$

Isolating δ,

$$\delta < \frac{1}{1+\mu}$$

Thus, we need to show that $\delta < \frac{1}{1+\mu}$ whenever $z \geq 0$. In other words, $\delta < \frac{1}{1+\mu}$ does not need to hold when $z < 0$. Rearranging $z \geq 0$, we get

$$0 < -(\mu+d)\,\delta^2+(2\,d-2\,\mu-1)\,\delta+1+\mu-d$$

The right hand side is a 2. order polynomial with two roots. The coefficient $-(\mu+d)$ is negative and the polynomial is therefore concave. Thus, δ must be between the two roots:

$$\frac{2\,d-2\,\mu-1-\sqrt{-8\,d\mu+8\,\mu^2+8\,\mu+1}}{2\,(d+\mu)} < \delta < \frac{2\,d-2\,\mu-1+\sqrt{-8\,d\mu+8\,\mu^2+8\,\mu+1}}{2\,(d+\mu)}$$

Hence, we must show that

$$\frac{2\,d-2\,\mu-1+\sqrt{-8\,d\mu+8\,\mu^2+8\,\mu+1}}{2\,(d+\mu)} < \frac{1}{1+\mu}$$

Rearranging,

$$(1+\mu)\,\sqrt{-8\,d\mu+8\,\mu^2+8\,\mu+1} < -2\,d\mu+2\,\mu^2+5\,\mu+1$$

The left hand side is positive and we can apply $f(x) = x^2$ which is increasing on $[0, \infty)$ on both sides.

$$(1+\mu)^2\left(-8\,d\mu+8\,\mu^2+8\,\mu+1\right) < \left(-2\,d\mu+2\,\mu^2+5\,\mu+1\right)^2$$

Reducing,

$$0 < 4\,\mu\,(d+\mu)\left(d\mu - \mu^2 + 1 - \mu\right)$$

Dividing out $4\,\mu\,(d+\mu)$ and rearranging,

$$0 < \mu(d-\mu) + 1 - \mu$$

$0 < \mu < d < 1$ due to Assumption 1 and the expression is therefore true. Thus, $\frac{\partial c^*}{\partial d} < 0$. □

References

[1] Axelrod, R., 1984, The Evolution of Cooperation, Basic Books

[2] Baliga, S., Sjöström, T., "Arms Races and Negotiations", *Review of Economic Studies*, Vol. 77 (2004), pp. 351-369.

[3] Baliga, S., Sjöström, T., "The Strategy of Manipulating Conflict", *The American Economic Review*, Vol. 106(2) (2012).

[4] Bas, M. A., Coe, A. J., "Nuclear Russian Roulette: A Model of Proliferation and Preventive War", 2015, Yale University, Working paper.

[5] Calvert, R., 1993, "Communication in Institutions: Efficiency in a Repeated Prisoner's Dilemma with Hidden Information." In W. Barnett, M. Hinich, and N. Schofield, eds., Political Economy: Institutions, Information, Competition, and Representation. Cambridge University Press.

[6] Carlsson, van Damme, 1993, "Global Games and Equilibrium Selection", *Econometrica*, 61, 989-1018.

[7] Chassang, S., Padro i Miquel, G., 2010, "Conflict and Deterrence under Strategic Risk " *The Quarterly Journal of Economics* 125 (4): 1821-58

[8] Debs, A., Monteiro, N., 2014, "Known Unknowns: Power Shifts, Uncertainty, and War.", *International Organization* 68 (1): 1-31.

[9] Downs, George W., David M. Rocke, and Randolph M. Siverson, 1986. Arms Races and Cooperation." In Cooperation under Anarchy Ed. Kenneth A. Oye. Princeton: Princeton University Press.

[10] Fearon, J., "Bargaining Enforcement and International Cooperation", *International Organization*, Vol. 52, No. 2 (Spring, 1998), pp. 269-305

[11] Friedman, J. W., 1971, "A Non-cooperative Equilibrium for Supergames". *Review of Economic Studies* 38 (1): 1-12

[12] Hymans, J, 2006, The Psychology of Nuclear Proliferation, Cambridge University Press

[13] Jervis, R., 1976, Perception and Misperception in International Politics, Princeton University Press.

[14] Jervis, R., 1978, "Cooperation Under the Security Dilemma", *World Politics*, 30, 167-214.

[15] Keohane, R. O., 1984, After Hegemony: Cooperation and Discord in the World Political Economy, Princeton University Press.

[16] Kydd, A., 1997, "Game Theory and the Spiral Model", *World Politics*, 49(3): 371-400.

[17] Morris, S., Shin, H., 2003, "Global Games: Theory and Applications", in M. Dewatripont, L. Hansen and S. Turnovsky (eds.) Advances in Economics and Econometrics (Proceedings of the Eighth World Congress of the Econometric Society) (Cambridge, UK: Cambridge University Press).

[18] Potter, W. C., Mukhatzhanova, G., 2010, Forecasting Nuclear Proliferation in the 21^{st} Century, Role of Theory, Stanford University Press

[19] Schelling, T., 1960, The Strategy of Conflict, Harvard University Press, New ed. 1990

[20] Spaniel, W., 2015, "Bargaining over the Bomb: The Successes and Failures of Nuclear Negotiations", PhD thesis, University of Rochester.

Military Alliances and the Power Curse

Allan Anders Balsgaard*

February 2016

Abstract

This paper studies the strategic contradictions between allied states using a simple dynamic model with incomplete information. We consider a weak state and its stronger ally faced with a common threat to their security. A coordinated response to the threat is preferred by both states, but mistrust makes it difficult for the weaker state to rely on the stronger ally for protection. Therefore, the weaker state may be compelled to take action unilaterally. To forestall the adverse effects of an uncoordinated response, the stronger state may prefer to preempt. Hence, our model describes how a weaker ally can force a strong power to take undue military action. It also provides a theoretical mechanism for the *power curse* in international relations.

*Email: allanbalsgaard@gmail.com

Introduction

In international relations, the phrases 'conflict of interests' and 'strategic contradictions' are mostly used to describe the relationship between a state and its adversary. These phrases, however, can with equal justification be used to characterize the relationship between states allied with one another.

As the proverb 'the enemy of my enemy is my friend' suggests, alliances between states can exist whenever there is an overlap of interests. A military alliance needs not reflect deep friendship and may exist despite considerable contradictions of interest. History provides numerous examples of such alliances. A well-known and compelling example was the alliance between the U.S. and Islamic fighters during the Soviet War in Afghanistan (1979-89). The U.S. provided extensive logistic and military support for the Islamic insurrection in Afghanistan to inflict losses on the Soviet military. The alliance lasted a whole decade – despite the fact that the U.S. and Islamic fighters had widely different goals.

Even among close allies there may be serious conflict of interests as the relationship between NATO members illustrates. Despite having fought on the same side in the War in Afghanistan (2001-14), newspapers were rife with stories of disputes between NATO members over who bears the heaviest burden in Afghanistan in terms of deployment of military personnel[1]. In particular, the U.S. was discontent with the passive role played by the Germans.

The dispute between Israel and the U.S. over Iran's nuclear program is another example of a serious conflict of interests between close allies. Prior to the Iran nuclear deal framework (2015), Israel pressured the U.S. to launch a military strike on Iran's nuclear facilities in order to prevent it from acquiring nuclear weapons. The U.S. on the other hand, was rather reluctant to embark on yet another military adventure in the Middle East. Israel was uncertain whether American promises to carry out a military strike if negotiations failed were credible. The dispute created a serious strain in the alliance, with Israel threatening to attack unilaterally.

[1]Blair, David, "Germany failing to fight Taliban, U.S. Claims", *The Telegraph*, Feb. 1., 2008; Gebauer, Matthias, "NATO Dispute over Afghanistan Forces: It's All Smiles in Vilnius, but only for the Photo Op", *Der Spiegel*, Feb. 8, 2008

Conflicts of interests among allies is a widespread phenomenon and numerous other examples could be given to illustrate this. Previous studies of the strategic contradictions between allies have mainly been concerned with conflicts over burden-sharing and the tendency to free-ride (Sandler & Hartlay, 2001). The War in Afghanistan (2001-14) is a recent example of this. Others have shown how the conflict over burden-sharing can be mitigated through competition among the allies (Niou & Zeigel, 2015). However, to our knowledge, no one has studied how mutual mistrust between the allied states can hamper the ability of these states to coordinate a response to a common threat and how this mistrust can lead to a premature attack.

In this model, there are two allied states – a weak and a strong – facing a common threat to their security. Both states prefer that the stronger state engages the threat, but asymmetry in military capabilities means that they have different timetables. The stronger state prefers to delay direct military engagement for as long as possible to allow for negotiations to take place and possibly succeed. The weaker state can only counter the threat in the short run and therefore has to rely on the stronger state for security in the long run. The stronger state is unable to commit to engaging the threat in the future. This inability to commit to military action gives rise to a strategic dilemma, in which the weak state must decide whether to engage the threat unilaterally or rely on the stronger state to take action later. The stronger state is faced with a dilemma too, in which it must either engage the threat immediately or wait until a later period, hoping that the weaker state does not act on its own. These strategic dilemmas are described using a simple dynamic model with asymmetric information. We derive the conditions for cooperation between the allied states.

Our theoretical framework provides an explanation for the *power curse*. The *power curse* is an interesting stylized fact in international relations according to which strong states easily become burdened by their own military superiority. It is the tendency of a great power to get unduly embroiled in military operations, which drain it for resources and over-stretches its military (Kennedy, 1987). It is the paradox that great powers can seem powerless in certain respects. A weaker

ally can impose its will on the great power by threatening to take matters into its own hands. Therefore, a strong military can be a curse rather than a blessing, as a strong military compels the strong power to get involved militarily against its will (Gallarotti, 2011; Nye, 2003). Hence, our model provides a theoretical underpinning for the power curse.

The paper proceeds as follows: In the following section, we review some related literature. Then, we present the model and derive a pure strategy perfect Bayesian equilibrium, which specifies when cooperation is possible. A number of case studies in support of the model are then provided. The last section summarizes our findings and concludes.

Related Literature

One strand international relations literature has been focusing on arms races. Models of repeated interactions – such as the repeated prisoner's dilemma (Axelrod, 1984) – belong to this category. These models have yielded some important insights about the sustainability of peace and cooperation[2]. Incorporating private information and allowing for negotiations, Baliga and Sjöström (2004) examined the conditions under which an arms race driven by a spiral of fear can be avoided. In other models, additional realism has been added by allowing states to both arm and go to war with one another (Kydd, 1997; Sartori and Meirowitz, 2008; Jackson and Morelli, 2009).

Another strand of international relations literature (Fearon, 1995; Powell, 2006) deals with rationalist explanations of war: Why do states resort to costly fighting, when resources could be divided through peaceful bargaining, which would make both sides better off? Two explanations are suggested: Asymmetric information and lack of commitment. Others argue that agency problems are an important explanation for wars (Jackson and Morelli, 2007). Agency problems arise when the preferences of a country's leader differ from the preferences of the population.

A common feature of the above theories is their focus on two or more military

[2]This paper is inspired by my master thesis *Nuclear Brinkmanship and Preventive War*, 2012, in which I analyzed a related, but simpler model.

opponents and their efforts to attack, coerce and out-arm one another. The focus of our model, on the other hand, is the conflict of interests between allied states and their failure to coordinate on a response to a common enemy. This focus allows us to provide a theoretical mechanism for the power curse.

Our paper also contributes to the literature on nuclear proliferation in international relations theory. Rather than exploring the causes of nuclear proliferation or what can be done to prevent it (Sagan, 1997), our paper examines how other states react to it and the circumstances under which states are likely to resort to preventive war to counter a threat.

Model

Two allied states – state S (the stronger state) and state W (the weaker state) wish to coordinate their attack on a common adversary. We use a simple dynamic model with two-sided asymmetric information to represent the strategic dilemmas of the states. Before the game starts, nature determines the types of state S and state W. S is a *hawk* (denoted an S-hawk) with probability h_S and a *dove* (denoted an S-dove) with probability $1 - h_S$ where $0 < h_S < 1$. W is a *hawk* (denoted a W- hawk) with probability h_W and a *dove* (denoted a W-dove) with probability $1 - h_W$, where $0 < h_W < 1$. Types are private information.

In period 1 of the game, S can either *attack* or *wait*. If S attacks, the game ends, S receives the payoff A^1_{SH} if he is a hawk and A^1_{SD} if he is a dove. W receives the payoff A^1_{WH} if he is a hawk and A^1_{WD} if he is a dove. If S waits, W observes S's choice and decides whether to attack or wait. If W attacks, the game ends, W receives the payoff B_{WH} if he is a hawk and B_{WD} if he is a dove. S receives the payoff B_{SH} if he is a hawk and B_{SD} if he is a dove. If W waits, the game proceeds to period 2, in which only S is capable of attacking. In period 2, S observes W's choice and decides whether or not to attack. If S attacks, S receives the payoff A^2_{SH} if he is a hawk and A^2_{SD} if he is a dove. W receives the payoff A^2_{WH} if he is a hawk and A^2_{WD} if he is a dove. If S does not attack there is peace and S receives the payoff P_{SH} if he is a hawk and P_{SD} if he is a dove. W receives the payoff P_{WH}

if he is a hawk and P_{WD} if he is a dove.

W ranks the different outcomes in the following way:

$$P_{WH} < B_{WH} < A^1_{WH} < A^2_{WH} \tag{1}$$

$$B_{WD} < P_{WD} < A^1_{WD} < A^2_{WD} \tag{2}$$

Both of W's types agree that an attack by S in period 2 is better than an attack by S in period 1 since the common adversary may decide to acquiesce to the demands of S and W through negotiations as time passes. Both of W's types also agree that an attack by S in period 1 is better than an attack by W. W is not as powerful militarily as S and an attack by W is more likely to fail. However, they disagree whether an attack by W is better than no attack at all. W-hawks prefer an attack by W over no attack at all and W-doves prefer no attack at all over an attack by W.

S ranks the different outcomes in the following way:

$$P_{SH} < A^1_{SH}, B_{SH} < A^1_{SH} < A^2_{SH} \tag{3}$$

$$B_{SD} < A^1_{SD} < A^2_{SD} < P_{SD} \tag{4}$$

Both of S's types agree that an attack by S in period 2 is better than an attack by S in period 1 since the passing of time may get the adversary to acquiesce through peaceful negotiations. Both of S's types also agree that an attack by S in period 1 is better than an attack by W since S fears that an unilateral attack by W may fail and lead to unwanted consequences. However, they disagree on whether an attack by S is better than no attack at all. S-hawks prefer an attack by S over no attack at all and S-doves prefer no attack at all over an attack by S.

Analysis

We let p denote the probability with which S attacks in period 2 and q the probability with which W attacks in period 1. Hence, $p = 1$ means that S is certain to attack in period 2. b is W's belief of S's type where b is the probability that S is a hawk and $0 \leq b \leq 1$

Types are private information. Everything else is common knowledge. Figure 1 provides an illustration of the game.

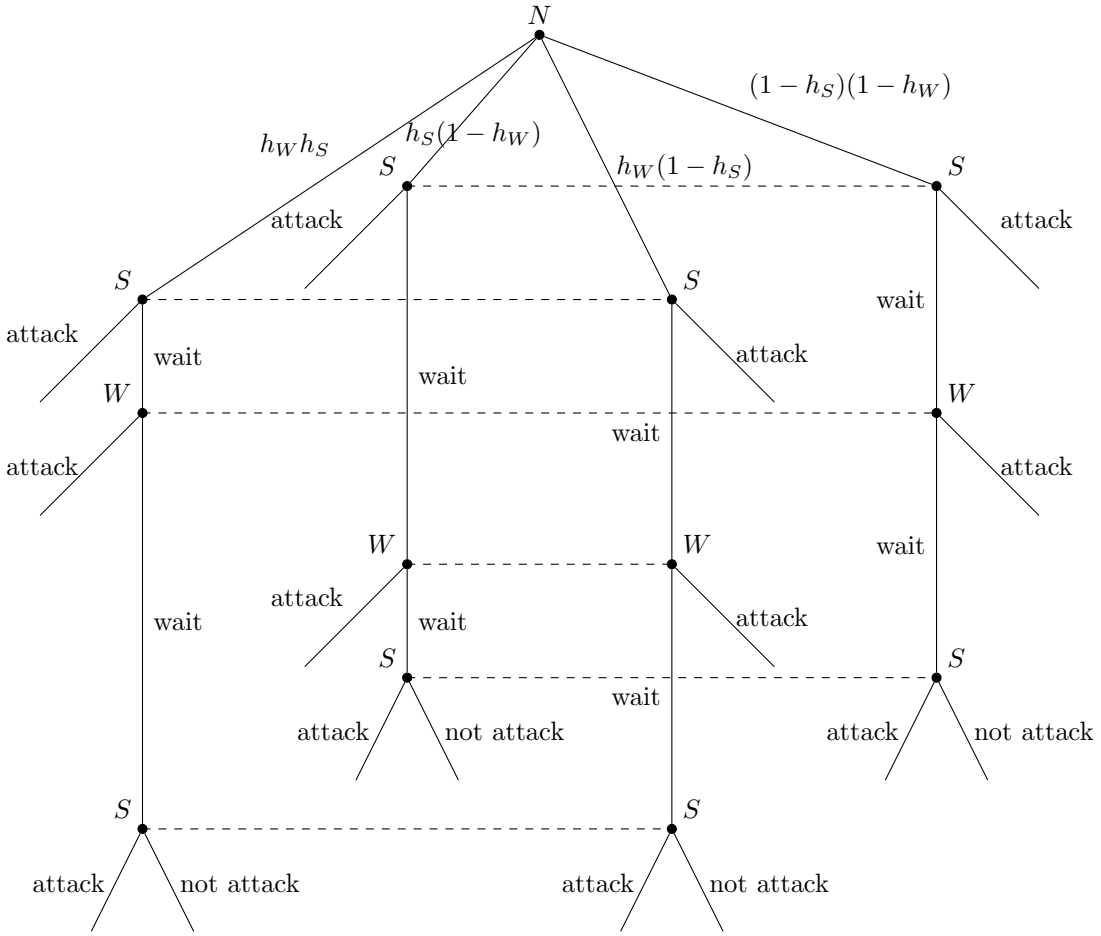

Figure 1: **Game Tree**

We determine $\hat{q}_H$ – the probability that W attacks unilaterally, which makes an S-hawk indifferent between attacking in period 1 and waiting.

$$\hat{q}_H B_{SH} + (1 - \hat{q}_H) A^2_{SH} = A^1_{SH} \tag{5}$$

The left hand side is S's expected payoff from waiting if W attacks with probability $\hat{q}_H$. The right hand side is S's payoff from attacking in period 1. Isolating $\hat{q}_H$,

$$\hat{q}_H = \frac{A^2_{SH} - A^1_{SH}}{A^2_{SH} - B_{SH}} \tag{6}$$

Observe that $0 < \hat{q}_H < 1$ since $A^2_{SH} - A^1_{SH} < A^2_{SH} - B_{SH}$ due to S's preferences (3). If $q < \hat{q}_H$, S waits in period 1, whereas S attacks if $q > \hat{q}_H$. If $q = \hat{q}_H$, S is indifferent between attacking and waiting, but for simplicity we assume that S attacks. Thus, S attacks if and only if $q \geq \hat{q}_H$.

We determine $\hat{q}_D$ – the probability that W attacks, which makes an S-dove indifferent between attacking in period 1 and waiting.

$$\hat{q}_D B_{SD} + (1 - \hat{q}_D) P_{SD} = A^1_{SD} \tag{7}$$

Isolating $\hat{q}_D$,

$$\hat{q}_D = \frac{P_{SD} - A^1_{SD}}{P_{SD} - B_{SD}} \tag{8}$$

Observe that $0 < \hat{q}_D < 1$ since $P_{SD} - A^1_{SD} < P_{SD} - B_{SD}$ due to S's preferences (4). If $q < \hat{q}_D$, S waits in period 1, whereas S attacks if $q > \hat{q}_D$. If $q = \hat{q}_D$, S is indifferent between attacking and waiting, but for simplicity we assume that S attacks. Thus, S attacks if and only if $q \geq \hat{q}_D$.

We can distinguish between two cases depending on the relative size of $\hat{q}_H$ and $\hat{q}_D$. If $\hat{q}_H < \hat{q}_D$, S-hawks are more likely to attack in period 1 than S-doves are given that W attacks unilaterally with a given probability. In this sense, S-hawks are more eager to attack than S-doves are – not only in period 2 – but also in period 1. On the other hand, if $\hat{q}_D < \hat{q}_H$, S-doves are more eager to attack in period 1 given than S-hawks are. Given a probability that W attacks unilaterally, S-doves are more likely to attack than S-hawks are. This case seems less intuitive than the previous. Whereas only S-hawks prefer to attack in period 2, S-hawks are less likely to attack in period 1 than are S-doves. Hence, depending on the ranking of $\hat{q}_H$ and $\hat{q}_D$, S-doves may behave more aggressively in period 1 than S-hawks.

We determine $\hat{p}_H$ – the probability that S attacks in period 2, which makes a W-hawk indifferent between attacking unilaterally and waiting.

$$\hat{p}_H A^2_{WH} + (1 - \hat{p}_H) P_{WH} = B_{WH} \tag{9}$$

Isolating $\hat{p}_H$,

$$\hat{p}_H = \frac{B_{WH} - P_{WH}}{A^2_{WH} - P_{WH}} \tag{10}$$

Observe that $0 < \hat{p}_H < 1$ since $B_{WH} - P_{WH} < A^2_{WH} - P_{WH}$ due to W's preferences (1). If $p < \hat{p}_H$, W attacks preemptively in period 1, whereas W waits if $p > \hat{p}_H$. If $p = \hat{p}_H$, W is indifferent between attacking and waiting, but for simplicity we assume that W attacks. Thus, W attacks if and only if $p \leq \hat{p}_H$.

Equilibrium

We solve for perfect Bayesian equilibria since the model is dynamic with asymmetric information. A strategy profile and a set of beliefs are a perfect Bayesian equilibrium if strategies are sequentially rational given beliefs and beliefs are updated using Bayes' rule. In the following, we explore different combinations of pure strategies in order to determine which of them are perfect Bayesian equilibria. As we will see, the prior beliefs h_S and h_W are crucial for S and W's decision to attack in period 1.

We let $\{(a, a), (a, w)\}$ be a short-hand notation for a set of strategies in which S attacks in period 1 and W only attacks if he is a *hawk*. Thus, the first vector (a, a) in the set denotes the strategies chosen by an S-hawk and an S-dove in period 1, respectively. And the second vector (a, w) in the set denotes the strategies chosen by a W-hawk and a W-dove in period 1, respectively. To keep things simple, S's strategy in period 2 is not included in the notation since the strategy is exclusively determined by S's type. Also note that there is no strategic trade-off for a W-dove since a W-dove never wants to attack unilaterally.

Proposition 1. *The following gives a characterization of the pure strategy perfect Bayesian equilibria of the game given the prior beliefs h_S and h_W. The first scheme is for $\hat{q}_H < \hat{q}_D$ and the second for $\hat{q}_D < \hat{q}_H$.*

$\hat{q}_H < \hat{q}_D$	$0 < h_S \leq \hat{p}_H$	$\hat{p}_H < h_S < 1$
$\hat{q}_D \leq h_W \leq 1$	$\{(a,a),(a,w)\}$	$\{(a,a),(a,w)\}$, $\{(w,w),(w,w)\}$
$\hat{q}_H \leq h_W < \hat{q}_D$	$\{(a,w),(a,w)\}$	$\{(a,w),(a,w)\}$, $\{(w,w),(w,w)\}$
$0 \leq h_W < \hat{q}_H$	$\{(w,w),(a,w)\}$	$\{(w,w),(w,w)\}$

$\hat{q}_D < \hat{q}_H$	$0 < h_S \leq \hat{p}_H$	$\hat{p}_H < h_S < 1$
$\hat{q}_H \leq h_W \leq 1$	$\{(a,a),(a,w)\}$	$\{(a,a),(a,w)\}$, $\{(w,w),(w,w)\}$
$\hat{q}_D \leq h_W < \hat{q}_H$		$\{(w,w),(w,w)\}$
$0 \leq h_W < \hat{q}_D$	$\{(w,w),(a,w)\}$	$\{(w,w),(w,w)\}$

Proof. Solving the game backwards, we first describe the behavior in period 2. Subsequently, we describe the behavior in period 1.

Period 2

In period 2, S decides whether or not to attack. By (3), we know that attacking is a dominant strategy for S-hawks. And by (4) we know that S-doves prefer not to attack. Hence, the probability that S attacks in period 2 is the probability that S is a hawk.

Period 1

In the following, we analyze the different types of strategies and determine the parameter values in the (h_S, h_W)-space for which a given strategy is valid. For each strategy, we show that S's strategy is a best response to W's strategy and vice versa for each set of prior beliefs.

$\{(a,a),(a,w)\}$:

S: Since only W-hawks attack, the probability that W attacks is h_W. Therefore, both types of S only attack if $\hat{q}_D \leq h_W$ and $\hat{q}_H < h_W$.

W: If both types of S attack in period 1, the game never reaches the information set where W decides whether to attack or wait. Consequently, when $b \leq \hat{p}_H$, it is optimal for a W-hawk to attack in period 1. Hence, attacking is only optimal for a W-hawk if he is sufficiently convinced that S is not going to attack in period 2.

Thus, $\{(a,a),(a,w)\}$ is part of a perfect Bayesian equilibrium whenever $\hat{q}_D \leq h_W$, $\hat{q}_H < h_W$ and $b \leq \hat{p}_H$.

$\{(a,w),(a,w)\}$:

We show that S's strategy is a best response given W's strategy and that W's strategy is a best response given S's strategy. Notice that for this particular type of strategy, it is crucial whether $\hat{q}_H < \hat{q}_D$ or $\hat{q}_D < \hat{q}_H$.

- $\hat{q}_H < \hat{q}_D$, S: Only W-hawks attack. Thus, the probability that W attacks is equal to the probability that W is a hawk, h_W. Therefore, S-hawks attack and S-doves wait if and only if $\hat{q}_H \leq h_W < \hat{q}_D$.
- $\hat{q}_H < \hat{q}_D$, W: Since only S-hawks attack in period 1, W is certain that S is a dove if S waited, i.e. $b = 0$. Therefore, W is certain that S will not attack in period 2. Consequently, W-hawks attack in period 1.

Thus, $\{(a,w),(a,w)\}$ is part of a perfect Bayesian equilibrium whenever $\hat{q}_H \leq h_W < \hat{q}_D$. Notice that $\{(a,w),(a,w)\}$ cannot be an equilibrium strategy if $\hat{q}_D < \hat{q}_H$. Since S-hawks attack and S-doves wait in period 1, we must have that $\hat{q}_H < h_W$ and $h_W < \hat{q}_D$ at the same time. This is clearly in contradiction with $\hat{q}_D < \hat{q}_H$. Hence, $\{(a,w),(a,w)\}$ cannot be part of a perfect Bayesian equilibrium.

$\{(w,a),(a,w)\}$: This strategy set cannot be part of a perfect Bayesian equilibrium. Since only S-doves attack in period 1, W is sure that S is a hawk if S waited, i.e. $b = 1$. Therefore, W waits since S is certain to attack in period 2. Hence, $\{(w,a),(a,w)\}$ cannot be part of an equilibrium.

$\{(w,w),(a,w)\}$: Again, we show that S's strategy is a best response given W's strategy and that W's strategy is a best response given S's strategy.

S: Given that W-hawks attack, S is only willing to wait if the share of W-hawks is sufficiently low, i.e. if $h_W < \hat{q}_H$ and $h_W < \hat{q}_D$.

W: Neither of S's type attacks in period 1. Using Bayes' rule, W's belief in period 1 is $b = h_S$. Hence, W-hawks attack preemptively in period 1 only if the share of S-hawks is low enough $h_S \leq \hat{p}_H$.

Thus, $\{(w,w),(a,w)\}$ is part of a perfect Bayesian equilibrium whenever $h_S \leq \hat{p}_H$, $h_W < \hat{q}_H$ and $h_W < \hat{q}_D$.

$\{(w,w),(w,w)\}$: Again, we show that S's strategy is a best response given W's strategy and that W's strategy is a best response given S's strategy.

S: Given that W never attacks in period 1, S prefers to wait regardless of W's type, i.e. regardless how large h_W is.

W: S waits in period 1. Using Bayes' rule, W's belief in period 1 thus becomes $b = h_S$. Hence, W-hawks only wait in period 1 if the share of S-hawks is sufficiently high, i.e. $h_S > \hat{p}_H$.

Thus, $\{(w,w),(w,w)\}$ is part of a perfect Bayesian equilibrium whenever $h_S > \hat{p}_H$.

$\{(a,a),(w,w)\}$: This strategy set cannot be part of a perfect Bayesian equilibrium since attacking in period 1 can never be optimal for S if W is certain to wait. Anticipating that W is going to wait, S instead prefers to wait.

$\{(a,w),(w,w)\}$: This strategy set also cannot be a part of a perfect Bayesian equilibrium since attacking in period 1 can never be optimal for an S-hawk if W is certain to wait.

$\{(w,a),(w,w)\}$: This strategy set also cannot be a part of a perfect Bayesian equilibrium since attacking in period 1 can never be optimal for an S-dove if W is certain to wait. □

Our analysis shows how the likelihood of a preemptive attack by W is affected by the share of S-hawks. If h_S is low, S is unlikely be a hawk and thus unlikely to attack in period 2. Therefore, W-hawks are unwilling to rely on S and attack unilaterally. If W-hawks are sufficiently common, the fear of an attack by W forces S to preempt in period 1. Hence, if h_S is low and h_W is high, S attacks in period 1. This corresponds to the equilibrium, in which $\{(a,a),(a,w)\}$ is an equilibrium strategy. In this equilibrium, there is an attack in period 1 despite that neither of the states wants it. Both states prefer that S attacks in period 2 rather than in period 1, but this Pareto-improvement is not possible. S prefers to delay an attack until period 2, but the risk of an attack by W forces state S to attack preemptively in period 1. Thus, mutual suspicion leads to an inefficient outcome with a premature attack.

If h_W is also low and W-hawks therefore are less common, S is willing to wait despite the risk that W attacks unilaterally. In this equilibrium, the outcome is not predetermined. Sometimes W preempts in period 1 and sometimes the game proceeds to period 2 where S attacks if he is a hawk and does not attack if he is a dove. This corresponds to the equilibrium strategy $\{(w,w),(a,w)\}$.

Now let us increase h_S so that S is very likely to be a hawk. When h_S is high, W is more confident that S will attack in period 2. This confidence in S makes W willing to wait in period 1 rather than attack unilaterally. Anticipating W's decision to wait, S also decides to wait in period 1 rather than to attack. Hence, if hawks are sufficiently common among S, there is no preemptive attack in period 1, both states wait and S attacks in period 2 if he is a hawk and does not attack if he is a dove. This corresponds to the equilibrium, in which $\{(w,w),(w,w)\}$ is an equilibrium strategy. Notice that the share of W-hawks is not important for the feasibility of this outcome. Since W prefers to wait regardless of type, the share of W-hawks does not affect S's decision to attack or wait.

The equilibrium strategy, in which $\{(a,a),(a,w)\}$ is chosen, provides an illus-

tration of the power curse, because in this equilibrium the weaker state forces the stronger state to attack prematurely. The weaker state fears to leave the responsibility for attacking the common enemy to the stronger state since the stronger state may turn out to be a dove that does not attack. If the mistrust to the stronger state is sufficiently strong, the weaker state decides to attack unilaterally instead of waiting for the stronger state to do the job in a later period. Anticipating the weaker state's preemption, the stronger state is forced to attack preemptively in order to prevent the weaker state from initiating an attack that might fail or lead to an undesired result. Therefore, we speak of the power curse. The weaker state forces the stronger state to engage in military endeavors that it had preferred to avoid altogether. Hence, possessing a powerful military can force a powerful state to undertake undue military engagements.

Case Studies

In the following, case studies from recent history are presented to demonstrate the ability of our model to account for real world events. The first example concerns the dispute between the U.S. and Israel over the timing of a preventive military strike on Iran's nuclear program. The subsequent examples deal with the relationship between the U.S. and its allies during the Cold War.

Destroying Iran's nuclear program

The Iran nuclear deal framework that was signed by Iran and the U.S. in spring 2015 ended a long crisis over Iran's nuclear program. Before the Iran nuclear deal framework, there were considerable tensions between U.S. and Israel. These close allies found it hard to coordinate their actions and make a proper response to Iran's attempt to acquire nuclear weapons.

In November 2011, the IAEA (International Atomic Energy Agency) released a report stating that Iran had been conducting experiments aimed at acquiring a nuclear weapons capability[3]. Since this revelation, Iran's nuclear program became

[3]Implementation of the NPT Safeguards Agreement and relevant provisions of Security Coun-

an important topic in the international community. The U.S., Israel and the EU imposed harsh sanctions on Iran in order to thwart an Iranian bomb. Israel was especially fierce in its rhetoric. At the AIPAC conference in 2012, the Israeli Premier, Netanyahu, gave a speech in which he compared Iran's nuclear program to Holocaust and called for military action (March 5, 2012). Later in an interview, he said:

> If you don't make the decision and don't succeed in preventing this [an Iranian nuke], to whom will you explain this – to the historians? To the generations before you, and the generations that won't come after you?[4]

By comparing the threat of an Iranian nuclear weapon to the systematic extermination of Jewish population during WWII by Nazi Germany, Netanyahu had put himself in a situation from which there was no retreat[5]. From that point on, to refrain from military action would have made him appear weak and severely undermined his credibility.

It was unclear whether Iran was prepared to back down and whether Netanyahu's statement did have any effect. On the one hand, the Iranian regime appeared relatively undeterred. They condemned Israel's threats and threatened retaliation if attacked[6]. The nuclear program represented a considerable investment for them and they were very unwilling to back down. A nuclear weapons capability would provide the Iranian regime with enormous strategic advantages. Nuclear weapons would allow the regime to project power throughout the region and, furthermore, be a guarantee against foreign intervention and attempts of regime change.

---

cil resolutions in the Islamic Republic of Iran, *IAEA*, November 18, 2011

[4]Lis, "Netanyahu: Strike on Iran's nuclear facilities possible within months", *Haaretz*, March 9, 2012

[5]Aluf Benn,"By conjuring the Holocaust, Netanyahu brought Israel closer to war with Iran", *Haaretz*, March 6, 2012

[6]Ravid, "Iran: Israel 'a barking dog' that will not dare attack Islamic Republic over nuclear program", *Haaretz*, March 17, 2012.

While cooperation and coordination with the U.S. was preferred, Israel wished for an airstrike to be carried out as soon possible due to the limits of Israel's military capabilities. Iran was engaged in an effort to disperse and fortify its nuclear facilities and the Israeli's ability to cause significant damage on the Iranian nuclear program was therefore gradually being diminished. The U.S. on the other hand was untouched by such limitations and therefore preferred to delay an attack for as long as possible, hoping that negotiations would succeed to halt the program. Both Israel and the U.S. preferred that an attack was launched by the U.S. A U.S. attack would be more likely to destroy the program. An Israeli attack would only have caused a moderate setback in the nuclear program. Furthermore, an Israeli attack on Iran would have been perceived as yet another aggression by Jews on Islam and therefore could have rejuvenated anti-Israeli sentiments among muslims throughout the Middle East. Moreover, an Israeli attack would almost certainly have resulted in Iranian retaliation against American forces in the Persian Gulf and beyond. If Israel had acted on its own, the U.S. would have been drawn into the conflict anyway. Hence, both countries preferred that an attack against the Iranian nuclear program was to be carried out by the U.S.

However, delegating all responsibility for destroying the Iranian nuclear program to the U.S. was problematic due to high levels of mutual distrust. Israel was uncertain whether it could rely on the Obama administration for protection. In Israel, Obama – a Nobel Peace Prize laureate .- was perceived to be a weak and unable contemplate the use of military power. This put Israel in a tough dilemma: Launch an attack itself or rely on Obama's promise to carry out an attack later? The Obama-administration was faced with a dilemma too: Attack now or wait and hope that the Israelis do not lose their nerve and launch an attack independently?

The crisis over Iran's nuclear program is compelling example of how a smaller ally tried to used the threat of attacking unilaterally to force the stronger ally to launch a military strike.

As it turned out, the U.S. decided to wait and negotiate instead of attacking. As a consequence hereof Israel came very close to launching an attack independently

in autumn 2012[7].

Operation Nickel Grass

In the 1973 Arab-Israeli war, Egypt and Syria forces launched a surprise attack on Israel while Israel was celebrating Yom Kippur, the holiest day in Judaism. During the first week of the fighting, Israel suffered heavy casualties and failed to repel the attackers. Fearing that the invading Egyptian and Syrian forces would overrun their country, the government deployed its nuclear arsenal as a weapon of last resort. These preparations were detected by the U.S. which responded by commencing an airlift to replace Israel's material losses. This became known as *Operation Nickel Grass*. In this way, the risk that Israel would use nuclear weapons forced the Americans to get involved, despite that the U.S. – fearing an Arab oil boycott – preferred to remain on the sideline (Cohen, 1999). Had Israel really been on the verge of total defeat, the U.S. probably would have intervened, but mutual mistrust meant that Israel was unwilling to put its destiny in the hands of the U.S. Hence, Israel's threat to escalate the conflict forced the U.S. to become deeply involved at a very early point in the conflict. Note that the opposite happened in the crisis over the Iranian nuclear program where the U.S. did not become involved. Instead, the U.S. eventually managed to diffuse the threat through negotiations. Whether Israel would have been forced to consider the use of nuclear weapons without American support is an open question. But by supporting Israel directly, the U.S. ran a high risk of drawing the Soviet Union into the conflict and provoking an oil boycott by the Arab world.

South Korean Nuclear Program

After the end of the Korean War in 1953, which divided Korea in two, the U.S. kept a large number of American soldiers in South Korea in order to deter North Korea from attempting to invade South Korea. However, the American defeat in Vietnam in 1975 led the U.S. to scale down its military presence in East and

[7]George Friedman, "Israel: The case against attacking Iran", Geopolitical Weekly, August 25, 2015, Stratfor.

Southeast Asia – including the Korean peninsula. In 1971, the withdrawal of 26.000 U.S. soldiers from South Korea made Seoul uncertain whether the U.S. would come to the rescue of South Korea if North Korea invaded again. South Korea responded to the withdrawal of American troops by attempting to acquire a plutonium reprocessing plant, which could have been used for building nuclear weapons. The uncertainty about the American commitment to defend South Korea led to South Korea's attempt to acquire its own nuclear weapons capability. In this way, South Korea attempted to force the U.S. to keep its military presence in South Korea. However, the U.S. managed to evade South Korea's pressure. The pressure exerted by the U.S. on the French company to cancel the delivery of the plutonium reprocessing plant eventually convinced the leadership in Seoul to abandon the nuclear program (O'Neil, 2013).

Conclusion

This paper studies the strategic contradictions between allied states, faced with a threat to their security. The states are distinguished by the level of military strength. The stronger ally can afford to delay engagement of the threat to allow for negotiations to take place, whereas the weaker ally is much less patient because it cannot cope effectively with the threat in the long run. Both states prefer that the stronger state engages the threat and the weaker state is therefore forced to rely on the stronger state if the threat is not engaged immediately and the crisis draws out. If the level of mistrust between the states is sufficiently high, a strategic dilemma of commitment is created. Whereas the stronger state fears that the weaker state will attempt to engage the threat unilaterally and prematurely, the weaker state fears that if it relies on the stronger state for protection, engagement of the threat may never take place.

We formalize this strategic dilemma using a simple dynamic model with asymmetric information. Solving the model for perfect Bayesian equilibria, we identify the conditions under which the allied states fail to make a coordinated response to the threat. Our model describes how a weaker, mistrustful ally is able to impose its strategic goals on its stronger ally. The prospect of a unilateral attack by the

weaker state compels the stronger ally to engaged the threat prematurely or in an unwarranted degree. Hence, our model provides theoretical underpinning for the power curse. In international relations, being a great power often can be a curse rather than a blessing. With great power comes great responsibility. And a large responsibility can compel states to become excessively involved in military conflicts.

References

[1] Axelrod, R., 1984, *The Evolution of Cooperation*, New York: Basic books.

[2] Baliga, S., Sjöström, T., Arms Races, *Review of Economic Studies*, (2004) 71, 351369

[3] Cohen, A., 1999, *Israel and the Bomb*, Columbia University Press.

[4] Gallarotti, G. M., "Soft power: What it is, why it's important, and the conditions for its effective use", *Journal of Political Power*, 2011, 4:1, 25–47.

[5] Gibbons, R., 1992, *A Primer in Game Theory*, Harvester Wheatscheaf, New York.

[6] Jackson, M. O., Morelli, M., Political Bias and War, *The American Economic Review*, 2007, Vol. 97, No. 4.

[7] Jackson, M. O., Morelli, M., Strategic Militarization and Wars, *Quarterly Journal of Political Science*, 2011, 4: 279-313

[8] Kennedy, P., 1989, *The Rise and Fall of the Great Powers*, Fontana Press

[9] Kydd, A., Game Theory and the Spiral Model, *World Politics*, 1997, Vol. 49, No. 3, 371-400

[10] Meirowitz, A., Sartori, A. E., Strategic Uncertainty as a Cause of War, *Quarterly Journal of Political Science*, 2008, 3: 327-352

[11] Niou, E., Zeigler, S., External Threats and Internal Rivalries: The Dynamics Alliance Formation. Paper presented at the American Political Science Association Conference 2015.

[12] Nye, J., 2005, *Soft Power: The Means to Success in World Politics*, Public Affairs.

[13] O'Neil, A., 2013, *Asia, the U.S. and Extended Nuclear Deterrence: Atomic Umbrellas in the Twenty-First Century*, Routledge.

[14] Powell, R., War as a Commitment Problem, *International Organization*, 2008, 3: 327-352

[15] Sandler, T., Hartlay, K., 2001, Economics of Alliances: The Lessons for Collective Actions. *Journal of Economic Literature*, Vol. XXXIX (September 2001), pp. 869896.

[16] Sagan, S., Why Do States Build Nuclear Weapons?: Three Models in Search of a Bomb, *International Security*, 1997, Vol. 21, No. 3, pp. 54–86.

[17] Snyder, G. "'Prisoner's dilemma' and 'Chicken' Models in International Politics.", *International Studies Quarterly*, 1971, 15, 66–103.

Does Better Information Make War Less Likely?

Allan Anders Balsgaard* Thomas Jensen†

February 2016

Abstract

We explore how the probability of war due to asymmetric information depends on the quality of information that states receive or collect about their opponents. In a version of the standard ultimatum bargaining model with private information we show that it is possible for the probability of war to increase when the incompletely informed state receives more precise (less noisy) information. Thus, better information does not necessarily make war less likely.

*University of Copenhagen. Email: allan.anders.balsgaard@econ.ku.dk
†University of Copenhagen. Email: thomas.jensen@econ.ku.dk

Introduction

Asymmetric information is one of the fundamental rationalist explanations of why states end up in costly war (Fearon, 1995). If two states bargain over the division of a contested resource and state A is uncertain about which outcomes state B is willing to accept, then A is facing a trade-off between getting as good a deal as possible and minimizing the risk of war. Therefore, A's optimal bargaining strategy may well lead to a positive probability of war. Without uncertainty about the opponent, a bargaining agreement that both states prefer to war can always be reached.[1]

In this paper we explore how the probability of war due to asymmetric information depends on the quality of information that states collect or receive about their opponents. Information about other states' military capabilities and the beliefs and intentions of their leaders will generally be noisy, but improved intelligence capabilities and international institutions can reduce the noise and thus make states better informed. We ask the question if less noisy information will always make war less likely.

We know that the probability of war is zero in the limit where states are completely informed. Thus, immediate intuition suggests that the probability of war will simply decrease (at least weakly) as states become better informed. The main message of this paper is that this intuition is not always correct. Reducing the level of noise in a state's information about its opponent will sometimes increase the probability of war.

The basis for our analysis is a simple version of the standard ultimatum bargaining model with private information about the costs of war (Fearon, 1995). Two states, A and B, bargain over a contested resource. A proposes a split of this resource, B can either accept this proposal or go to war. A is incompletely informed about B's cost of war, which can be low, medium, or high. We extend this model by letting A receive a noisy signal about B's cost. The quality of the signal is parametrized such that we can continuously move from the case of pure noise all the way to a fully informative signal, which is of course equivalent to

[1] Assuming that no other causes of war, such as commitment problems or indivisibility of the contested resource, are present.

complete information.

Within this model, we consider how the ex ante probability of war changes with signal quality. Our results identify several different situations where the probability of war can increase when A's signal become more precise (less noisy). In particular, suppose the probability of war is as high as it can be in the model when A's signal is pure noise. Thus, as we increase the signal quality, the probability of war must fall from the highest possible level to zero in the complete information limit. Even in this case the probability of war will not always be weakly decreasing in signal quality, i.e., increased signal quality will sometimes lead to a higher probability of war.

The main reason behind our findings is that when a state's information about its opponent is less noisy then it is more confident that the received information is correct. For example, if A's intelligence indicates that B has a high cost of war then A will believe more in this information the better its intelligence capabilities are. This is the case even when the intelligence happens to be wrong, which is less likely with better intelligence capabilities, but still possible. In other words, with less noisy information A is less likely to have "wrong" beliefs, but when beliefs are wrong then they are more so. This can cause A to make offers that make war more likely. It is important to note that A is a fully rational Bayesian updater. Its confidence in the received information is completely justified and our results do not in any way depend on irrational overconfidence or other behavioral biases.

A few other papers have studied theoretically how the probability of war varies with states' uncertainty about their opponents. As in the present paper, Reed (2003) uses a simple ultimatum bargaining model with private information[2] to explore how the bargaining outcome and the probability of war depends on the level of uncertainty for the incompletely informed state. A main result is that more precise information always leads to a lower probability of war. This seems to be in stark contrast with our findings. The reason is that Reed compares situations where the uncertainty of the incompletely informed state are given by distributions that differ in variance but has identical mean values. Thus, the results are better

[2]The challenger (state A) is incomplete informed about the probability of winning the war (p) rather than the cost of war for the defender (state B) as in our model. However, this difference in assumptions between Reed's model and ours is not what drives the difference in results.

suited for comparing different crises with varying levels of uncertainty (which is also the main purpose) rather than exploring how more precise information in a given crisis will affect the probability of war. The latter requires an explicitly defined information structure and that updating after all possible signals is considered, which makes the assumption of a fixed mean very restrictive.

Wittman (2009) presents a result similar to Reed's in a case with two-sided incomplete information and a double auction setup for the bargaining protocol. Schub (2015) uses a setup similar to Reed's as the starting point for studying the contrast between the consequences of lower uncertainty when it is due to perceptual errors as opposed to rational information processing. A main theoretical finding is that lower uncertainty due to rational updating always leads to a lower probability of war (as in Reed, 2003), while war may become more likely when uncertainty decreases because of perceptual errors.[3]

Finally, Kurazaki (2015) studies a model of signaling and misperception in crises and shows that the probability of war will in some cases increase with the (exogenous) probability that a signal is perceived correctly. Despite the fact that states are fully rational in our model, this result is related to our findings because the possibility of misperception in Kurazaki's model plays a somewhat similar role to that of noisy signals in our model. Still, the results are not directly comparable because the models are distinct in several ways, for example there is no endogenous signaling stage in our model and Kurazaki considers only binary choices for each state at each stage of the game.

The Model

Consider first a simple version of the standard ultimatum bargaining model with private information about the costs of war (Fearon, 1995). Two states, A and B, bargain over a continuously divisible resource of value 1. A proposes a split of this resource, $x \in [0, 1]$ for itself and $1 - x$ for B. B can either accept this proposal or go to war. The winner of the war gets all of the resource. A wins with probability

[3] Further, if the probability of war decreases with lower uncertainty due to perceptual errors then it will do so at a lower rate than if the reduction in uncertainty is due to rational updating.

$p \in (0,1)$ and B wins with the residual probability $1-p$.

War is costly. A's cost of war is $c_A > 0$. B's cost of war can be low ($c_B^L > 0$), medium ($c_B^M > c_B^L$), or high ($c_B^H > c_B^M$). B knows its own cost when it decides whether to accept A's proposal or go to war. When making its proposal, A knows only that the probabilities of the three possible costs for B are $q_L, q_M, q_H > 0$ with $q_L + q_M + q_H = 1$.

Each state is risk neutral, so its final utility is equal to its share of the resource if the proposal is accepted and to its probability of winning minus its cost of war if not. We assume that B will accept the proposal if indifferent. All aspects of the game except B's realized cost of war is common knowledge.

Solving this model is straightforward. To avoid corner solutions, assume that $c_B^H < 1-p$. Suppose B is type $t \in \{L, M, H\}$, which means that its cost of war is c_B^t. Then it will accept A's offer precisely if $1-x \geq 1-p-c_B^t$, which is equivalent to

$$x \leq p + c_B^t.$$

If a type t accepts A's proposal then all higher types will also accept. It is easy to see that only three proposals are relevant for A. Either it will propose $x^L = p + c_B^L$ (which all types will accept), $x^M = p + c_B^M$ (which the types M and H will accept), or $x^H = p + c_B^H$ (which only type H will accept). A's expected final utility from each of these proposals are

$$u_A(x^L) = p + c_B^L, \tag{1}$$

$$u_A(x^M) = (q_H + q_M)(p + c_B^M) + q_L(p - c_A), \text{ and} \tag{2}$$

$$u_A(x^H) = q_H(p + c_B^H) + (q_L + q_M)(p - c_A). \tag{3}$$

A will make the proposal that maximizes its expected final utility. In case of equality we assume that A will choose the optimal proposal that leads to the lowest probability of war, i.e., the lowest optimal x.

As is well known, this model highlights the trade-off for A between reaching a better bargaining deal and risking costly war. If, among the relevant proposals, A offers less for B, then it will be better off if B accepts but run a higher risk of ending in war. The equilibrium probability of war will be positive if x^L is not an

optimal proposal for A, i.e., if $u_A(x^L) < \max\{u_A(x^M), u_A(x^H)\}$. For example, this is the case if c_A and c_B^L are sufficiently close to zero.[4]

Introducing an Informative Signal

As explained in the introduction, the main purpose of this paper is to explore how the probability of war depends on the quality of states' information about their opponents. Therefore we now extend the simple model presented above by letting A receive an informative but noisy signal about B's type. We will parametrize the quality of the signal, such that we can continuously vary it from being not informative at all (pure noise) to fully informative. Thus we can track the probability of war as we start in a situation where A has only its ex ante information given by the probabilities q_L, q_M, q_H and then gradually approach the complete information benchmark where war is not possible.

To keep matters simple, we will restrict attention to a situation where the ex ante belief of A is that all types of B are equally likely. That is, $q_L = q_M = q_H = \frac{1}{3}$. Before making its proposal, A receives a signal $s \in \{l, m, h\}$ that is correlated with B's type. The distribution of s conditional on t is given by

$$\Pr(l|L) = \Pr(m|M) = \Pr(h|H) = \theta \text{ and}$$
$$\Pr(m|L) = \Pr(h|L) = \Pr(l|M) = \Pr(h|M) = \Pr(l|H) = \Pr(m|H) = \frac{1-\theta}{2},$$

where $\theta \in [\frac{1}{3}, 1]$ measures the quality of the signal. Thus, the signal indicates the correct type of B with probability θ and each of the wrong types of B with probability $\frac{1}{2}(1-\theta)$.

[4] If $c_B^L \to 0$ then $u_A(x^L) \to p$ and if $c_A \to 0$ then $u_A(x^M) \to p + (q_H + q_M)c_B^M$ and $u_A(x^H) \to p + q_H c_B^H$. Thus, for c_A, c_B^L sufficiently close to zero, both $u_A(x^M)$ and $u_A(x^H)$ will be higher than $u_A(x^L)$.

By Bayes rule[5] the after-signal beliefs of A are given by

$$\Pr(L|l) = \Pr(M|m) = \Pr(H|h) = \theta \text{ and}$$
$$\Pr(M|l) = \Pr(H|l) = \Pr(L|m) = \Pr(H|m) = \Pr(L|h) = \Pr(M|h) = \frac{1-\theta}{2}.$$

So unless $\theta = \frac{1}{3}$, in which case the signal is completely uninformative, A will always come to believe strictly more in the type indicated by the signal. If $\theta = 1$ A will know B's type with certainty.

The model is almost as easy to solve as the basic model without a signal.[6] For B the situation is exactly the same, so a type t will accept A's proposal x if and only if $x \leq x^t = p + c_B^t$. A can condition its proposal on its received signal. For each $s \in \{l, m, h\}$ A's problem has the same structure as in the basic model, we just have to replace the ex ante probabilities q_L, q_M, q_H with the updated probabilities $\Pr(L|s), \Pr(M|s), \Pr(H|s)$. For example, if A receives the signal l then it will compare the expected final utilities (1)-(3) with $q_L = \theta$ and $q_M = q_H = \frac{1}{2}(1-\theta)$ and choose the x^t providing the highest utility.

We will write A's optimal proposal as

$$x^*(s, \theta)$$

to indicate that it depends on the received signal and signal quality. Note that, since the signal is not informative at all for $\theta = \frac{1}{3}$, the optimal proposal will not depend on the received signal in this case:

$$x^*(l, \frac{1}{3}) = x^*(m, \frac{1}{3}) = x^*(h, \frac{1}{3}).$$

When $\theta = 1$, A knows B's type with certainty and will therefore offer just enough

[5] Because $q_L = q_M = q_H = \frac{1}{3}$ we get, for all $s \in \{l, m, h\}$ and $t \in \{L, M, H\}$,

$$\Pr(t|s) = \frac{q_t \Pr(s|t)}{q_L \Pr(s|L) + q_M \Pr(s|M) + q_H \Pr(s|H)} = \Pr(s|t).$$

[6] Formally, our solution concept is Perfect Bayesian Equilibrium. However, there is no signaling aspect in this game because A moves first and has no private information so it can simply be solved by backwards induction.

for this type to accept:

$$x^*(l,1) = x^L,\ x^*(m,1) = x^M,\ x^*(h,1) = x^H.$$

For all $\theta < 1$ war will happen with some positive probability unless A proposes x^L after receiving any of the three signals.

In the appendix we provide the general solution to the model by specifying A's optimal proposal $x^*(s,\theta)$ for all $s \in \{l,m,h\}$ and $\theta \in [\frac{1}{3},1]$. We split the presentation into three cases depending on whether A's proposal for $\theta = \frac{1}{3}$ is x^L, x^M, or x^H (and specify the restrictions on the parameters c_A, c_B^L, c_B^M, and c_B^H for each of these cases). Suppose, for example, that we are in the first of these cases and that A has received the signal h. Then we can plot $u_A(x^L)$, $u_A(x^M)$, and $u_A(x^M)$ (with $q_L = q_M = \frac{1}{2}(1-\theta)$ and $q_H = \theta$) as functions of θ to find A's optimal proposal for any value of θ. Figure 1 displays an example where, as we increase θ, A first changes its proposal from x^L to x^M (at $\theta = \theta_1'$) and then to x^H (at $\theta = \theta_1''$).

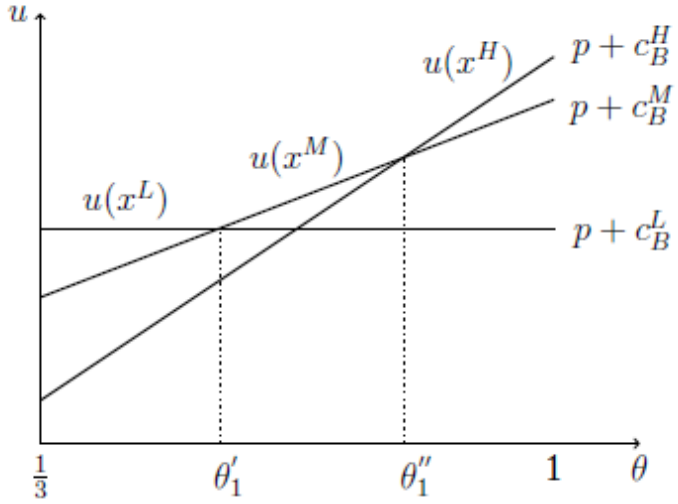

Figure 1: An example of $u_A(x^L)$, $u_A(x^M)$, and $u_A(x^M)$ as functions of θ in a situation where A chooses x^L for $\theta = \frac{1}{3}$ and has received the signal h

The Probability of War

We know that war is possible only when A is incompletely informed about B's cost of war. Here we ask the question if war is always less likely when A's information is more precise. In other words, is the ex ante probability of war in the model always (weakly) decreasing in θ?

To be able to write out the ex ante probability of war, first define the function $\text{War}(s,t,\theta)$:

$$\text{War}(s,t,\theta) = \left\{ \begin{array}{l} 1 \text{ if } x^*(s,\theta) > x^t \\ 0 \text{ if } x^*(s,\theta) \leq x^t \end{array} \right\}.$$

So $\text{War}(s,t,\theta)$ is equal to one if there will be war when A receives the signal s, B is of type t, and the quality of the signal is θ. Otherwise $\text{War}(s,t,\theta)$ is equal to zero. With this definition, the ex ante probability of war as a function of θ can be written

$$\begin{aligned} \text{ProbWar}(\theta) &= \sum_{\substack{s=l,m,h \\ t=L,M,H}} q_t \Pr(s|t)\text{War}(s,t,\theta) \\ &= \frac{1}{3} \sum_{\substack{s=l,m,h \\ t=L,M,H}} \Pr(s|t)\text{War}(s,t,\theta). \end{aligned}$$

Note that there will never be war when B is a type H because it is never optimal for A to make a proposal x above x^H. That is, $\text{War}(s,H,\theta) = 0$ for all s,θ. Thus, all terms with $t = H$ are redundant in the sum above.

With our first result we demonstrate that the probability of war is not always weakly decreasing in the quality of A's signal. So it is indeed possible that better information leads to a higher probability of war.

Proposition 1 *Suppose A makes the proposal x^L for $\theta = \frac{1}{3}$. Then ProbWar(θ) is* not *weakly decreasing in θ.*

Proof. When A proposes x^L all B-types will accept. Thus we have $\text{ProbWar}(\frac{1}{3}) = 0$. When $\theta = 1$ it is strictly optimal for A to make the proposal that just satisfies the B-type it is facing. Since the utilities (1)-(3) are continuous in q_L, q_M, q_H, A

will make the same proposal when θ is sufficiently close to 1. That is,

$$x^*(l,\theta) = x^L,\ x^*(m,\theta) = x^M,\ x^*(h,\theta) = x^H \text{ for } \theta \text{ close to } 1.$$

Thus, if θ is close to one then there will be war if $(s,t) = (m,L)$, (h,L), or (h,M). These combinations of signal and type are possible as long as $\theta < 1$, so we conclude that $\text{ProbWar}(\theta) > 0$ for θ close to but below one.[7] Since $\text{ProbWar}(\frac{1}{3}) = 0$, this immediately implies that $\text{ProbWar}(\theta)$ is not weakly decreasing. □

Figure 2 displays an example of $\text{ProbWar}(\theta)$ in the case where A proposes x^L for $\theta = \frac{1}{3}$. The jumps happen when A's proposal after one or more of the signals changes. $x^*(m,\theta)$ and $x^*(h,\theta)$ change from x^L to x^M at $\theta = \theta_1'$ and $x^*(h,\theta)$ changes from x^M to x^H at $\theta = \theta_1''$.

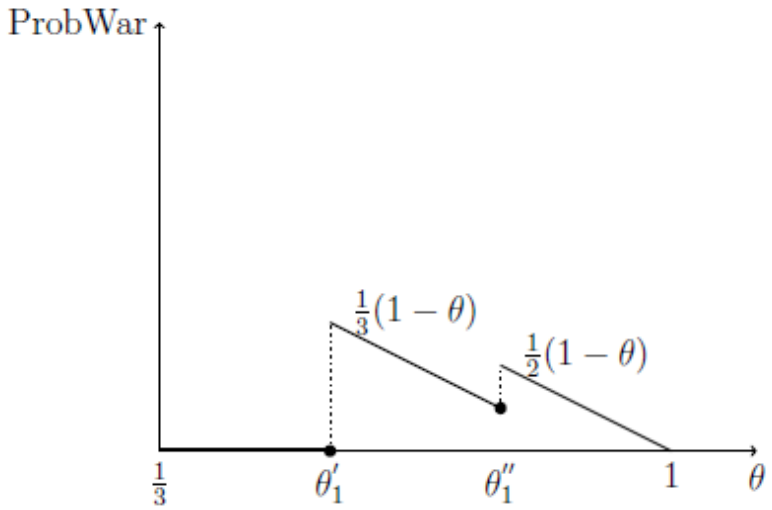

Figure 2: An example of $\text{ProbWar}(\theta)$ in the case where A proposes x^L for $\theta = \frac{1}{3}$

While Proposition 1 provides a negative answer to our general question of whether the ex ante probability of war is *always* weakly decreasing as we move towards the complete information benchmark, the fact that it depends on starting

[7] More precisely, for such θ the ex ante probability of war is

$$\frac{1}{3}(\Pr(m|L) + \Pr(h|L) + \Pr(h|M)) = \frac{1}{2}(1-\theta).$$

in a situation where the probability of war is zero makes it of limited appeal. After all, the question is primarily interesting because war can happen with asymmetric information while it cannot in the complete information benchmark. So we are more interested in situations where we start (at $\theta = \frac{1}{3}$) with a positive probability of war.

Proposition 2 shows that it is also possible for the probability of war to increase with θ when $\text{ProbWar}(\frac{1}{3}) > 0$.

Proposition 2 *Suppose A makes the proposal x^M for $\theta = \frac{1}{3}$. Then, generically,*[8] *ProbWar(θ) is* not *weakly decreasing in θ.*

The full proof can be found in the appendix. The main argument is similar to the one used in the proof of Proposition 1 and is easily explained. As θ gets close enough to one, A will propose x^H after receiving the signal h. At the $\theta < 1$ where $x^*(h, \theta)$ changes from x^M to x^H, $\text{War}(h, M, \theta)$ will jump from zero to one because a B-type M will accept x^M but not x^H. Thus, unless $x^*(l, \theta)$ or $x^*(m, \theta)$ changes to a lower value at the same θ, $\text{ProbWar}(\theta)$ will make an upward jump. Figure 3 displays one possibility for $\text{ProbWar}(\theta)$ in the case where A proposes x^M for $\theta = \frac{1}{3}$. $x^*(l, \theta)$ changes from x^M to x^L at $\theta = \theta_2'$ and $x^*(h, \theta)$ changes from x^M to x^H at $\theta = \theta_2''$.

[8] This means that the conclusion holds for almost all relevant parameter specifications, i.e., that the set of parameter specifications such that it does not hold has Lebesgue measure zero. See the proof for details.

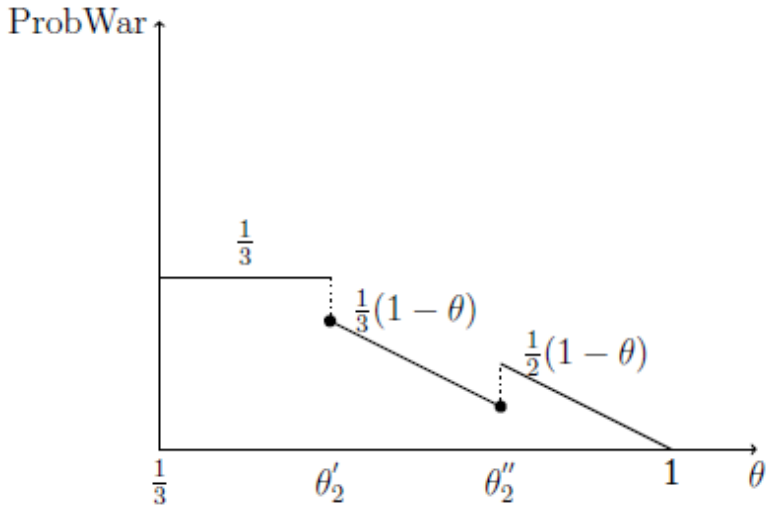

Figure 3: An example of ProbWar(θ) in the case where A proposes x^M for $\theta = \frac{1}{3}$

Our main and final result shows that even if we start (at $\theta = \frac{1}{3}$) in a situation where the probability of war is as high as it can possibly be in the model, then it can still be the case that the probability of war is not a weakly decreasing function of θ. If A makes the proposal x^H when $\theta = \frac{1}{3}$ then the B-types L and M will go to war and thus the ex ante probability of war is $\frac{2}{3}$. Since it is never optimal for A to make a lower proposal than x^H, this is the highest possible probability of war in the model. In this case an increase in ProbWar(θ) does not simply follow from the fact that $x^*(s,\theta) = x^*(s,1)$ for θ sufficiently close to one. The adjustment to the complete information proposals will not necessarily lead to an increase in the probability of war when we start at x^H. For example, if the only changes in proposals as we increase θ are that $x^*(l,\theta)$ changes from x^H to x^L and that $x^*(m,\theta)$ changes from x^H to x^M then it is easy to see that ProbWar(θ) is weakly decreasing.

So how is it possible for the probability of war to increase with θ if A proposes x^H for $\theta = \frac{1}{3}$? Suppose A has received the signal m and consider its proposal as θ increases. It is possible that $x^*(m,\theta)$ first changes from x^H to x^L and then to x^M. The latter change will lead to an increase in ProbWar(θ) if there are no changes in $x^*(l,\theta)$ or $x^*(h,\theta)$ at the same value of θ.

The precise result is stated in Proposition 3. The proof can be found in the

appendix.

Proposition 3 *For any balance of power parameter $p \in (0,1)$ and any cost of war $0 < c_A < \frac{1-p}{2}$ for A there exists an open set of costs (c_B^L, c_B^M, c_B^H) for B such that*

- *A makes the proposal x^H when $\theta = \frac{1}{3}$ and*
- *ProbWar(θ) is* not *weakly decreasing.*

Note that there is an upper bound on A's cost of war in the proposition. This condition is necessary for x^H to be optimal for A at $\theta = \frac{1}{3}$. With this proposal A runs the highest risk of ending in war, which provides the expected utility $p - c_A$. This can only be optimal if c_A is not too high. Also note that the proof explicitly specifies the set of costs for B mentioned in the proposition. The fact that the set is open shows that the two bullet points are not only satisfied in knife-edge situations.

Figure 4 displays an example of the probability of war function in the case where A proposes x^H for $\theta = \frac{1}{3}$. $x^*(l,\theta)$ and $x^*(m,\theta)$ change from x^H to x^L at $\theta = \bar{\theta}_3$ and $x^*(m,\theta)$ changes from x^L to x^M at $\theta = \theta_3'$.

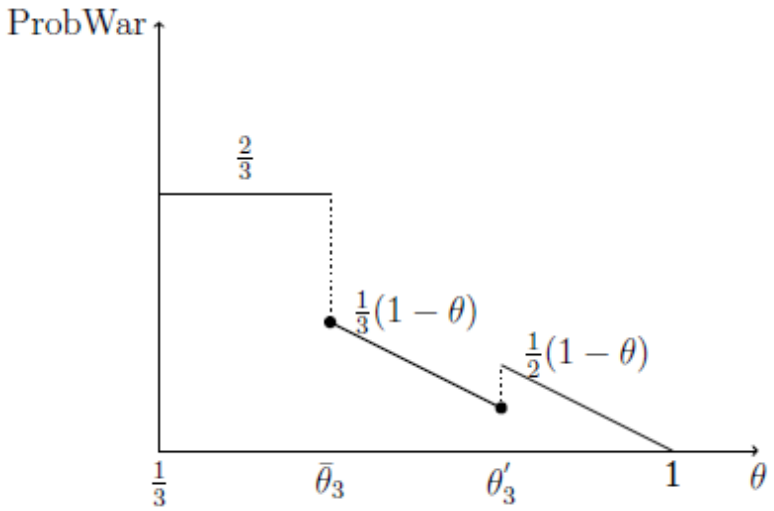

Figure 4: An example of ProbWar(θ) in the case where A proposes x^H for $\theta = \frac{1}{3}$

Conclusion

We have demonstrated that, in a simple version of the standard ultimatum bargaining model with private information about the costs of war, the ex ante probability of war can sometimes increase when the incompletely informed state receives more precise (less noisy) information about its opponent. So even though the model is closer to the complete information benchmark when information is more precise, the predicted probability of war will not necessarily be closer to zero.

Existing research on the link between the level of uncertainty about the characteristics of opponents and the probability of war (e.g. Reed, 2003) has shown that war always becomes less likely when the uncertainty is reduced in a way that preserves the mean value of states' beliefs. Thus, a main contribution of this paper is to highlight the difference between this approach, which is primarily useful for comparing different crises with varying levels of uncertainty, and our approach, which is more suitable for exploring how more precise information will affect the outcome of a given crisis.

Appendix

General solution to the model: $x^*(s,\theta)$ for all s,θ

We split the presentation into three cases depending on whether A's proposal for $\theta = \frac{1}{3}$ is x^L, x^M, or x^H.

1. **A makes the proposal x^L for $\theta = \frac{1}{3}$.**

   The conditions for x^L to be optimal for $\theta = \frac{1}{3}$ are easily derived from the equations (1)-(3) with $q_L = q_M = q_H = \frac{1}{3}$:

   $$c_B^L \geq \frac{2c_B^M - c_A}{3} \text{ and } c_B^L \geq \frac{c_B^H - 2c_A}{3}. \tag{4}$$

   In this case $x^*(s,\theta)$ is:

   $$\begin{aligned}
   x^*(l,\theta) &= x^L \text{ for all } \theta, \\
   x^*(m,\theta) &= \left\{ \begin{array}{l} x^L \text{ for } \theta \leq \theta_1' \\ x^M \text{ for } \theta > \theta_1' \end{array} \right\}, \\
   x^*(h,\theta) &= \left\{ \begin{array}{l} x^L \text{ for } \theta \leq \theta_1' \\ x^M \text{ for } \theta_1' < \theta \leq \theta_1'' \\ x^H \text{ for } \theta > \theta_1'' \end{array} \right\} \text{ if } \theta_1' < \theta_1'', \\
   x^*(h,\theta) &= \left\{ \begin{array}{l} x^L \text{ for } \theta \leq \bar{\theta}_1 \\ x^H \text{ for } \theta > \bar{\theta}_1 \end{array} \right\} \text{ if } \theta_1' \geq \theta_1'',
   \end{aligned}$$

   where

   $$\theta_1' = \frac{2c_B^L - c_B^M + c_A}{c_B^M + c^A},\ \theta_1'' = \frac{c_B^M + c_A}{2c_B^H - c_B^M + c_A}, \text{ and } \bar{\theta}_1 = \frac{c_B^L + c_A}{c_B^H + c_A}.$$

   At $\theta = \theta_1'$ A is indifferent between the proposals x^L and x^M after receiving $s = m$ or $s = h$. At $\theta = \theta_1''$ A is indifferent between x^M and x^H after receiving $s = h$. At $\theta = \bar{\theta}_1$ A is indifferent between x^L and x^H after receiving $s = h$.

2. **A makes the proposal x^M for $\theta = \frac{1}{3}$.**

The conditions for x^M to be optimal for $\theta = \frac{1}{3}$ are:

$$c_B^M > \frac{3c_B^L + c_A}{2} \text{ and } c_B^M \geq \frac{c_B^H - c_A}{2}. \tag{5}$$

In this case $x^*(s,\theta)$ is:

$$\begin{aligned} x^*(l,\theta) &= \left\{ \begin{array}{l} x^M \text{ for } \theta < \theta_2' \\ x^L \text{ for } \theta \geq \theta_2' \end{array} \right\}, \\ x^*(m,\theta) &= x^M \text{ for all } \theta, \\ x^*(h,\theta) &= \left\{ \begin{array}{l} x^M \text{ for } \theta \leq \theta_2'' \\ x^H \text{ for } \theta > \theta_2'' \end{array} \right\}, \end{aligned}$$

where

$$\theta_2' = \frac{c_B^M - c_B^L}{c_B^M + c^A} \text{ and } \theta_2'' = \frac{c_B^M + c^A}{2c_B^H - c_B^M + c_A}.$$

At $\theta = \theta_2'$ A is indifferent between the proposals x^L and x^M after receiving $s = l$. At $\theta = \theta_2''$ A is indifferent between x^M and x^H after receiving $s = h$.

3. A **makes the proposal** x^H **for** $\theta = \frac{1}{3}$**.**

The conditions for x^H to be optimal for $\theta = \frac{1}{3}$ are:

$$c_B^H > 3c_B^L + 2c_A \text{ and } c_B^H > 2c_B^M + c_A. \tag{6}$$

In this case $x^*(s,\theta)$ is:

$$\begin{aligned} x^*(l,\theta) &= \left\{ \begin{array}{l} x^H \text{ for } \theta < \bar{\theta}_3 \\ x^L \text{ for } \theta \geq \bar{\theta}_3 \end{array} \right\}, \\ x^*(m,\theta) &= \left\{ \begin{array}{l} x^H \text{ for } \theta < \bar{\theta}_3 \\ x^L \text{ for } \bar{\theta}_3 \leq \theta \leq \theta_3' \\ x^M \text{ for } \theta > \theta_3' \end{array} \right\} \text{ if } \bar{\theta}_3 \leq \theta_3', \\ x^*(m,\theta) &= \left\{ \begin{array}{l} x^H \text{ for } \theta < \theta_3'' \\ x^M \text{ for } \theta \geq \theta_3'' \end{array} \right\} \text{ if } \bar{\theta}_3 > \theta_3', \\ x^*(h,\theta) &= x^H \text{ for all } \theta, \end{aligned}$$

where

$$\theta_3' = \frac{2c_B^L - c_B^M + c_A}{c_B^M + c_A},\ \theta_3'' = \frac{c_B^H - c_B^M}{c_B^H + c_B^M + 2c_A},\ \text{and}\ \bar{\theta}_3 = \frac{c_B^H - 2c_B^L - c_A}{c_B^H + c_A}.$$

At $\theta = \theta_3'$ A is indifferent between the proposals x^L and x^M after receiving $s = m$. At $\theta = \theta_3''$ A is indifferent between x^M and x^H after receiving $s = m$. At $\theta = \bar{\theta}_3$ A is indifferent between x^L and x^H after receiving $s = l$ or $s = m$.

Proofs

Proof of Proposition 2. Consider $x^*(s, \theta)$ in the case where A makes the proposal x^M for $\theta = \frac{1}{3}$ (see case 2 in the previous subsection). At $\theta = \theta_2''$ (immediately after, to be precise), $x^*(h, \theta)$ changes from x^M to x^H. This means that Probwar(θ) jumps upwards by the amount $q_H \Pr(m|H) = \frac{1}{6}(1 - \theta_2'')$ if there are no changes in $x^*(l, \theta)$ or $x^*(m, \theta)$ at $\theta = \theta_2''$. This is the case if $\theta_2' \neq \theta_2''$. I.e., we have that Probwar(θ) is not weakly increasing unless $\theta_2' = \theta_2''$, which is equivalent to

$$(c_B^M - c_B^L)(2c_B^H - c_B^M + c_A) - (c_B^M + c^A)^2 = 0.$$

Since the set of roots for a polynomial in $\mathbb{R}^n$ has Lebesgue measure zero, it follows that Probwar(θ) is not weakly increasing for almost all cost specifications $(c_A, c_B^L, c_B^M, c_B^H)$ (with the usual restrictions $c_A > 0$ and $0 < c_B^L < c_B^M < c_B^H < 1 - p$) such that A proposes x^M for $\theta = \frac{1}{3}$. □

Proof of Proposition 3. Let $p \in (0, 1)$ and $c_A \in (0, \frac{1-p}{2})$. Consider the set C of costs (c_B^L, c_B^M, c_B^H) for B (with the usual restriction $0 < c_B^L < c_B^M < c_B^H < 1 - p$) such that x^H is optimal for A for $\theta = \frac{1}{3}$ and $\bar{\theta}_3 < \theta_3'$ (see case 3 in the previous subsection). These two conditions are equivalent to

$$c_B^L < \frac{c_B^H - 2c_A}{3} \text{ and } c_B^L > \frac{c_B^H c_B^M - (c_A)^2}{c_B^H + c_B^M + 2c_A}. \tag{7}$$

The first inequality is equivalent to the first inequality in (6), the second inequality is equivalent to $\bar{\theta}_3 < \theta_3'$. The second inequality in (6) follows from these two inequalities.

For $(c_B^L, c_B^M, c_B^H) \in C$ we have that ProbWar(θ) is not weakly decreasing because it jumps upwards at $\theta = \theta_3'$ ($x^*(m, \theta)$ changes from x^L to x^M). So it suffices to show that the set is non-empty and open. Because all conditions for (c_B^L, c_B^M, c_B^H) to belong to C are given by strict inequalities it is easy to see that the set is open. To see that it is not empty, first choose a number y with $2 < y < \frac{1-p}{c_A}$, which is possible because $c_A < \frac{1-p}{2}$. If we let $c_B^M = \frac{1}{y} c_A$ and $c_B^H = y c_A$ then the inequalities in (7) become

$$c_B^L < \frac{y-2}{3} c_A \text{ and } c_B^L > 0.$$

Thus we have that $(c_B^L, \frac{1}{y} c_A, y c_A) \in C$ for all c_B^L in the interval $(0, \min\{\frac{1}{y}, \frac{y-2}{3}\} c_A)$, which shows that C is not empty. $\square$

References

[1] Fearon, J.D. 1995. "Rationalist Explanations for War." *International Organization* 49(3): 379-414

[2] Kurizaki, S. 2015. "Signaling and Perception in International Crises: Two Approaches." *Journal of Theoretical Politics*, forthcoming (available at http://jtp.sagepub.com/)

[3] Reed, W. 2003. "Information, Power, and War." *American Political Science Review* 97(4): 633-641

[4] Schub, R. 2015. "Are You Certain? Leaders, Overprecision, and War." Working Paper (available at http://scholar.harvard.edu/schub/research)

[5] Wittman, D. 2009. "Bargaining in the Shadow of War: When is Peaceful Resolution Most Likely." *American Journal of Political Science* 53(3): 588-602