The Calculus of the Security Dilemma[*]

Avidit Acharya[†] **Kristopher W. Ramsay**[‡]

March 2013

Abstract

Some scholars known as *offensive realists* claim that in the uncertainty of world politics, trust and cooperation between states is extremely unlikely. Others, such as *defensive realists*, claim that rational states are capable of finding ways to counteract the complications created by misperceptions and distrust, and to reduce uncertainty to levels where it no longer inhibits cooperation. In this paper, we construct a formal model to show how in some situations cooperation between states is indeed very unlikely: even in the presence of minor misperceptions, states fail to cooperate. We then ask whether diplomacy (modeled as cheap talk) is able to remedy the failure. We show that in many situations, allowing the countries to communicate prior to taking their actions does *not* enable them to cooperate.

JEL Classification Codes: C72, F51

Key words: international conflict, cooperation, security dilemma, realism, cheap talk, incomplete information, higher order uncertainty

---

[*] We would like to thank Sylvain Chassang, Robert Keohane, Andrew Kydd, Adam Meirowitz, Stephen Morris, Satoru Takahashi, two anonymous referees, and seminar participants at Caltech, the University of Rochester, and the Princeton Conference on Signaling and Perceptions for valuable comments. We are especially grateful to Mark Fey for his help and suggestions.

[†] W. Allen Wallis Institute of Political Economy, Harkness Hall, University of Rochester, Rochester NY 14627-0158 USA (email: aachary3@z.rochester.edu)

[‡] Department of Politics, 038 Corwin Hall, Princeton University, Princeton NJ 08544-1012 USA (email: kramsay@princeton.edu).

1 Introduction

Introduced by Herz (1950) and Butterfield (1951), the *security dilemma* in international relations theory describes the obstacles that two countries face in achieving peace and cooperation. In recent years, debate over the scope and severity of the security dilemma has resurfaced amongst various factions of scholars that fall within the *realist* school of thought. On the one hand, *offensive realists* like Mearsheimer (2001) argue that in the anarchy of world politics, fears about the intentions of rival states may drive even two security-seeking states away from cooperation. On the other hand, *defensive realists* like Glaser (1995) respond to the pessimism of offensive realists by questioning the strength of the connections between anarchy, uncertainty and cooperation. In particular, defensive realists claim that two security-seeking states should not find it difficult to cooperate if they recognize each other as security-seeking, and while uncertainty about a state's motivations can complicate matters, uncertainty alone does not imply the dire predictions of offensive realism (Glaser 1997).[1]

An important contribution to the security dilemma debate is a paper by Kydd (1997*a*), which to our knowledge is the first formal treatment of incomplete information in the security dilemma. This paper laid the foundations for a book titled *Trust and Mistrust in International Relations*, in which Kydd (2005) argues that Bayesian game theory is well-suited to analyze the problems of trust that are at the heart of the security dilemma. Kydd (2005) proposes a new theory, which he calls *Bayesian realism*, as an alternative to offensive and defensive realism. In Bayesian realism, states have different preferences for revising the status quo and the level of trust between them is variable, as opposed to offensive and defensive realism in which states are always security-seeking. Using a signaling framework, Kydd (2005) shows that trustworthy states are often able to separate themselves from untrustworthy ones; and, in a dynamic setting, he shows how rational states can use costly gestures to reduce distrust to manageable levels, even when it is very high to start.

In this paper, we build on Kydd's (2005) premise that problems of trust are at the heart of the security dilemma. However, our model of uncertainty and distrust in the security dilemma differs in several important ways from Kydd's (2005) model and other previous work. First, while Kydd (2005) analyzes situations with common

---

[1] Other perspectives, such as those of *motivational realists*, are presented by Schweller (1996), and Kydd (1997*b*), who provides a thorough review of the arguments appearing in the literature.

knowledge of the fundamentals and uncertainty about countries' preferences, we focus explicitly on uncertainty about the strategic fundamentals. In our model, countries receive informative but noisy signals regarding the advantage of unilateral defection. Specifically, we consider a situation where there is some (small) uncertainty about whether the strategic situation is described by a Prisoner's Dilemma or by a Stag Hut. Previous theory, like Kydd's (2005), has modeled situations like World War II, where revisionist Germany is dissatisfied with an arrangement that provides it power that is incommensurate with its material and military status. Our model, alternatively, is of the security dilemma as it arises in situations like World War I (see, e.g., Van Evera 1999, Ch. 7). Here, the relevant uncertainty is about the state of military technology, the relative benefits to offensive military action, and the incentives to reciprocate cooperation.

Second, our model supports the argument of offensive realists that even when states know that they are each security-seeking, trust can be so low that cooperation becomes impossible. We show how two countries fail to cooperate even when each is certain that the other is trustworthy, and they are both certain that they are both certain that they are both trustworthy. One might wonder where the uncertainty enters our model if the countries can be this certain. We show that this uncertainty enters the model in the *higher order beliefs* of the countries: although a country may be certain that both countries are trustworthy, and certain that the other country is also certain of this, it may not be certain that the other country is certain that it is certain ... and so on, that both are trustworthy.[2] Rather than take this kind of higher order uncertainty literally, we view it as a metaphor for the deep fears, suspicions, and doubts that leaders have about how trustworthy their counterparts are, and how their counterparts may perceive their own perceptions of the strategic environment—exactly the kinds of fears and suspicions that lead offensive realists to question the possibility of cooperation in the anarchy of world politics.[3]

A number of previous papers have made a point similar to some we make. For example, Chassang and Padró i Miquel (2009*a*) study the role of fear, and the evolution of conflict, in a dynamic model of defensive weapons procurement. In a related

---

[2] Here we think of trustworthy states as those whose best response to cooperation in a given strategic setting is to reciprocate cooperation.

[3] Under this interpretation, our argument is somewhat related to Butterfield's (1951) *irreducibility dilemma* that argues that no leader can ever know what is in the mind of other leaders.

paper, Chassang and Padró i Miquel (2010) study a dynamic exit game with a noisy signal structure and show that there is an important link between strategic risk and the possibility of cooperation.[4] These authors build on the work of Carlsson and van Damme (1993), who introduced a perturbation that can be used to select equilibria in games with multiple equilibria. Our paper differs from these papers in important ways.[5] Most important of these is the fact that our framework is tractable enough that we can study the effect cheap talk on the set of equilibrium outcomes. Thus, following Fearon (1995) and others,[6] we model diplomacy as cheap talk, and ask whether diplomacy can make cooperation possible when it would otherwise not be possible. The same question is also asked by Baliga and Sjöström (2004), who analyze a security dilemma and show that cheap talk can increase the probability of cooperation when players are uncertain about the arming costs of their adversaries. However, unlike Baliga and Sjöström (2004), who focus on idiosyncratic costs (a private values case), we study a situation where countries might misperceive the fundamentals of the strategic environment (a common values case), and we show that in many situations cheap talk *cannot* remedy their failure to cooperate.

Like Baliga and Sjöström (2004) and Baliga and Morris (2002), however, we are unable to provide general results on cheap talk that hold across the class of games that we study.[7] Nevertheless, all of our results are negative, and go against the grain of prior work. For instance, Example 2 in Baliga and Morris (2002) shows that with correlated types, cheap talk can improve upon the no-cooperation outcome that obtains in its absence, even when only one side is permitted to speak. Moreover, the side that speaks uses only two messages. In contrast, we show in Proposition 2 that no matter how correlated the types are in our model, cheap talk does not change the set of equilibrium outcomes when only one side is permitted to speak

---

[4] Chassang and Padró i Miquel (2009*b*) use similar methods to show how mutual fears may aggravate the effect of negative economic shocks on civil conflict intensity.

[5] For one, our framework is tractable enough that we can generalize the argument of previous models to a broader class of games. Second, we implicitly show that the importance of *risk-dominance* in the Carlsson-van Damme approach is an artifact of symmetries built into their information structure, and that risk-dominance is not a necessary condition for analogous results to hold under a more general class of information structures. (See Harsanyi and Selten, 1988, for the definition of risk-dominance.)

[6] See also Ramsay (2011), Sartori (2002) and Smith (1998).

[7] Baliga and Sjöström (2004) study a particular game, while Baliga and Morris (2002) construct three examples to demonstrate that their results on cheap talk do not generalize outside the class of incomplete information games that they study.

and the message space is finite. Similarly, Example 3 in Baliga and Morris (2002) and Theorem 2 of Baliga and Sjöström (2004) study models with private values. These authors show that if both players can send one of two messages, then mutual cooperation occurs in equilibrium with positive probability. In contrast, Proposition 3 of this paper finds conditions for our common values environment such that no side cooperates with more than zero probability in any equilibrium of the game even when *both* countries are permitted to send one of two messages. Although none of our results directly contradict previous findings, they do indicate the limitations in our understanding of cheap talk's effect in games of incomplete information. More importantly, they advance our intuition about cheap talk's effect in games with higher order uncertainty.

The common values environment that we study in this paper is what sets it apart from previous formal work on the security dilemma.[8] Yet, our approach has a clear motivation in many of the classic works on the security dilemma. For example, Quester (1977), Jervis (1978) and Levy (1984) all stress the importance of uncertainty about the "strategic fundamentals" of the security environment in explaining the occurrence of conflict. These authors place the balance between offensive and defense capabilities, given current state of military technology—or what the literature calls the *offense-defense balance*—at the heart of their analysis. Lynn-Jones (1996, p.665) defines the offense-defense balance, more precisely, to be the amount of resources that a state must invest in offense to offset an adversary's investment in defense. The offense is said to be advantaged if it is beneficial to launch a surprise attack on the adversary rather than to defend one's territory (Jervis 1978, p.187). These and other authors claim that the historical relevance of the offense-defense balance is undeniable. For example, Van Evera (1999, Ch.7) recounts how differing perceptions of the offense-defense balance contributed to the start of World War I.

Our starting point, then, is to place states' uncertainty regarding the strategic environment at the center of our analysis of the security dilemma. Specifically, we consider an environment with structural uncertainty about the state of the offense-defense balance, and we analyze the effect of this uncertainty on the possibility of bilateral cooperation. Our modeling approach builds directly on the approach established by the previous literature.

---

[8] Previous work, e.g. Kydd (2005) and Baliga and Sjöström (2004), focuses on private values and the use of costly or costless signaling to achieve cooperation.

$$\begin{array}{c|r|c} \multicolumn{3}{c}{\text{country } -i} \\ & C & D \\ \hline C & W_i & 0 \\ \hline D & W_i + a_i & w_i \end{array} \quad \text{(rows: country } i\text{)}$$

Prisoner's Dilemma

$$\begin{array}{c|r|c} \multicolumn{3}{c}{\text{country } -i} \\ & C & D \\ \hline C & W_i & 0 \\ \hline D & W_i - b_i & w_i \end{array}$$

Stag Hunt

Figure 1: The Security Dilemma

2 Model

Following Jervis (1978), we begin by supposing that relations between two countries are described either by a Stag Hunt or by a Prisoner's Dilemma. Figure 1, which we reproduce here almost exactly as it appears in Jervis (1978, p.171), depicts the two possibilities: two countries $i = 1, 2$ must decide whether to cooperate C or defect D, but there is uncertainty as to whether their payoffs are given by the left payoff matrix or by the right.[9] The figure depicts the payoffs to each country i when facing the opponent country $-i$. Whether the true payoffs are given by the left or right matrix is determined by a state variable $s \in \mathbb{R}$, which represents the offense-defense balance. If $s \leq 0$ then the offense is advantaged, and the payoffs are given by the left matrix, a Prisoner's Dilemma. If $s > 0$ then the defense is relatively more advantaged, and the payoffs are given by the right matrix, a Stag Hunt. Throughout the paper, we maintain the assumptions that $W_i > w_i > 0$ and $a_i, b_i > 0$ for both $i = 1, 2$. These assumptions guarantee that the left payoff matrix in Figure 1 is indeed a Prisoner's Dilemma, and the right matrix is a Stag Hunt, as they are labeled.

The state s is a realization of some distribution π over $S \subseteq \mathbb{R}$. We assume that conditional on s, each country i receives a private signal $x_i \in \mathbb{R}$ drawn from a distribution $G_i(\cdot|s)$. We refer to the triple (π, G_1, G_2) where $G_i = \{G_i(\cdot|s)\}_{s \in S}$, $i = 1, 2$, as the *information structure*. Given the information structure, the set of possible signals for country i is $X_i = \bigcup_{s \in S} \operatorname{supp} G_i(\cdot|s)$, and a pure strategy for country i is a function $\alpha_i : X_i \to \{C, D\}$.

---

[9] The only difference is that Jervis (1978) provided a preference ordering over outcomes for each matrix, while we consider parametric payoffs that satisfy his ordering.

We now make some assumptions about the information structure. Given the information structure, let $H_i(\cdot|x_i)$ denote player i's posterior distribution over the state (updated by Bayes rule), conditional on receiving signal x_i. Also, let $F_i(\cdot|x_i)$ denote country i's posterior distribution over the possible signals received by the other country $-i$ (again, updated by Bayes rule), conditional on receiving signal x_i. Then, we assume

(A1) $\exists t \in \mathbb{R}$ s.t. $\forall i = 1, 2$,

(i) if $t \geq x_i \in X_i$ then $\operatorname{supp} H_i(\cdot|x_i) \subseteq (-\infty, 0]$.

(ii) if $t \leq x_i \in X_i$ then $\exists \varepsilon > 0$ s.t. $F_i(x_i - \varepsilon|x_i) > b_i/(w_i + b_i)$.

Part (i) of the assumption states that if a country receives a small enough signal, then it is certain (believes with probability 1) that it is playing the Prisoner's Dilemma in the left matrix of Figure 1. Part (ii) states that if a country i receives a high signal, believing that the payoffs are given by the Stag Hunt, then it believes that the other country's signal is lower than its signal with probability larger than $b_i/(w_i + b_i)$. This technical assumption implies the substantive assumption that defecting is not too "risky." If country i conjectured that its opponent plays a strategy that prescribes defection for signals smaller than its own, then country i would have a strict incentive to also defect.

Rather than describing a particular game, we have so far characterized a class of games $\mathcal{G}$ that we call *security dilemma* games. Holding fixed the players, $i = 1, 2$, their common action set, $\{C, D\}$, and the parameters, $(w_i, W_i, a_i, b_i)_{i=1,2}$, a security dilemma game $\Gamma \in \mathcal{G}$ is fully described by its information structure (π, G_1, G_2) satisfying assumption (A1). We now complete the description of some games that fall in the class $\mathcal{G}$.

2.1 Examples

The games described below have different information structures, but all of them are security dilemma games satisfying assumption (A1).

Game Γ^A. Suppose that the prior distribution of the state variable s is the improper uniform prior on $\mathbb{R}$.[10] Country 1 observes the state perfectly, so it always

---

[10] The assumption of an improper prior is nonstandard, but poses no difficulties, since the players' interim beliefs are well-defined. (See, e.g., Morris and Shin 2003.)

receives the signal $x_1 = s$, while country 2 receives a signal x_2 that is uniformly distributed on the interval $[s - \xi, s + \xi]$ with ξ arbitrarily small. This implies that conditional on receiving signal x_2, country 2 believes that the state is uniformly distributed on $[x_2 - \xi, x_2 + \xi]$. It also implies that conditional on its signal x_i, each country believes that the other country's signal is uniformly distributed on $[x_i - \xi, x_i + \xi]$. Part (i) of assumption A1 is satisfied, with $t = -\xi$, and part (ii) is satisfied if

(A2) $b_i < w_i$, $i = 1, 2$.

This is because for the game we have just described, $F_i(\cdot|x_i)$ is continuous at $x_{-i} = x_i$ with $F_i(x_i|x_i) = 1/2$ for all $x_i \in X_i = \mathbb{R}$, $i = 1, 2$. Note that (A2) implies that mutual defection is risk-dominant in the Stag Hunt of Figure 1, but that (A1) itself does not contain any implicit assumption regarding risk-dominance.

Game Γ^B. Again, suppose that the prior distribution of s is the improper uniform prior on $\mathbb{R}$. However, this time assume that each country observes the state with some noise: conditional on the state s, each country receives a private signal independently drawn from the uniform distribution over $[s - \xi, s + \xi]$ with ξ arbitrarily small. This implies that conditional on its signal x_i, each country believes that s is distributed uniformly on $[x_i - \xi, x_i + \xi]$. Conditional on its signal x_i, each country i believes that the other country's signal is distributed according to the tent-shaped density

$$f(x_{-i}|x_i) = \begin{cases} \frac{1}{2\xi}\left(1 - \frac{x_i - x_{-i}}{2\xi}\right) & \text{if } x_i - 2\xi \leq x_{-i} \leq x_i \\ \frac{1}{2\xi}\left(1 + \frac{x_i - x_{-i}}{2\xi}\right) & \text{if } x_i < x_{-i} \leq x_i + 2\xi \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Assumption (A1) is satisfied (again, with $t = -\xi$) if (A2) holds, for the same reason as in game Γ^A.

Game Γ^E. The state space is $S = \{-1, 0, 1, 2, ..., \infty\}$. The prior probability of state $s \in S$ is given by $\left(\frac{1}{2}\right)^{2+s}$. If the state is s, then each country i independently observes signal $x_i = s$ with probability $q \in (0, 1)$, and signal $x_i = s + 1$ with probability $1 - q$. Therefore, conditional on signal x_i, country i believes the state is x_i with probability $\frac{q}{2-q}$ and believes that the state is $x_i - 1$ with complementary probability. Conditional on signal x_i, country i believes that the other country's signal is $x_i - 1$ with probability $q\left(1 - \frac{q}{2-q}\right)$, $x_i + 1$ with probability $(1 - q)\left(\frac{q}{2-q}\right)$ and x_i with remaining probability. Part (i) of assumption (A1) is satisfied with $t = 0$. For part (ii), we need

(A3) $q\left(1-\frac{q}{2-q}\right) > \max\left\{\frac{b_1}{w_1+b_1}, \frac{b_2}{w_2+b_2}\right\}$.

Since we can set b_1 and b_2 arbitrarily small, we can always find parameters such that this inequality holds.

In the game Γ^A, a country with signal larger than ξ is certain that the payoffs are given by the Stag Hunt. A country with signal larger than 2ξ is certain that the other country is certain that the payoffs are given by the Stag Hunt. A country with signal larger than 3ξ is certain that the other country is certain that it is certain that the payoffs are given by the Stag Hunt. And so on. The games Γ^B and Γ^E have analogous belief structures, all of which are similar to Rubinstein's (1989) email game. Yet, we show that cooperation is not possible in any equilibrium of any game in the class $\mathcal{G}$.

3 Main Result

We now state and prove our main result: no matter how high the signals of the countries, and no matter how precise their observations of the state, there is no equilibrium of any security dilemma game in which any type of either country cooperates.

Theorem 1. *Every security dilemma game $\Gamma \in \mathcal{G}$ has a unique equilibrium in which all types of both countries defect.*

Proof. First note that for every game $\Gamma \in \mathcal{G}$, it is an equilibrium for all types of both countries to defect. To show that there are no other equilibria, suppose that there is an equilibrium in which a nonempty set of types $\mathcal{C}_i \subseteq X_i$ of some country i cooperate.

If $\mathcal{C}_i \neq \emptyset$ for some country i, then the number $x^* = \inf \mathcal{C}_1 \cup \mathcal{C}_2$ exists, and by assumption (A1) we have $x^* \geq t$. In addition, either $x^* = \inf \mathcal{C}_1$ or $x^* = \inf \mathcal{C}_2$, or both. Let j be any country such that $x^* = \inf \mathcal{C}_j$. Now, there are two possibilities: (i) $x^* \in \mathcal{C}_j$ and (ii) $x^* \notin \mathcal{C}_j$. Suppose $x^* \in \mathcal{C}_j$ and that country j receives signal x^*. Given country $-j$'s equilibrium strategy, let φ denote the probability with which country j believes that country $-j$ will cooperate. Since $x^* \geq t$, assumption (A1) implies that country j believes that country $-j$'s signal is smaller than x^* with probability larger than $b_j/(w_j + b_j)$. But, by definition of x^*, all types of country $-j$ below x^* defect. So $\varphi < w_j/(w_j + b_j)$. This implies that the expected payoff to country j from cooperating is

$$\varphi W_j < \varphi W_j + w_j - \varphi(w_j + b_j) = \varphi(W_j - b_j) + (1 - \varphi)w_j \qquad (2)$$

where the quantity on the right side of the inequality is country j's expected payoff from defecting. Therefore, we have shown that the type x^* of country j can profitably deviate to defection: a contradiction.

Next, suppose that $x^* \notin \mathcal{C}_j$. By construction, we can choose a type $\tilde{x} > x^*$ that is close enough to x^* so that $\tilde{x} \in \mathcal{C}_j$ and the type $\tilde{x}$ of country j believes with probability at most $\tilde{\varphi} < w_j/(w_j + b_j)$ that country $-j$ will cooperate. We can then use an argument similar to the one above to show that this type of country j could profitably deviate to defection: again, a contradiction. □

Reconstructing the Belief Structure

The logic of Theorem 1 can be explained by reconstructing the countries' beliefs associated with the information structure of a particular security dilemma game. Consider the game Γ^A whose information structure was described in Section 2.1. Assume (A2) so that Γ^A belongs to the class $\mathcal{G}$, and for expositional purposes assume, in addition, that the payoffs are symmetric:

(A4) $(w_i, W_i, a_i, b_i) = (w, W, a, b)$, $i = 1, 2$.

For country 1 to cooperate, it must believe with at least probability $p = w/(w + b) > 1/2$ that its opponent will also cooperate. Now, recall that a country that receives a signal smaller than $-\xi$ must defect. Therefore, country 1 must believe that country $-i$'s signal is larger than $-\xi$ with probability at least p. In the terminology of Monderer and Samet (1989) country 1 must "p-believe" that country 2's signal is larger than $-\xi$. For this to be true, country 1's signal must be weakly larger than the threshold $x^0 = -2\xi(1-p)$. This threshold is calculated by finding the value of x^0 such that the length of the interval $[-\xi, x^0 + \xi]$ is p times the length of the interval $[x^0 - \xi, x^0 + \xi]$, which is 2ξ.

Now, observe that p-believing that country 2 received a signal larger than $-\xi$ is only a necessary condition for country 1 to cooperate. It is not sufficient. In fact, we need country 1 to p-believe the following event as well:

country 2 p-believes that country 1's signal is larger than $-\xi$.

Otherwise, if country 2 does not p-believe that country 1's signal is larger than $-\xi$, then country 2 cannot be expected to cooperate. And if country 2 does not cooperate, then country 1 does not have an incentive to cooperate either. But then for country 2 to p-believe that country 1's signal is larger than $-\xi$, country 2's signal must be at least x^0. Therefore, for country 1 to p-believe that country 2 p-believes that country 1's signal is larger than $-\xi$, country 1's signal must be weakly larger than the threshold $x^1 = \xi(2p-1) - 2\xi(1-p)$. This threshold is calculated by finding the value of x^1 such that the length of the interval $[x^0, x^1 + \xi]$ is $2\xi p$.

Again, however, the conditions that country 1 p-believes that country 2's signal is larger than $-\xi$ *and* p-believes that country 2 p-believes that country 1's signal is larger than $-\xi$ are together still only necessary for country 1 to cooperate, not sufficient. Country 1 must also p-believe that country 2 p-believes that country 1 p-believes that country 2's signal is larger than $-\xi$. Otherwise, country 1 cannot expect country 2 to expect country 1 to cooperate, will therefore not expect country 2 to cooperate, and thus it will not be in country 1's interest to cooperate. In fact, for country $i = 1, 2$ to cooperate it must p-believe each of the following infinite sequence of events:

(0) $-i$'s signal is larger than $-\xi$
(1) $-i$ p-believes that i's signal is larger than $-\xi$
(2) $-i$ p-believes that i p-believes that $-i$'s signal is larger than $-\xi$
(3) $-i$ p-believes that ...
(4) ... ad infinitum

Proceeding inductively, one can show that if country i p-believes the (0)th through (n)th one of these statements, its signal must be at least

$$x^n = n\xi\,(2p-1) - 2\xi(1-p) \tag{3}$$

Since $\xi > 0$ and $p > 1/2$, this quantity is unboundedly increasing in n. Consequently, there is no signal value for which country i p-believes every element of the infinite sequence of events listed above. As a result, there is no signal value for which country i cooperates.

Reconstructing the belief structure also enables us to clarify the importance of assumption A1(ii) in proving Theorem 1. This assumption guarantees that the sequence of thresholds x^n is increasing and converges to $+\infty$. For example, in the game Γ^A, suppose that (A2) holds with reverse inequality so that A1(ii) is violated. Then

$p < 1/2$, and the sequence of x^n decreases, converging to $-\infty$. Therefore, the iterative procedure above fails, and there may be equilibria in which the countries cooperate.

4 Cheap Talk Diplomacy

Theorem 1 above shows that without the opportunity to communicate, two countries playing a security dilemma game $\Gamma \in \mathcal{G}$ are incapable of cooperating in equilibrium. In this section we ask whether the opportunity to communicate enables cooperation.

Consider the following modification to a game $\Gamma \in \mathcal{G}$. Suppose that after both countries observe their private signals, each is able to make a public announcement. Both countries can then make their decisions of whether or not to cooperate dependent on the pair of announcements. Let M_i denote the nonempty set of available messages for country i. A pure strategy for country i is a pair (μ_i, σ_i) such that $\mu_i : X_i \to M$ is its message rule and $\sigma_i : M_1 \times M_2 \times X_i \to \{C, D\}$ is its action rule. Note that each country can condition its action on its signal and on the pair of announcements. We have now defined a new game $\hat{\Gamma}$, which we call the *cheap talk extension* of Γ. Let $\hat{\mathcal{G}}(\Gamma)$ denote the class of games that are cheap talk extensions of the game Γ.[11]

To study the effect of cheap talk, we make the following assumption, which states that there are positive spillovers to cooperation in the Stag Hunt payoff matrix of Figure 1:

(A5) $b_i < W_i - w_i$, $i = 1, 2$

Assumption (A5) implies that each country would always like the other country to cooperate regardless of whether it intends to do so itself.[12]

Unfortunately, we are not able to provide general results that hold across all cheap talk extensions of games in $\mathcal{G}$. Instead, we study various cheap-talk extensions to the games described in Section 2.1. We begin by proving our simplest result, which relies on an argument due to Baliga and Morris (1998). These authors showed that pre-play communication has no effect on the equilibrium outcome of Rubinstein's (1989)

---

[11] Note that this is a large class, since in describing a cheap talk extension $\hat{\Gamma}$, we have not specified the sets M_1 and M_2. For example, M_1 could be finite while M_2 is infinite, or they could both be finite, or one could be a singleton while the other one is infinite, etc.

[12] Cheap talk extensions to Bayesian games with binary action positive spillovers were first studied by Baliga and Morris (2002).

email game. Not surprisingly, the same is true for the game Γ^E, which is similar to the email game.

Proposition 1. *Assume (A3) and (A5). Then, in every equilibrium outcome of every cheap talk extension $\hat{\Gamma}^E \in \hat{\mathcal{G}}(\Gamma^E)$, all types of both countries defect.*

Proof. Suppose to the contrary, that some type x_i of either country $i = 1, 2$ cooperates in some equilibrium of a game $\hat{\Gamma}^E \in \hat{\mathcal{G}}(\Gamma^E)$. Fix the equilibrium, and let x_j^* be the smallest type of either country that cooperates, with j denoting the country associated with this type. Let m_j^* be the equilibrium message sent by x_j^*, and let M_{-j}^* be the set of messages of the other country that induce x_j^* to cooperate. (In other words, $\sigma_j(m_j^*, m_{-j}, x_j^*) = C$ for all $m_{-j} \in M_{-j}^* \neq \emptyset$.) Since it is strictly dominant for types -1 and 0 to defect, we must have $x_j^* \geq 1$. Next, by assumption (A5), the type $x_{-j}^* = x_j^* - 1$ of country $-j$ must send a message $m_{-j}^* \in M_{-j}^*$. But, by definition of x_j^*, the type $x_{-j}^* = x_j^* - 1$ of country $-j$ chooses to defect. Therefore, conditional on receiving message m_{-j}^*, the type x_j^* believes that country $-j$ will defect with probability weakly larger than $q\left(1 - \frac{q}{2-q}\right) > b_j/(w_j + b_j)$, which holds by assumption (A3). Therefore, country j cannot cooperate after message profile (m_j^*, m_{-j}^*), establishing the intended contradiction. □

Proposition 1 shows that cheap talk is ineffective when added to the game Γ^E. Does this result also hold for games Γ^A and Γ^B? We do not provide a complete answer to this question, but our results below suggest that communication is difficult, if not impossible. First, consider the case of one-sided messages. Let $\overline{\mathcal{G}}(\Gamma^\ell)$ denote the (sub)class of cheap talk extensions of the game Γ^ℓ, $\ell = A, B$, such that M_1 is finite and M_2 is a singleton. In these games, only player 1 has the opportunity to communicate, and may do so with a finite set of messages (which we allow to be arbitrarily large). The next proposition states that there are no equilibria in which communication takes place in any game in this class.

Proposition 2. *Assume (A2) and (A5). Then, in every equilibrium outcome of every cheap talk extension $\overline{\Gamma}^\ell \in \overline{\mathcal{G}}(\Gamma^\ell)$, $\ell = A, B$, all types of both countries defect.*

Proof. We prove this only for the case where $\ell = A$. The case where $\ell = B$ is conceptually identical, but more tedious.

Let $p_i = w_i/(w_i + b_i)$ and note that $p_i \in \left(\frac{1}{2}, 1\right)$, $i = 1, 2$, by assumption (A2). Also note that $X_1 = X_2 = \mathbb{R}$. For any set $X \subseteq \mathbb{R}$ let

$$\Phi_X(x) = \frac{1}{2\xi} \cdot \lambda(X \cap [x - \xi, x + \xi]) \tag{4}$$

where $\lambda(\cdot)$ is Lesbegue measure. In words, this is the probability mass that type x of a country assigns to the event that the signal received by the other country falls in the set X, unconditional on the message profile. Observe that

$$x' \leq x \;\Rightarrow\; \Phi_X(x') \leq \Phi_X(x) \qquad \forall X \subseteq [x, \infty) \tag{5}$$

Now, fix an equilibrium $((\mu_1, \sigma_1), (\mu_2, \sigma_2))$. Since M_2 is a singleton, we can set $M_2 = \{m_2\}$, so that $\mu_2(x) = m_2$ for all $x \in X_2$. Define the sets

$$\begin{aligned} \mathcal{C}_1^m &= \{x \in X_1 \mid \mu_1(x) = m,\ \sigma_1(m, m_2, x) = C\} \\ \mathcal{D}_1^m &= \{x \in X_1 \mid \mu_1(x) = m,\ \sigma_1(m, m_2, x) = D\}, \qquad m \in M_1 \end{aligned} \tag{6}$$

These are the sets of country 1 types that send message m and respectively cooperate and defect. Also, define the set

$$\mathcal{C}_2 = \{x \in X_2 \mid \exists x_1 \in [x - \xi, x + \xi] \text{ s.t. } \sigma_2(\mu_1(x_1), m_2, x) = C\} \tag{7}$$

This is the set of country 2 types that cooperate in some equilibrium outcome of the game. Note that if $\mathcal{C}_1^m = \emptyset$ for all $m \in M_1$ then $\mathcal{C}_2 = \emptyset$. So to prove the result, it suffices to show that $\mathcal{C}_1^m = \emptyset$ for all $m \in M_1$. To that end, suppose for the sake of contradiction that $\mathcal{C}_1^m \neq \emptyset$ for some $m \in M_1$ and let $x_1^* = \inf \bigcup_{m \in M_1} \mathcal{C}_1^m$. Then, $\mathcal{C}_2 \neq \emptyset$, so we can define $x_2^* = \inf \mathcal{C}_2$. We know that $x_1^*, x_2^* \geq -\xi$. The contradiction is then established in three steps.

Step 1. $x_1^* - \xi \leq x_2^* \leq x_1^* - \xi(2p_1 - 1)$.

If $x_2^* < x_1^* - \xi$ then there exists a type $x_2 < x_1^* - \xi$ of country 2 such that $\sigma_2(\mu_1(x_1), m_2, x_2) = C$ for some $x_1 \in [x_2 - \xi, x_2 + \xi]$. But, since $x_2 + \xi < x_1^*$, all country 1 types in the interval $[x_2 - \xi, x_2 + \xi]$ defect. Therefore, the type x_1 cannot exist, and we must have $x_1^* - \xi \leq x_2^*$.

If $x_1^* - \xi(2p_1 - 1) < x_2^*$ then there exists a type $x_1 \in \left[x_1^*, x_2^* + \xi(2p_1 - 1)\right)$ such that $\sigma_1(\mu_1(x_1), m_2, x_1) = C$. But by definition of x_2^*, the type x_1 of country 1 believes with probability at most $\frac{1}{2\xi}(x_1 + \xi - x_2^*) < \frac{1}{2\xi}\left(x_2^* + \xi(2p_1 - 1) + \xi - x_2^*\right) = p_1$ that country 2 will cooperate. So type x_1 of country 1 will defect. So $x_2^* \leq x_1^* - \xi(2p_1 - 1)$.

Step 2. $\exists \tilde{x} \in [x_1^* - \xi, x_1^*]$ s.t. $\sigma_2(\mu_1(\tilde{x}), m_2, x_2) = D$ for all $x_2 \in [x_2^*, x_1^*]$.

Assume not. Let $M_1^* = \{m \in M_1 \mid \exists x_1 \in [x_1^* - \xi, x_1^*] \text{ s.t. } \mu_1(x_1) = m\}$ be the set of messages sent by country 1 types between $x_1^* - \xi$ and x_1^*. For each $m \in M_1^*$, let $x^m \in [x_2^*, x_1^*]$ be a type such that $\sigma_2(m, m_2, x^m) = C$. (The hypothesis is that the type x^m exists for each $m \in M_1^*$.) Then, because the type x^m, $m \in M_1^*$, cooperates after seeing message m from country 1, we must have

$$\Phi_{\mathcal{C}_1^m}(x^m) \geq p_2 \left(\Phi_{\mathcal{C}_1^m}(x^m) + \Phi_{\mathcal{D}_1^m}(x^m) \right) \qquad \forall m \in M_1^*. \tag{8}$$

Summing over $m \in M_1^*$, and rearranging, we get

$$\sum_{m \in M_1^*} \Phi_{\mathcal{C}_1^m}(x^m) \geq \frac{p_2}{1 - p_2} \sum_{m \in M_1^*} \Phi_{\mathcal{D}_1^m}(x^m) \tag{9}$$

Also, note that by definition of M_1^*, and because $[x_1^* - \xi, x_1^*] \subset [x^m - \xi, x^m + \xi]$, we have

$$\sum_{m \in M_1^*} \Phi_{\mathcal{D}_1^m}(x^m) \geq \frac{1}{2} \tag{10}$$

Combining this with (9), and the fact that $p_2 > \frac{1}{2}$, we arrive at

$$\sum_{m \in M_1^*} \Phi_{\mathcal{C}_1^m}(x^m) > \frac{1}{2} \tag{11}$$

However, notice that we must have

$$\sum_{m \in M_1^*} \Phi_{\mathcal{C}_1^m}(x^m) \leq \sum_{m \in M_1^*} \Phi_{\mathcal{C}_1^m}(x_1^*) = \Phi_{\bigcup_{m \in M_1^*} \mathcal{C}_1^m}(x_1^*) \leq \frac{1}{2} \tag{12}$$

The second inequality holds because $\mathcal{C}_1^m \subseteq [x_1^*, \infty)$ for all $m \in M_1^*$, by definition of x_1^*. The equality holds because $\{\mathcal{C}_1^m\}_{m \in M_1}$ is by definition a collection of mutually disjoint sets. The first inequality holds by the property in (5), since $x^m \leq x_1^*$ for all $m \in M_1^*$. But then (11) and (12) contradict each other.

Step 3. The type $\tilde{x}$ (from Step 2) of country 1 has a profitable deviation.

Consider a type $x_1 \in [x_1^*, x_1^* + \varepsilon) \cap \left(\bigcup_{m \in M_1} \mathcal{C}_1^m \right)$ where ε is small, i.e. $0 < \varepsilon < \xi(2p_1 - 1)$. Let $\mu_1(x_1) = m_1$. Since $\Phi_{\mathcal{C}_2}(x_1) \geq p_1 > \frac{1}{2}$, the fact that ε is small implies that there is a type $x_2 \in [x_2^*, x_1^*]$ such that $\sigma_2(m_1, m_2, x_2) = C$. Therefore, $\mu_1(\tilde{x}) = \tilde{m} \neq m_1$ by definition of type $\tilde{x}$. However, observe that by sending message $\tilde{m}$ type $\tilde{x}$ of country 1 can expect country 2 to cooperate with probability at most

$$\frac{1}{2\xi}(\tilde{x} + \xi - x_1^*) = \frac{1}{2} - \frac{1}{2\xi}(x_1^* - \tilde{x}) \tag{13}$$

But by deviating to the message m_1, type $\tilde{x}$ of country 1 can expect country 2 to cooperate with probability at least

$$\begin{aligned} p_1 - \frac{1}{2\xi}\left[(x_1+\xi)-(\tilde{x}+\xi)\right] &= p_1 - \frac{1}{2\xi}(x_1-\tilde{x}) \\ &\geq p_1 - \frac{1}{2\xi}(x_1^* + \varepsilon - \tilde{x}) \\ &> \frac{1}{2} - \frac{1}{2\xi}(x_1^* - \tilde{x}) \qquad (14) \end{aligned}$$

So the type $\tilde{x}$ of country 1 can expect a country 2 to cooperate with strictly larger probability after sending message m_1 than after sending message $\tilde{m}$. Therefore, by (A5), it is profitable for the country 1 type $\tilde{x}$ to deviate to message m_1. □

As we mentioned in the introduction, Example 2 in Baliga and Morris (2002) studies the effect of cheap talk in a setting with correlated types. These authors show that with cheap talk, mutual cooperation may be part of an equilibrium outcome even when there are positive spillovers to cooperation. Moreover, the equilibrium that they construct has only one player sending one of two messages. In contrast, observe that in the game $\overline{\Gamma}^A$, the players' types become perfectly correlated as $\xi \to 0$. Yet, Proposition 2 establishes that neither side will ever cooperate even when player 1 can send one of a large but finite number of messages.

Restricting only one side to speak may not be reasonable if the goal is to model diplomacy as cheap talk. Example 3 of Baliga and Morris (2002) and Theorem 2 of Baliga and Sjöström (2004) study models with uncorrelated types. Both papers show that mutual cooperation may be an equilibrium outcome when both players are allowed to speak, even when there are positive spillovers to cooperation. Moreover, in proving this, both papers construct equilibria in which no player uses more than two messages. What happens when we allow for two sided cheap talk with binary message spaces? Let $\breve{\mathcal{G}}(\Gamma^\ell)$ denote the (sub)class of cheap talk extensions of the game Γ^ℓ, $\ell = A, B$, such that $|M_1| = |M_2| = 2$. In this class of games, both players can speak, but may only send one of two messages. So, for example, they may announce their intended action. In this setting, we can construct trivial equilibria in which mutual cooperation is an equilibrium outcome. Consider, for example, the strategy profile where all types of country i, except the type ξ, send message $m_i \in M_i$ and the type ξ sends message $m_i' \neq m_i$; and all types of both countries defect except when the message profile is (m_1', m_2'), in which case they cooperate. This strategy profile is part of an equilibrium, but conditional on any state s, the probability of cooperation by either country is always zero. The next proposition shows that if b_1 and b_2 are low enough, then it can never be greater than zero.

Proposition 3. *Fix a game* $\breve{\Gamma}^\ell \in \breve{\mathcal{G}}(\Gamma^\ell)$, $\ell = A, B$. *For each country* $i = 1, 2$ *there exists a threshold* $\bar{b}_i > 0$ *such that if* $b_i < \bar{b}_i$, $i = 1, 2$, *then conditional on any state* s, *both countries defect with probability 1 in every equilibrium of the game* $\breve{\Gamma}^\ell$.

Proof. We prove this only for $\ell = A$. (Again, the case $\ell = B$ is conceptually identical, but more tedious.) So fix a game $\breve{\Gamma}^A \in \breve{\mathcal{G}}(\Gamma^A)$, and let

$$\bar{b}_i = \min\{W_i - w_i, w_i/7\} \tag{15}$$

We show that if $b_i < \bar{b}_i$, $i = 1, 2$, then conditional on any state s, both countries defect with probability 1 in every equilibrium of the game $\breve{\Gamma}^A$. Note that if $b_i < \bar{b}_i$, $i = 1, 2$, where $\bar{b}_i$ is given by (11), then both (A2) and (A5) are satisfied.

Now, fix an equilibrium, and for each $i = 1, 2$, define the set

$$\begin{aligned} \mathcal{C}_i = \{x \in X_i \mid \exists \varepsilon > 0 \text{ and } m_{-i} \in M_{-i} \text{ s.t.} \\ \forall x_i \in [x, x + \varepsilon),\ \sigma_i(\mu_i(x_i), m_{-i}, x_i) = C\} \end{aligned} \tag{16}$$

Assume for the sake of contradiction that $\mathcal{C}_i \neq \emptyset$ for some $i = 1, 2$. Let $x^* = \inf \mathcal{C}_1 \cup \mathcal{C}_2$, and note that $x^* \geq -\xi$. Let j be any country for which $x^* = \inf \mathcal{C}_j$. Fix $\varepsilon > 0$ small and consider a type $\tilde{x}_j \in [x^*, x^* + \varepsilon) \cap \mathcal{C}_j$. Let $m_{-j} \in M_{-j}$ denote the message such that $\sigma_j(\mu_j(\tilde{x}_j), m_{-j}, \tilde{x}_j) = C$. Then, there must be a type $\tilde{x}'_j \in \mathcal{C}_j$ such that

$$\tilde{x}'_j \leq x^* + \varepsilon + 2\xi \frac{b_j}{w_j + b_j} \tag{17}$$

and $\sigma_j(\mu_j(\tilde{x}'_j), m'_{-j}, \tilde{x}'_j) = C$, where $m'_{-j} \neq m_{-j}$. Otherwise, by (A5), all country $-j$ types in the interval $\left(\tilde{x}_j - \xi, x^* + \varepsilon + 2\xi\frac{b_j}{w_j + b_j} - \xi\right]$ would send message m_{-j}. Therefore, conditional on message profile $(\mu_j(\tilde{x}_j), m_{-j})$, the type $\tilde{x}_j$ would believe that country $-j$ will defect with at least probability

$$\frac{1}{2\xi}\left(\varepsilon + 2\xi\frac{b_j}{w_j + b_j}\right) = \frac{b_j}{w_j + b_j} + \frac{\varepsilon}{2\xi} \tag{18}$$

and therefore would choose to defect after the message profile $(\mu_j(\tilde{x}_j), m_{-j})$, a contradiction.

Now, because each country has only two messages, there are two cases: (i) at least a measure $\xi/2$ of country $-j$ types in the interval $[x^* - \xi, x^*]$ send message m_{-j}, or (ii) at least a measure $\xi/2$ of country $-j$ types in the interval $[x^* - \xi, x^*]$ send message m'_{-j}. In case (i), conditional on message profile $(\mu_j(\tilde{x}_j), m_{-j})$ the type $\tilde{x}_j$ believes that country $-j$ defects with probability at least

$$\frac{1}{2\xi}\left(\frac{\xi}{2} - \varepsilon\right) = \frac{1}{4} - \frac{\varepsilon}{2\xi} \tag{19}$$

In case (ii), conditional on message profile $(\mu_j(\tilde{x}'_j), m'_{-j})$, the type $\tilde{x}'_j$ believes that country $-j$ will defect with probability at least

$$\frac{1}{2\xi}\left[\frac{\xi}{2} - \left(\varepsilon + 2\xi\frac{b_j}{w_j + b_j}\right)\right] = \frac{1}{4} - \frac{\varepsilon}{2\xi} - \frac{b_j}{w_j + b_j} \tag{20}$$

If $b_j < w_j/7$ then we can choose ε small enough that the probability thresholds in (19) and (20) are both strictly larger than $b_j/(w_j + b_j)$, which establishes the intended contradiction. This means that we must have $\mathcal{C}_1 = \mathcal{C}_2 = \emptyset$. Thus, conditional on any state s, both countries defect with probability 1. $\square$

The threshold in (15) that we used in proving Proposition 3 is sufficient, but may not be necessary for our no-cooperation result to hold. As we suggested earlier, we are not sure how far the result in Proposition 3 generalizes, as it is not obvious how to generalize our proof strategy. In any case, Proposition 3 provides a counterpoint to previous results (i.e., Example 3 of Baliga and Morris 2002 and Theorem 2 of Baliga and Sjöström 2004) that show that cheap talk can have a significant effect in two player binary action games of incomplete information with players using only two messages.

What explains the difference between our results and previous results on cheap talk? In Baliga and Sjöström (2004), for a given player the relative payoff to cooperating when the other player defects (or cooperates) is independent of the other player's type. Because some types get high returns to defecting while others pay small costs to cooperating and being defected on, this results in a variation in incentive compatibility constraints that enables cheap talk to be partially informative. In our model, however, this type of variation does not occur, despite types being correlated. While in our model, correlation does not create the requisite variation in incentive constraints to eliminate the incentives for deception, it does (by a clever construction) in Example 2 of Baliga and Morris (2002).

5 Final Remarks

Our results support the logical validity of offensive realism as a paradigm of world politics, and they demonstrate its consistency with a rational theory of international cooperation. However, it would be a mistake to interpret our results as providing an unqualified endorsement of offensive realism. This is because our model is silent about when assumptions (A1) and (A5) are accurate descriptions of real-life situations. For example, as we suggested in the introduction, our assumptions are consistent with the situation that precipitated World War I but not with the situation that precipitated World War II. An explanation for when the World War I situation arises is an important question for research, but is outside the scope of our paper. Moreover, it is straightforward to show that if assumption (A1)(ii) is violated then mutual cooperation is an equilibrium outcome.[13] Similarly, if assumption (A1) is satisfied but (A5) is violated, then cheap talk enables cooperation in some games where it would otherwise not be possible.[14] These observations are hardly surprising given the existing literature. Having said that, it is difficult to make sense of a violation of assumption (A5) in our context: If (A5) does not hold, then a defecting country would (weakly) gain from its opponent defecting rather than cooperating. (So while it is plausible that assumption (A1) might not be empirically descriptive, it is less plausible that assumption (A5) is violated.)

Finally, our paper leaves many questions unanswered. First, we narrowly focused our attention on costless signaling because we were interested in studying the effectiveness of diplomacy. We have left open the question of what would happen if we allowed for costly signals (though we suspect that there are many situations in which costly signaling would be effective). However, we should mention that although it

---

[13] Suppose for concreteness that $(w_i, W_i, a_i, b_i) = (4, 12, 4, 8)$, $i = 1, 2$ and the information structure is given by the game Γ^B. Here (A2) is violated so mutual cooperation is risk dominant. Then, it is easy to verify that there is a symmetric equilibrium in which all types (weakly) above $(1 + 2/\sqrt{3})\xi$ cooperate and all types below this threshold defect.

[14] Suppose for concreteness that $(w_i, W_i, a_i, b_i) = (8, 12, 4, 5)$, $i = 1, 2$ and the information structure is given by the game Γ^B. Here, (A2) is satisfied so mutual defection is the only equilibrium outcome. However, in the cheap talk extension $\overline{\Gamma}^B \in \overline{\mathcal{G}}(\Gamma^B)$ in which country 1 can send one of a finite number of messages, mutual cooperation is an equilibrium outcome. In fact, let $\kappa_1 = 1 - 2\sqrt{30}/9$ and $\kappa_2 = 2\sqrt{6}/3 - (1 + 2\sqrt{30}/9)$. Then, it is easy to verify that the following is an equilibrium to this game: (i) all country 1 types (weakly) above $\kappa_1\xi$ send message m and cooperate, and all other types send message $m' \neq m$ and defect, and (ii) all country 2 types (weakly) above $\kappa_2\xi$ cooperate if and only if the message is m, and all other types defect after every message.

may be possible to generate positive results on cooperation with costly signals (as Kydd 2005 has done), such results would not tell us when we could expect diplomacy to be effective in the presence of fears created by higher order uncertainty. Second, we have left open the question of whether there exists a foundation for our model in which the information structure in this paper arises endogenously when players strategically acquire information. Third, and most importantly, we have left open the question of whether there exists a cheap talk extension to a game in $\mathcal{G}$ in which cheap talk can be effective. By focusing our attention on cheap talk, we were able to compare our results to the influential results of Baliga and Morris (2002) and Baliga and Sjöström (2004). These authors have contributed a great deal to our understanding of the effect of cheap talk on games of incomplete information, and in doing so have advanced our understanding of the effectiveness of diplomacy for maintaining peace in security environments plagued by uncertainty and misinformation. However, our results show that more work is required to achieve a general understanding of the effect of cheap talk in games of incomplete information.

References

[1] Baliga, Sandeep and Stephen Morris. (1998). "Coordination, Spillovers, and Cheap-Talk." Cowles Foundation Discussion Paper #1203, Yale University.

[2] Baliga, Sandeep and Stephen Morris. (2002). "Coordination, Spillovers, and Cheap-Talk." *Journal of Economic Theory* 105(2): 450-468.

[3] Baliga, Sandeep and Tomas Sjöström. (2004). "Arms Races and Negotiations." *Review of Economic Studies*, 71(2): 351–369.

[4] Butterfield, Herbert. (1951). *History and Human Relations.* London, UK: Collins.

[5] Carlsson, Hans and Eric van Damme. (1993). "Global Games and Equilibrium Selection," *Econometrica*, 61(5): 989-1018.

[6] Chassang, Sylvain and Gerard Padró i Miquel. (2009*a*). "Defensive Weapons and Defensive Alliances," *American Economic Review: Papers and Proceedings* 99(2): 282-286.

[7] Chassang, Sylvain and Gerard Padró i Miquel. (2009*b*). "Economic Shocks and Civil War," *Quarterly Journal of Political Science* 4(3): 211-228.

[8] Chassang, Sylvain and Gerard Padró i Miquel. (2010). "Conflict and Deterrence under Strategic Risk," *Quarterly Journal of Economics* 125(4): 1821-1858.

[9] Fearon, James D. (1995). "Rationalist Explanations of War," *International Organization*, 49(3): 379-414.

[10] Glaser, Charles. (1995). "Realists as Optimists: Cooperation as Self-Help." *International Security* 19(3): 50-90.

[11] Glaser, Charles. (1997). "The Security Dilemma Revisited," *World Politics*, 50(1): 171-201.

[12] Harsanyi, John and Reinhard Selten. (1988). *A General Theory of Equilibrium Selection in Games*, Cambridge, MA: the MIT Press.

[13] Herz, John H. (1950). "Idealist Internationalism and the Security Dilemma" *World Politics*, 2(2): 157-180.

[14] Jervis, Robert. (1978). "Cooperation under the Security Dilemma," *World Politics*, 30(2): 167-214.

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

[16] Kydd, Andrew. (1997*b*). "Sheep in Sheep's Clothing: Why Security Seekers Do Not Fight Each Other," *Security Studies*, 7(1): 114-155.

[17] Kydd, Andrew. (2005). *Trust and Mistrust in International Relations*, Princeton, NJ: Princeton University Press.

[18] Levy, Jack S. (1984). "The Offense/Defense Balance of Military Technology: A Theoretical and Historical Analysis," *International Studies Quarterly*, 28(3): 219-238.

[19] Lynn-Jones, Sean. (1996). "Offense-Defense Theory and Its Critics," *Security Studies*, 4(4): 660-691.

[20] Mearsheimer, John. (2001). *The Tragedy of Great Power Politics*, W.W. Norton.

[21] Monderer, Dov and Dov Samet. (1989). "Approximating Common Knowledge with Common Beliefs," *Games and Economic Behavior*, 1: 170-190.

[22] Morris, Stephen and Hyun Song Shin. (2003). "Global Games: Theory and Applications," in M. Dewartripont, L. Hansen and S. Turnovsky (eds.) *Advances in Economics and Econometrics (Proceedings of the Eighth World Congress of the Econometric Society)*, Cambridge: Cambridge University Press.

[23] Quester, George H. (1977). *Offense and Defense in the International System*, New York, NY: Wiley.

[24] Ramsay, Kristopher W. (2011). "Cheap Talk Diplomacy, Voluntary Negotiations and Variable Bargaining Power," *International Studies Quarterly*, 55(4): 1003-1023.

[25] Rubinstein, Ariel. (1989). "The Electronic Mail Game: Strategic Behavior under 'Almost' Common Knowledge," *American Economic Review*, 79(3): 385-391.

[26] Sartori, Anne. (2002). "The Might of the Pen: A Reputational Theory of Communication in International Disputes," *International Organization,* 56(1): 123-151.

[27] Smith, Alastair. (1998). "International Crises and Domestic Politics," *American Political Science Review,* 92(3): 623-638.

[28] Schweller, Randall. (1996). "Neorealism's Status Quo Bias: What Security Dilemma?," *Security Studies,* 5(3): 385-423.

[29] Van Evera, Stephen. (1999). *Causes of War: Power and the Roots of Conflict*, Ithaca, NY: Cornell University Press.