The Strategy of Manipulating Conflict*

Sandeep Baliga
Northwestern University

Tomas Sjöström
Rutgers University

June 24, 2010

Abstract

Two decision-makers choose hawkish or dovish actions in a conflict game with incomplete information. The decision-making can be manipulated by "extremists" who send publicly observed cheap-talk messages. The power of extremists depends on the nature of the underlying conflict game. If actions are strategic complements, a "hawkish extremist" (provocateur) can increase the likelihood of conflict by sending messages which trigger a "fear-spiral" of hawkish actions. This reduces the welfare of both decision-makers. If actions are strategic substitutes, a "dovish extremist" (pacifist) can send messages which cause one decision-maker to back down and become more dovish. This reduces his welfare but benefits the other decision-maker. The hawkish extremist is unable to manipulate the decision-makers if actions are strategic substitutes, and the pacifist is equally powerless if actions are strategic complements.

1 Introduction

Conflicts are often inflamed by actions taken by agents with extreme agendas. For example, Ariel Sharon's symbolic visit to the Temple Mount in September 2000 helped spark the Second Intifada (Hefetz and Bloom [34]). Terrorist attacks on India's parliament in December 2001, mounted by extremists

*This paper replaces our earlier paper "Decoding Terror." We thank Jim Jordan for early discussions which stimulated us to write this paper. We also thank Alessandro Lizzeri and many seminar participants for comments, and Julie Chen, Kane Sweeney and Banu Olcay for excellent research assistance. Any errors are our responsibility.

sponsored by Pakistan's intelligence service I.S.I., created a crisis which benefitted Pakistani extremists by diverting attention from the "war on terror" (Rabasa et al. [51]). In these examples, political insiders (Sharon, I.S.I.) deliberately triggered conflicts by inciting their perceived enemies (Palestinians, India). Terrorist organizations such as E.T.A. behave in a similarly provocative fashion.

At the other end of the ideological spectrum, pacifists want their key audience to *renounce* violence. The Campaign for Nuclear Disarmament (C.N.D.) was initiated by Bertrand Russell during the Cold War. The goal was unilateral nuclear disarmament under the slogan "better red than dead":

> "If no alternative remains except Communist domination or the extinction of the human race, the former alternative is the lesser of two evils" (Russell quoted in Rees [53]).

We will study how different kinds of extremists can manipulate conflicts by sending "messages". But the logic of extremist communication must depend on the nature of the underlying conflict. Following the literature, we will distinguish two kinds of conflicts:

> "World War I was an unwanted spiral of hostility...World War II was not an unwanted spiral of hostility-it was a failure to deter Hitler's planned aggression." Joseph Nye (p. 111, [48].).

Stag hunt and chicken are stylized representations of these two kinds of strategic interactions (Jervis [41]). In stag hunt games, actions are strategic complements. This captures the idea that fear can cause aggression and escalate into conflict, as in Hobbes's "state of nature" or Jervis's "spiralling model". In contrast, chicken is a model of preemption and deterrence, where actions are strategic substitutes, and fear makes a player back down. We will study the ability of extremists to manipulate both kinds of conflicts.[1]

[1] Baliga and Sjöström [7] show how the payoff matrices of stag hunt and chicken games can be derived from a bargaining game with limited commitment to costly conflict. Suppose H represents an invasion of a disputed territory. If only one player chooses H then he has an advantageous bargaining position and gets most of the territory. If nobody invades the disputed territory, then it is divided more equitably. Whether actions are strategic substitutes or complements is decided by what happens if *both* players choose H. If this means a high probability of a war which neither side wants then actions are strategic substitutes. But if the probability of a war is low, actions may be strategic complements instead.

Our formal model is based on the conflict game of Baliga and Sjöström [5]. There are two countries, A and B. In country $i \in \{A, B\}$, a decision-maker called player i chooses a dovish action D or a hawkish action H. Player i may be interpreted as the median voter, or some other pivotal political decision-maker in country i. The hawkish action might be an act of war, accumulation of weapons, or any other aggressive action. It may involve selecting a hawkish agent who will take aggressive actions against the other country. For example, the median voters in Israel and Palestine have to decide whether to support Hamas or Fatah, or Likud or Kadima, respectively.

Each player $i \in \{A, B\}$ can be a dominant strategy dove, a dominant strategy hawk, or a "moderate" whose best response depends on the opponent's action. Player A doesn't know player B's type, and vice versa. Baliga and Sjöström [5] discussed how fear of the opponent can make moderates choose the hawkish action when the actions are strategic complements. Now our main purpose is to understand how extremists can inflame this spiral of fear. In addition, we generalize the model by allowing actions to be strategic substitutes as well as complements.

If the conflict game has strategic complements, then the moderates are "coordination types" who behave as in a stag hunt game: they want to match the action of the opponent. This can trigger an escalating spiral of fear, as in the classic work of Schelling [54] and Jervis [41]. But if the conflict game has strategic substitutes, then the moderates are "opportunists" (anti-coordination types) who behave as in a game of chicken: they choose H if they think the opponent will choose D, but are intimidated and back down (choose D) if they believe the opponent will choose H. Whether actions are strategic complements or substitutes, under fairly mild assumptions on the distribution of types, the conflict game without cheap-talk has a unique equilibrium, referred to as the *communication-free equilibrium*.

Why might real-world decision makers allow themselves to be manipulated by third parties such as Sharon or the C.N.D.? To study this question, we add a third player called "the extremist" to the conflict game. Before players A and B make their decisions, the extremist (player E) sends a publicly observed cheap-talk message. A visit to the Temple Mount may be a real-world example of a such a message. Of course, "cheap-talk" is an idealization. In reality, the message has to be sufficiently dramatic to be noticed above the background noise and daily concerns of media and politicians. For some real-world extremists, this may require a costly message, like an act of violence. Our basic model abstracts from this by assuming messages are

publicly observed at no cost. However, as an extension, we verify that our results are robust to messages being costly to send and receive.

The extremist may be at the center of politics in country A, or the leader of an extremist movement located in, or with influence in, country A. We assume the extremist's true preferences are commonly known. We consider two cases: a hawkish extremist ("provocateur") who wants player A to choose H, and a dovish extremist ("pacifist") who wants player A to choose D. Both kinds of extremists prefer that the opposing player B chooses D.

Political insiders, like Ariel Sharon and the I.S.I., presumably have privileged information about pivotal decision-makers in their home countries. But even extremists who are outsiders, moving about the population, may discover the preferences of the country's pivotal decision-maker, e.g., the degree of religious fervor of the average citizen.[2] We simplify by assuming the extremist has *perfect* information about the true preferences (type) of the pivotal decision maker in country A, player A. Thus, in our model, player E knows player A's (equilibrium) reaction to cheap talk messages, which greatly simplifies the analysis. The model could be generalized to give the extremist only some noisy signal of player A's type, but then player E could not be sure of player A's reaction. This would add complications but no new insights.

Our main interest is in *communication equilibria,* where extremist cheap-talk is effective in the sense of influencing the equilibrium decisions of players A and B. It is at first surprising that such equilibria exist. Since the extremist's payoff function is common knowledge, it is commonly known exactly what he wants player A and B to do. Player A knows what player E knows, and the only possible reason why player A would allow himself to be manipulated by player E's message is that the message might influence player B. But, for the reasons discussed by Aumann [3], it is not obvious why this should be the case. Namely, player B might consider that player E's message lacks credibility: player E cannot signal what he wants players A and B to do, since this is commonly known anyway. Nevertheless, communication equilibria exist. Under some assumptions, there is even a *unique* communication equilibrium. Importantly, even if multiple communication equilibria exist, they always have the same structure and the same welfare implications.

If cheap-talk is effective, then some message m_1 will make player B more

[2]Organizations like the E.T.A. and Hezbollah have both political and military wings which blurs the distinction between insiders and outsiders.

likely to choose H. A hawkish extremist is willing to send message m_1 only if player A also becomes more likely to choose H. Such co-varying actions must be strategic complements. On the other hand, a dovish extremist is willing to send m_1 only if player A becomes more likely to choose D. Such negative correlation occurs when actions are strategic substitutes. This argument implies that if the underlying game has strategic complements, then only a hawkish extremist can communicate effectively. By sending message m_1, the hawkish extremist triggers an unwanted (by players A and B) spiral of fear and hostility, making both players A and B more likely to choose H. Conversely, if the underlying game has strategic substitutes, then only a dovish extremist can communicate effectively. By sending message m_1, the dovish extremist makes player B more likely to choose H and causes player A to back down and choose D. In all scenarios, communication is effective because it ensures *both* A and B change their actions.

With strategic complements, message m_1 can be interpreted as a provocation by a hawkish extremist which increases the tension between players A and B. It occurs only when player A is a "weak moderate" who would have chosen D in the communication-free equilibrium. The provocation makes player B more likely to play H, and player A switches to H in response. Although the hawkish extremist does not want player B to choose H, he is willing to pay this cost to encourage player A to switch to H. For example, Ariel Sharon's visit to the Temple Mount seemed designed to derail the Israeli-Palestinian peace process (Hefetz and Bloom [34]). The December 2001 attack on the Indian Parliament was apparently designed by members of the I.S.I. to increase tension between India and Pakistan, causing Pakistani troops to shift from the Afghanistan border to the Indian border (see Aneja [2], Riedel [52] and Singh [55]). Similarly, E.T.A. and other terrorist organizations seem to deliberately provoke a repressive response by the state, making moderates more sympathetic to the terrorist organization's agenda.[3]

[3]According to Woodsworth [58], E.T.A.'s model of armed action was a

> "spiral of action-repression-action, which operates along the following lines: 1) E.T.A. carries out a provocative violent action against the political system; 2) the system responds with repression against "the masses"; 3) the masses respond with a mixture of panic and rebellion."

According to *The Management of Savagery* [1], Al Qaeda's objective is to provoke American attacks against the Islamic world that will make moderate Muslims turn against the U.S. and its allies:

Provocation is counter-productive if player E knows that player A is a dominant strategy type who will not change his behavior. Therefore, the absence of a provocation indicates that player A might be a dominant strategy hawk. This "bad news" makes player B more likely to choose H than he would be in the communication-free equilibrium (although not as likely as following an act of provocation). Thus, with strategic complements, players A and B are more likely to choose H in the communication equilibrium, *whether or not a provocation actually occurs,* than in the communication-free equilibrium. Because each decision-maker always wants the other to choose D, the communication-free equilibrium interim Pareto dominates the communication equilibrium for players A and B. Eliminating the hawkish extremist would make all types of players A and B strictly better off. This includes player A's most hawkish types - even though their preferences are actually aligned with the hawkish extremist. When the preferences are aligned in this way, the extremist will not behave provocatively, which alarms player B. [4] Without the hawkish extremist, player B would not be so alarmed.

With strategic substitutes, message m_1 can be interpreted as a "peace rally" organized by the pacifist (dovish extremist). It occurs only when player A is a "tough moderate" who would have chosen H in the communication-free equilibrium. The communication equilibrium has a "better red than

> "Force America to abandon its war against Islam by proxy and force it to attack directly so that the noble ones among the masses....will see that their fear of deposing the regimes because America is their protector is misplaced and that when they depose the regimes, they are capable of opposing America if it interferes." Abu Bakr Naji, *The Management of Savagery* ([1] p. 24)

This document, apparently composed by strategic thinkers within Al Qaeda, also describes how the conflict with the Islamic world will destroy the American empire.

> "It is just as the American author Paul Kennedy says: 'If America expands the use of its military power and strategically extends more than necessary, this will lead to its downfall.' " (Naji [1], p. 18).

[4] The fact that the *absence* of terrorism is informative is reminiscent of Sherlock Holmes's "curious incident of the dog in the night-time" (Conan Doyle [18]):

Gregory (Scotland Yard detective): "Is there any other point to which you would wish to draw my attention?"

Holmes: "To the curious incident of the dog in the night-time."

Gregory: "The dog did nothing in the night-time."

Holmes: "That was the curious incident."

dead" flavour: following a peace rally in country A, player B becomes more aggressive, and player A backs down. In fact, whether or not a peace rally occurs, player B is more likely to choose H in the communication equilibrium than in the communication-free equilibrium, and this unambiguously makes player A worse off. Thus, player A would like to ban peace protests if he could. On the other hand, because they induce player A to choose D, peace protests make player B better off.

We study several extensions of the basic model. It is straightforward to allow costly messages, e.g., "terrorism". Moreover, suppose player B appears weak if he does not react aggressively to a terrorist attack, and appearing weak is costly. This amplifies the fear spiral (if actions are strategic complements) and reinforces our basic results. A terrorist attack can trigger the fear spiral even if player E knows neither player A's nor player B's true type. However, without knowing player A's true type, the extremist cannot be sure of player A's reaction. Terrorism backfires if player A sticks to D while player B switches to H. Thus, in this extension, provocation pays only under certain parameter values.

Finally, we consider what happens if player B can make (publicly observed) offensive or defensive investments before the conflict game is played. When the conflict game has strategic complements, player B naturally "over-invests" in defensive capability in order to encourage the opponent to choose D. With strategic substitutes, the strategic effect is more subtle. Intuition suggests that it is optimal to invest in offensive rather than defensive weapons, in order to force the opponent to back down and choose D. This intuition is not valid in the presence of a dovish extremist. When player B's defensive capability increases, the dovish extremist in country A becomes more inclined to engage in peace protests, and as we have seen, this benefits player B. As a result, player B actually over-invests in defensive capability even with strategic substitutes.

2 Related Literature

Our work is related to several strands of literature. Most closely related is Jung [42], who considers communication by a hawkish "Ministry of Propaganda" in a version of the Baliga and Sjöström [5] model. The leader of one country has two possible types, and the Ministry of Propaganda knows the

true type, while the other leader has only one possible type. In the absence of communication, there would be multiple equilibria. Communication serves to refine the set of equilibria, and for this purpose it is crucial that messages are not cheap-talk; the Ministry of Propaganda cares about maintaining a reputation for being accurate, so its payoff depends directly on its message. But its communication is not effective in our sense: whatever the announcement of the Ministry of Propaganda, both leaders choose H with probability one, which is also equilibrium behavior in the absence of communication. In contrast, we study cheap-talk equilibria which do not replicate the outcome of any communication-free equilibrium. This requires two-sided incomplete information and a richer type-space.

Bueno de Mesquita and Dickson [15] and de Figueiredo and Weingast [25] develop models of provocation and terrorism. If a terror act by Hamas, say, is met by costly indiscriminate violence by Israel, this is a signal of the nature of the Israeli regime. Kydd and Walter [43] study "spoiling" where terrorists force an opponent to exit peace negotiations. If terror acts by Hamas are not met by costly suppression by the Palestinian Authority, this a signal of the nature of the Palestinian regime. These authors use the insights of the classic literature on signaling games (Spence [56]) to study the informational content of the actions of the *targets* of extremism. We focus instead on the informational content of the messages sent by extremists, and show how they can influence decision-makers. We assume messages are cheap-talk, but consider costly messages as an extension.

The seminal paper on cheap-talk is Crawford and Sobel [21]. Many articles study cheap-talk in two-player games, with no third party trying to manipulate the outcome. For example, Farrell and Gibbons [29] and Matthews and Postlewaite [47] study cheap-talk before bargaining and auctions. Baliga and Morris [4] study cheap-talk before a game with one-sided incomplete information. Ordershook and Palfrey [49] study the impact of debate before voting and agenda-setting. Matthews [46] gives veto power to the sender and finds, like we do, that at most two messages are sent in equilibrium.

In the language of the cheap-talk literature, our model has one sender (the extremist) and two receivers. Farrell and Gibbons [28] and Goltsman and Pavlov [32] present models of cheap-talk with multiple audiences, but unlike our model there is no strategic interaction between the receivers. More closely related is Levy and Razin [45], where a leader of a democratic country sends a cheap-talk message which is received by two audiences: his own citizens and the decision maker in the other country. The leader and his citizens share

the same state-contingent preferences, but only he knows the true state. If the leader's message is informative, it will directly influence his own citizens. He would prefer to send them private messages (not overheard in the other country), but this is assumed to be impossible. In our model, the extremist's preferences are not aligned with either receiver's, and private messages would be of no value, because we consider a different kind of manipulation: the extremist tries to *indirectly* influence the behavior of player A by provoking player B.

Our work differs in a more fundamental way from the work above. In signalling models beginning with Spence [56], the sender's information directly influences his cost of sending messages. In cheap-talk models beginning with Crawford and Sobel [21], the sender's information directly influences his payoff from the receiver's action. But in our model, neither is true. In the case of a hawkish extremist, it is commonly known that he strictly wants player A to choose H (whatever player B does), and that he strictly wants player B to choose D (whatever player A does). In view of Aumann's [3] argument on the ineffectiveness of cheap-talk in stag hunt games, it may seem surprising that extremist cheap-talk can be effective here. The underlying logic, already explained in the introduction, differs from the previous literature.

Our model of the reciprocal fear of surprise attack is related to the theory of global games introduced by Carlsson and van Damme [22]. Chassang and Padro-i-Miguel [23], [24] use this theory to formalize the logic of mutual fear when information is highly correlated. Edmond [27] considers a global game where the citizens can overthrow a dictator by coordinating on a revolution, but the dictator increases his chances of survival by jamming the citizens' signals about how likely it is that a revolution will succeed. Bueno de Mesquita [14] studies a related model where the level of violence inflicted by uninformed extremists generates information for the population. Our model is not a global game, but manipulation of global games of conflict by public messages is an interesting topic for future research.

Depending on whether actions are strategic complements or substitutes, aggression either begets or deters aggression. A growing empirical literature on the Israeli-Palestinian conflict addresses this point, although the findings are not very conclusive. Berrebi and Klor ([10], [11]) find that terrorism increases support for the right-wing party Likud in Israel and that there is more terrorism when the left-wing party Labor is in power. Jaeger and Paserman ([37], [38], [39]) find that Palestinian violence or suicide attacks lead to increased violence by Israel, but Israeli violence either has no effect

or possibly a deterrent effect. Jaeger et al. [40] find that major events in the conflict, such as the First Intifada, radicalized young Palestinians, but more moderate Israeli violence does not have a permanent effect.

There is a vast theoretical literature on terrorism which is less related to our work, including studies on the link between the quality of terrorist recruits and the state of the economy (Bueno de Mesquita [13]), public goods provision by terrorist organizations (Berman [9] and Iannaccone and Berman [36]), optimal organization of terror and counter-terror networks (Bar-Isaac and Baccara [8] and Goyal and Vigier [33]) and delegation of counter-terrorism to a third party (Padro-i-Miguel and Yared [50]). Bueno de Mesquita [16] and Kydd and Walter [44] provide excellent surveys of these and other issues.

3 The Model

3.1 The Conflict Game without Cheap Talk

Two decision makers, players A and B, simultaneously choose either a hawkish (aggressive) action H or a dovish (peaceful) action D. As mentioned in the introduction, we interpret player $i \in \{A, B\}$ as the pivotal political decision-maker in country i. The payoff for player $i \in \{A, B\}$ is given by the following payoff matrix, where the row represents his own choice, and the column represents the choice of player $j \neq i$.

$$\begin{array}{cc} & \begin{array}{cc} H & D \end{array} \\ \begin{array}{c} H \\ D \end{array} & \begin{array}{cc} -c_i & \mu - c_i \\ -d & 0 \end{array} \end{array} \tag{1}$$

We assume $d > 0$ and $\mu > 0$, so player j's aggression imposes a cost on player i. For simplicity, d and μ are the same for each player. Notice that d captures the cost of being caught out when the opponent is aggressive, while μ represents a benefit from being more aggressive than the opponent. The game has *strategic complements* if $d > \mu$ and *strategic substitutes* if $d < \mu$.

Player $i \in \{A, B\}$ has a cost c_i of taking the hawkish action, referred to as his "type". Neither player knows the other player's type. The two types c_A and c_B are random variables independently drawn from the same distribution. Let F denote the continuous cumulative distribution function, with support $[\underline{c}, \overline{c}]$, and where $F'(c) > 0$ for all $c \in (\underline{c}, \overline{c})$. Notice that the two

players are symmetric *ex ante* (before their types are drawn). When taking an action, player A knows c_A but not c_B, while player B knows c_B but not c_A.

Player i is a *dominant strategy hawk* if H is a dominant strategy ($\mu \geq c_i$ and $d \geq c_i$ with at least one strict inequality). Player i is a *dominant strategy dove* if D is a dominant strategy ($\mu \leq c_i$ and $d \leq c_i$ with at least one strict inequality). Player i is a *coordination type* if H is a best response to H and D a best response to D ($\mu \leq c_i \leq d$). Player i is an *opportunistic type* if D is a best response to H and H a best response to D ($d \leq c_i \leq \mu$). Notice that coordination types exist only in games with strategic complements, and opportunistic types exist only in games with strategic substitutes. Assumption 1 states that the support of F is big enough to include dominant strategy types of both kinds.

Assumption 1 If the game has strategic complements then $\underline{c} < \mu < d < \overline{c}$.
If the game has strategic substitutes then $\underline{c} < d < \mu < \overline{c}$.

The possibility that the opponent might be a dominant strategy type creates a spiral or multiplier effect. With strategic complements, the possibility that the opponent is a dominant strategy hawk causes coordination types who are "almost dominant strategy hawks" (i.e., types close to μ) to play H. This in turn causes "almost-almost dominant strategy hawks" to play H, and an escalating spiral of aggression triggers further aggression (see Baliga and Sjöström [5]). Strategic substitutes generates a very different spiral. Opportunistic types with a cost close to d are "almost dominant strategy doves". The possibility that the opponent is a dominant strategy hawk makes these "almost dominant strategy doves" back off and play D. This emboldens opportunistic types who are "almost dominant strategy hawks" to play H, and so on.

To formalize this argument, suppose player i thinks player j will choose H with probability p_j. Player i's expected payoff from playing H is $-c_i + \mu(1-p_j)$, while his expected payoff from D is $-p_j d$. Thus, if he chooses H instead of D, his *net* gain is

$$\mu - c_i + (d-\mu)p_j \tag{2}$$

A *strategy* for player i is a function $\sigma_i : [\underline{c}, \overline{c}] \to \{H, D\}$ which specifies an action $\sigma_i(c_i) \in \{H, D\}$ for each cost type $c_i \in [\underline{c}, \overline{c}]$. In Bayesian Nash

equilibrium (BNE), all types maximize their expected payoff. Therefore, $\sigma_i(c_i) = H$ if the expression in (2) is positive, and $\sigma_i(c_i) = D$ if it is negative. If expression (2) is zero then type c_i is indifferent, but for convenience we will assume he chooses H in this case.

Player i uses a *cutoff strategy* if there is a *cutoff point* $x \in [\underline{c}, \overline{c}]$ such that $\sigma_i(c_i) = H$ if and only if $c_i \leq x$. Because the expression in (2) is monotone in c_i, all BNE must be in cutoff strategies. Therefore, it is without loss of generality to restrict attention to cutoff strategies. Any such strategy can be identified with its cut-off point $x \in [\underline{c}, \overline{c}]$. As there are dominant strategy doves and hawks by Assumption 1, all BNE must be interior: each player chooses H with probability strictly between 0 and 1.

If player j uses cutoff point x_j, the probability he plays H is $p_j = F(x_j)$. Therefore, using (2), player i's best response to player j's cutoff x_j is to choose the cutoff $x_i = \Gamma(x_j)$, where

$$\Gamma(x) \equiv \mu + (d - \mu)F(x). \tag{3}$$

The function Γ is the best-response function for cutoff strategies. If there is "enough uncertainty", then the spirals that underlie the best-response function generate a unique equilibrium. This is ensured by Assumption 2.

Assumption 2 $F'(c) < |\frac{1}{d-\mu}|$ for all $c \in (\underline{c}, \overline{c})$.

If F happens to be uniform, then there is maximal uncertainty (for a given support) and Assumption 2 is redundant. More precisely, with a uniform distribution, $F'(c) = 1/(\overline{c} - \underline{c})$, so Assumption 1 implies $F'(c) < |\frac{1}{d-\mu}|$. Of course, Assumption 2 is much weaker than uniformity.[5]

Theorem 1 *The conflict game without cheap-talk has a unique Bayesian Nash equilibrium.*

Proof. Equilibria must be in cutoff strategies, and must be interior by Assumption 1. The best response function Γ, defined by (3), is continuous,

[5] Assumption 2 is violated if the type distribution is highly concentrated around one point. In this case, multiple equilibria can easily exist, even if Assumption 1 holds. Notice that we are assuming types are independent. Since the complete information chicken and stag hunt games have multiple equilbria, a small amount of idiosycratic noise, as in Harsanyi's purification argument, will not refine the set of equilbria.

with $\Gamma(\underline{c}) = \mu > \underline{c}$ and $\Gamma(\bar{c}) = d < \bar{c}$, so it has a fixed-point $\hat{x} \in [\underline{c}, \bar{c}]$. If each player uses cut-off $\hat{x}$, the strategies form a BNE. It remains to show this BNE is unique. Notice that $\Gamma'(x) = (d - \mu)F'(x)$, so the best response function is upward (downward) sloping if actions are strategic complements (substitutes). In either case, a well-known sufficient condition for uniqueness is that best-response functions have slope strictly less than one in absolute value.[6] Assumption 2 implies that $0 < \Gamma'(x) < 1$ if $d > \mu$ and $-1 < \Gamma'(x) < 0$ if $d < \mu$. Hence, the best-response functions cross at most once and there is a unique equilibrium. ■

Proposition 1 shows that there exists a unique BNE, which we refer to as the communication-free BNE, whether actions are strategic substitutes or strategic complements (as long as Assumptions 1 and 2 hold). In equilibrium, player i chooses H if $c_i < \hat{x}$, where $\hat{x}$ is the unique fixed point of $\Gamma(x)$ in $[\underline{c}, \bar{c}]$ (see Figure 1 for the case of strategic complements). The symmetry of the game implies that both players use the same cutoff point. The equilibrium can be reached via iterated deletion of dominated strategies, and captures the escalating spiral of fear discussed by Schelling [54] and Jervis [41] (see Baliga and Sjöström [5] for further discussion).

3.2 Cheap-Talk

We now introduce a third player, player E. Player E is the "extremist", as discussed in the introduction. His payoff function is similar to player A's, with one exception: player E's cost type c_E differs from player A's cost type c_A. Thus, player E's payoff is obtained by setting $c_i = c_E$ in the payoff matrix (1), and letting the row represent player A's choice and the column player B's choice. There is no uncertainty about c_E. Formally, c_E is common knowledge among the three players.

Player E knows c_A but not c_B. More generally, the extremist might receive some signal of player A's type. To avoid unnecessary complications, we assume the signal is perfect, so player E knows c_A.

We consider two possibilities. First, if player E is a *hawkish extremist* ("provocateur"), then $c_E < 0$. To put it differently, $(-c_E) > 0$ represents a

[6] This condition is familiar from the IO literature. With upward-sloping best-response functions, as in Bertrand competition with product differentiation, the slope should be less than one. With downward-sloping best-response functions, as in Cournot competition, the slope should be greater than negative one. See Vives [**?**].

benefit the hawkish extremist enjoys if player A is aggressive. The hawkish extremist is guaranteed a strictly positive payoff if player A chooses H, but he gets a non-positive payoff when player A chooses D, so he always wants player A to choose H. Second, if player E is a *dovish extremist* ("pacifist"), then $c_E > \mu + d$. The most the dovish extremist can get if player A chooses H is $\mu - c_E$, while the worst he can get when player A chooses D is $-d > \mu - c_E$, so he always wants player A to choose D. Notice that, holding player A's action fixed, the extremist (whether hawkish or dovish) is better off if player B chooses D.

Before players A and B play the conflict game described in Section 3.1, player E sends a publicly observed cheap-talk message $m \in M$, where M is his message space. For interpretations of this message, see the introduction.[7] The time line is as follows.

1. The cost type c_i is determined for each player $i \in \{A, B\}$. Players A and E learn c_A. Player B learns c_B.

2. Player E sends a (publicly observed) cheap-talk message $m \in M$.

3. Players A and B simultaneously choose H or D.

Cheap-talk is *effective* if there is a positive measure of types that choose different actions at time 3 than they would have done in the unique communication-free equilibrium of Section 3.1. A Perfect Bayesian Equilibrium (PBE) with effective cheap-talk is a *communication equilibrium*. Clearly, if players A and B maintain their prior beliefs at time 3, then they must act just as in the unique communication-free equilibrium. Therefore, for cheap-talk to be effective, player E's message must reveal some information about player A's type.

A strategy for player E is a function $m : [\underline{c}, \bar{c}] \to M$, where $m(c_A)$ is the message sent by player E when player A's type is c_A. Without loss of generality, we assume each player $j \in \{A, B\}$ uses a "conditional" cut-off strategy: for any message $m \in M$, there is a cut-off $c_j(m)$ such that if player j hears message m, then he chooses H if and only if $c_j \leq c_j(m)$.

[7] In reality, extremists sometimes send costly messages, perhaps to "get the attention" of decision makers. In our model, player E would be willing to incur a cost to influence the outcome of the game. Unless these costs are prohibitively big, they would not change the nature of our arguments.

Lemma 1 *In communication equilibrium, it is without loss of generality to assume that M contains only two messages, $M = \{m_0, m_1\}$, where $c_B(m_1) > c_B(m_0)$.*

Proof. Suppose strategy μ is part of a BNE. Because unused messages can simply be dropped, we may assume that for any $m \in M$, there is c_A such that $m(c_A) = m$. Now consider any two messages m and m'. If $c_B(m) = c_B(m')$, then the probability player B plays H is the same after m and m', and this means each type of player A also behaves the same after m as after m'. Clearly, if all players behave the same after m and m', having two separate messages m and m' is redundant. Hence, without loss of generality, we can assume $c_B(m) \neq c_B(m')$ whenever $m \neq m'$.

Whenever player A is a dominant strategy type, player E will send whatever message minimizes the probability that player B plays H. Call this message m_0. Thus,

$$m_0 = \arg\min_{m \in M} c_B(m) \tag{4}$$

Message m_0 is the *unique* minimizer of $c_B(m)$, since (by the previous paragraph) $c_B(m) \neq c_B(m_0)$ whenever $m \neq m_0$.

Player E cannot always send m_0, because then messages would not be informative and cheap-talk would be ineffective (contradicting the definition of communication equilibrium). But, since message m_0 uniquely maximizes the probability that player B chooses D, player E must have some other reason for choosing $m(c_A) \neq m_0$. Specifically, if player E is a hawkish extremist (who wants player A to choose H) then it must be that type c_A would choose D following m_0 but H following $m(c_A)$; if player E is a dovish extremist (who wants player A to choose D) then it must be that type c_A would choose H following m_0 but D following $m(c_A)$. This is the only way player E can justify sending any other message than m_0.

Thus, if player E is a hawkish extremist, then whenever he sends a message $m_1 \neq m_0$, player A will play H. Player B therefore responds with H whenever $c_B < d$. That is, $c_B(m_1) = d$. But $c_B(m) \neq c_B(m')$ whenever $m \neq m'$, so m_1 is unique. Thus, $M = \{m_0, m_1\}$.

Similarly, if player E is a dovish extremist, then whenever he sends a message $m_1 \neq m_0$, player A will play D. Player B's cutoff point must therefore be $c_B(m_1) = \mu$. Again, this means $M = \{m_0, m_1\}$. ■

Notice that this lemma holds for both strategic substitutes and strategic complements, and for both dovish and hawkish extremists. It also does not require Assumption 2.

4 Cheap-Talk with Strategic Complements

In this section, we consider the case of strategic complements, $d > \mu$.

4.1 Doves can't Communicate Effectively

We first show that if player E is a dovish extremist, $c_E > \mu + d$, then he cannot communicate effectively when actions are strategic complements. From Lemma 1, $M = \{m_0, m_1\}$ with $c_B(m_1) > c_B(m_0)$. Thus, player B is more likely to choose H after m_1 than after m_0. The dovish extremist wants both players A and B to play D, so he would only choose m_1 if this message causes player A to play D. Formally, if $m(c_A) = m_1$, then we must have $c_A > c_A(m_1)$, so that type c_A chooses D when he hears message m_1. But if $c_A > c_A(m_1)$ for all c_A such that $m(c_A) = m_1$, then player B expects player A to play D for sure when player B hears m_1, so player B's cut-off point must be $c_B(m_1) = \mu$. But, with $d > \mu$, types below μ are dominant strategy types who always play H, so we cannot have $c_B(m_0) < \mu$, a contradiction. Thus, we have:

Proposition 1 *If player E is a dovish extremist and the game has strategic complements, then cheap-talk cannot be effective.*

When player A is a dominant strategy type, the dovish extremist will obviously send the message m_0 that minimizes the probability that player B chooses H. When player A is a coordination type, the dovish extremist again prefers m_0. Indeed, when actions are strategic complements, the message m_1 which makes player B more likely to play H only serves to make player A's coordination types more likely to play H. But such a spiral of fear and hostility is not desirable to a dovish extremist. Here, Aumann's [3] intuition applies: the dovish extremist will always send message m_0, and so is unable to influence the outcome of the conflict game. In particular, he cannot increase the probability of the "peaceful" outcome DD.

4.2 Hawkish Cheap-Talk

Now suppose player E is a hawkish extremist, $c_E < 0$, and the game has strategic complements. We will construct a communication equilibrium, where the hawkish extremist E uses cheap-talk to increase the risk of conflict

above the level of the communication-free equilibrium of Section 3.1. It is surprising that player E can do this, because c_E is commonly known. That is, it is commonly known that player E wants player B to choose D and player A to choose H. To understand the equilibrium intuitively, it helps to recall that $M = \{m_0, m_1\}$ by Lemma 1, where $c_B(m_1) > c_B(m_0)$, and interpret message m_1 as a "provocation" and message m_0 as "no provocation".

Say that player A is a *susceptible type* if he chooses H following message m_1, but D following m_0. The set of susceptible types is

$$S \equiv (c_A(m_0), c_A(m_1)].$$

The proof of Lemma 1 showed that if $m(c_A) = m_1$ then type c_A must be susceptible. Since the provocation makes player B more likely to choose H, player E will only behave provocatively if it causes player A to change his action from D to H. On the other hand, player E wants player A to choose H and therefore strictly prefers to provoke a conflict whenever player A is susceptible. That is, it is optimal for player E to set $m(c_A) = m_1$ if and only if $c_A \in S$. Accordingly, message m_1 signals that player A will choose H. As argued in the proof of Lemma 1, this implies $c_B(m_1) = d$. Therefore, if m_1 is sent then player B will choose H with probability $F(d)$, so player A prefers H if and only if

$$-c_A + (1 - F(d))\mu \geq F(d)(-d)$$

which is equivalent to $c_A \leq \Gamma(d)$. Thus, player A uses cut-off point $c_A(m_1) = \Gamma(d)$, where Γ is defined by (3).

It remains only to consider how players A and B behave when there is no provocation (message m_0). Let $y^* = c_A(m_0)$ and $x^* = c_B(m_0)$ denote the cutoff points in this case. Thus, if m_0 is sent then player B will choose H with probability $F(x^*)$, so player A prefers H if and only if

$$-c_A + (1 - F(x^*))\mu \geq F(x^*)(-d)$$

which is equivalent to $c_A \leq \Gamma(x^*)$. Thus, $y^* = \Gamma(x^*)$. When player B hears message m_0, he knows that player A is not a susceptible type. That is, c_A is either below y^* or above $\Gamma(d)$, and player A chooses H in the former case and D in the latter case. Therefore, player B prefers H if and only if

$$-c_B + \frac{1 - F(\Gamma(d))}{1 - F(\Gamma(d)) + F(y^*)}\mu \geq \frac{F(y^*)}{1 - F(\Gamma(d)) + F(y^*)}(-d) \qquad (5)$$

Inequality (5) is equivalent to $c_B \leq \Omega(y^*)$, where

$$\Omega(y) \equiv \frac{[1 - F(\Gamma(d))]\,\mu + F(y)d}{[1 - F(\Gamma(d))] + F(y)}$$

Thus, $x^* = \Omega(y^*)$.

To summarize, any communication equilibrium must have the following form. Player E sets $m(c_A) = m_1$ if and only if $c_A \in S = (y^*, \Gamma(d)]$. Player A's cut-off points are $c_A(m_0) = y^*$ and $c_A(m_1) = \Gamma(d)$. Player B's cut-off points are $c_B(m_0) = x^*$ and $c_B(m_1) = d$. Moreover, x^* and y^* must satisfy $y^* = \Gamma(x^*)$ and $x^* = \Omega(y^*)$. Conversely, if such x^* and y^* exist, then they define a communication equilibrium. We now show graphically that they do exist.

By Assumption 2, Γ is increasing with a slope less than one. Since $F(\underline{c}) = 0$ and $F(\bar{c}) = 1$, we have $\Gamma(\underline{c}) = \mu > \underline{c}$ and $\Gamma(\bar{c}) = d < \bar{c}$. Furthermore,

$$\Gamma(d) - \mu = F(d)\,(d - \mu) < d - \mu.$$

Therefore,

$$\Gamma(d) < d. \tag{6}$$

Also,

$$\Gamma(\mu) = \mu(1 - F(\mu)) + dF(\mu) > \mu$$

as $d > \mu$. Let $\hat{x}$ be the unique fixed point of $\Gamma(x)$ in $[\underline{c}, \bar{c}]$. Clearly, $\mu < \hat{x} < \Gamma(d)$ (see Figure 2).

Figure 2 shows three curves: $x = \Omega(y)$, $y = \Gamma(x)$ and $x = \Gamma(y)$. The curves $x = \Gamma(y)$ and $y = \Gamma(x)$ intersect on the 45 degree line at the unique fixed point $\hat{x} = \Gamma(\hat{x})$. Notice that

$$\Omega'(y) = \frac{F'(y)\,(d - \mu)\,(1 - F(\Gamma(d)))}{([1 - F(\Gamma(d))] + F(y))^2}$$

so Ω is increasing. It is easy to check that $\Omega(y) > \Gamma(y)$ whenever $y \in (\underline{c}, \Gamma(d))$. Moreover,

$$\Omega(\underline{c}) = \Gamma(\underline{c}) = \mu$$

and

$$\Omega(\Gamma(d)) = \Gamma(\Gamma(d)) < \Gamma(d)$$

where the inequality follows from (6) and the fact that Γ is increasing. These properties are shown in Figure 2. Notice that the curve $x = \Omega(y)$ lies to the

right of the curve $x = \Gamma(y)$ for all y such that $\underline{c} < y < \Gamma(d)$ (because $\Omega(y) > \Gamma(y)$ for such y), but the two curves intersect when $y = \underline{c}$ and $y = \Gamma(d)$.

As shown in Figure 2, the two curves $x = \Omega(y)$ and $y = \Gamma(x)$ must intersect at some (x^*, y^*), and it must be true that

$$\hat{x} < y^* < x^* < \Gamma(d) < d \tag{7}$$

By construction, $y^* = \Gamma(x^*)$ and $x^* = \Omega(y^*)$. Thus, a communication equilibrium exists.

Both players A and B are strictly more likely to choose H in communication equilibrium than in communication-free equilibrium. To see this, notice that in the communication-free equilibrium, each player's cutoff is $\hat{x}$. By (7), the cut-off points are strictly higher in communication equilibrium, whether or not a provocation occurs. Thus, whenever a player would have chosen H in the communication-free equilibrium, he necessarily chooses H in communication equilibrium. Moreover, after any message, there are types (of each player) who choose H, but who would have chosen D in the communication-free equilibrium. It follows that all types of players A and B are made worse off by communication, because each wants the opponent to choose D.

For player E, the welfare comparison across equilibria is ambiguous, because cheap-talk makes both players A and B more likely to choose H. Specifically, there are three cases. First, if either $c_A \leq \hat{x}$ or $c_A > \Gamma(d)$, then player A's action is the same in the communication equilibrium and in the communication-free equilibrium, but player B is more likely to choose H in the former, making player E worse off. Second, if $\hat{x} < c_A \leq y^*$, then player A would have chosen D in the communication-free equilibrium. In the communication equilibrium, there is no provocation when $\hat{x} < c_A \leq y^*$, but player A plays H rather than D, because player B is likely to choose H (the "dog that doesn't bark" effect). Third, if $y^* < c_A \leq \Gamma(d)$, then a provocation causes player A to play H, rather than D as in the communication free equilibrium. Player E gets a strictly positive payoff whenever player A chooses H, and a non-positive payoff whenever player A chooses D. Thus, player E is better off if player A switches to H.

The communication equilibrium is *unique* if the two curves $x = \Omega(y)$ and $y = \Gamma(x)$ have a *unique* intersection. This would be true, for example, if F were concave, because in this case both Ω and Γ would be concave. However, uniqueness also obtains without concavity, if a "conditional" version of

Assumption 2 holds. Intuitively, after m_0 is sent player B knows that player A's type is either below y^* or above $\Gamma(d)$. Thus, the continuation equilibrium must be the equilibrium of a "conditional" game (without communication) where it is commonly known that player A's type distribution has support $[\underline{c}, y^*] \cup (\Gamma(d), \bar{c}]$ and density

$$g(c) \equiv \frac{F'(c)}{1 - F(\Gamma(d)) + F(y^*)}$$

on this support. Furthermore, following m_0, player A's type y^* must be indifferent between choosing H and D. (If he strictly preferred H, type $y^* + \varepsilon$ would also prefer to send H following m_0, but then player E would prefer to send m_0 when player A's type is $y^* + \varepsilon$.) That is, in the "conditional" game, the cut-off type is y^*. Recall that Assumption 2 guarantees uniqueness in the "unconditional" communication-free game. The analogous condition which guarantees uniqueness in the "conditional" game is $g(y^*) < 1/(d - \mu)$. Thus, the "conditional" game has a unique equilibrium if the following "conditional" version of Assumption 2 holds:

$$\frac{F'(y)}{1 - F(\Gamma(d)) + F(y)} < \frac{1}{d - \mu} \tag{8}$$

for all $y \in (\underline{c}, \bar{c})$. More formally, it is easily verified that (8) implies $0 < \Omega'(y) < 1$. This implies, since $0 < \Gamma'(x) < 1$, that the two curves $x = \Omega(y)$ and $y = \Gamma(x)$ intersect only once, as indicated in Figure 2. Thus, as before, the requirement for uniqueness is that the distribution is sufficiently diffuse.[8]

In summary:

Theorem 2 *Suppose player E is a hawkish extremist and the game has strategic complements. A communication equilibrium exists. All types of players A and B prefer the communication-free equilibrium to any communication equilibrium. Player E is better off in communication equilibrium if and only if $\hat{x} < c_A \leq \Gamma(d)$. If (8) holds for all $y \in (\underline{c}, \bar{c})$ then there is a unique communication equilibrium.*

In the communication-free equilibrium, the probability of peace, in the sense that the outcome is DD, is $(1 - F(\hat{x}))^2$. In the communication equilibrium, DD happens with probability $(1 - \Gamma(d))(1 - F(x^*)) < (1 - F(\hat{x}))^2$.

[8] As a special case, suppose F is uniform on $[0, \bar{c}]$. Then (8) holds if $\bar{c}$ is big enough; more precisely if $\bar{c}^2 - d\bar{c} > (d - \mu)\, d$.

Thus, peace is less likely in the communication equilibrium than in the communication-free equilibrium.

To understand how the cut-off points can be uniformly higher with cheap-talk, we again interpret message m_1 as "provocation" and message m_0 as "no provocation". A provocation occurs when player A is a coordination type $c_A \in [y^*, \Gamma(d)]$ who would have played D in the communication-free equilibrium. Now, he plays H instead, and so does player B (except if he is a dominant strategy dove). The players behave aggressively following a provocation because they think the other will be aggressive, as in a "bad" equilibrium of a stag-hunt game. The fact that a provocation does *not* occur also triggers conflict, but for a different reason. In "the curious incident of the dog in the night-time" (Conan Doyle [18]), the dog did not bark at an intruder because the dog knew him well. Similarly, when player A's preferences are aligned with the hawkish extremist, there is no provocation. Hence, an "extremist who does not bark" signals the possibility that player A is a dominant strategy hawk. This information makes player B want to play H. Accordingly, the communication equilibrium has more conflict than the communication-free equilibrium, no matter which message is sent.

There is a stark contrast between the results in Baliga and Sjöström [5], where communication between the decision-makers prevented conflict, and the current results. In both cases, cheap-talk truncates the distribution of types, with a separate message sent for intermediate types and another for extreme types. In Baliga and Sjöström [5], separating out "tough" coordination types cuts the "fear-spiral" and prevents the whole population from being infected by fearfulness. The intermediate types themselves coexist peacefully. In contrast, communication by a hawkish extremist separates out "weak" coordination types, who would have played D in the communication-free equilibrium but are provoked into playing H. This brings conflict when peace could have prevailed. When there is no provocation in the communication equilibrium, the spiralling logic is even worse than before, because the absence of "weak" coordination types leads to a less favorable type-distribution.

4.3 Extensions of the Basic Model

4.3.1 Costly Messages

Our basic model assumes communication is pure cheap-talk. But in reality, in order to get the attention of decision makers, a costly message may be required. In this subsection, we show that the model is robust to assuming messages are real (costly) actions. To be specific, suppose players A and B are political leaders of countries A and B, and player E is a terrorist based in country A. The "provocative" message m_1 is a terrorist attack which imposes a cost $\tau_j > 0$ on player $j \in \{A, B, E\}$. The other message, m_0 or "no attack", involves no costs.

The terrorist does not internalize τ_A and τ_B, and as these costs are already incurred when players A and B move, they do not affect strategic behavior. We now argue that if τ_E is not prohibitively big, then the communication equilibrium exists as before. The terrorist's expected payoff from sending m_1 when player A is a susceptible type is $-c_E + (1 - F(d))\,\mu - \tau_E$, as player A plays H for sure and player B plays H unless he is a dominant strategy dove. If the terrorist instead sends m_0, then player A plays D and the terrorist's expected payoff is $-d\,(1 - F(x^*))$. Therefore, the terrorist prefers m_1 as long as

$$d\,(1 - F(x^*)) - c_E + (1 - F(d))\,\mu > \tau_E.$$

The left hand side is strictly positive, so if τ_E is not too big, the communication equilibrium of Section 4.2 still exists.

4.3.2 Renegotiation and Domestic Politics

Player E is willing to send costly message m_1 because it triggers a continuation equilibrium which is good for him but bad for players A and B. Indeed, following m_1 all types except dominant strategy doves choose H. By definition of equilibrium, no individual player can gain by deviating. However, a *joint* deviation by players A and B, an agreement to disregard the terror act, with its sunk costs τ_A and τ_B, could make them better off. That is, the equilibrium might not be "renegotiation-proof" in the sense that following message m_1 it may be common knowledge that both players A and B would prefer a different continuation equilibrium.[9] We now argue that

[9] Player E's message m reveals information about player A's type at time 2, and it is impossible for player B to simply "not listen". At time 3, when players A and B make their

a realistic modification of the model makes the communication equilibrium renegotiation-proof. To simplify the exposition, we set $\tau_E = 0$.

A political leader who is weak in the face of terrorism is less likely to stay in power. For example, Jimmy Carter lost the Presidential election in 1980 in part because he failed to deal effectively with the Iranian hostage crisis. To capture this, we modify the game by assuming player B gets an extra payoff $R > 0$ if he plays H after m_1, interpreted as the "rents" from increased popularity. Player B does *not* get R if he plays H after m_0, or if he plays D. After a terrorist attack, player B is a "conditional" dominant strategy hawk if $c_B \leq R + \mu$ and a "conditional" dominant strategy dove if $c_B \geq R + d$ (assuming $\bar{c} > R + d$ to rule out corner solutions).

The communication equilibrium of Section 4.2 is modified as follows. Player E sets $m(c_A) = m_1$ if and only if $c_A \in (y^*, \Gamma(R + d)]$. Player A's cut-off points are $c_A(m_0) = y^*$ and $c_A(m_1) = \Gamma(R + d)$. Player B's cut-off points are $c_B(m_0) = x^*$ and $c_B(m_1) = R + d$. As before, x^* and y^* must satisfy $y^* = \Gamma(x^*)$ and $x^* = \Lambda(y^*)$, where Γ is defined by (3), but Λ now depends on R as follows:

$$\Lambda(y) \equiv \frac{[1 - F(\Gamma(R + d))]\,\mu + F(y)d}{[1 - F(\Gamma(R + d))] + F(y)}$$

Again, it can be shown that the two curves $x = \Lambda(y)$ and $y = \Gamma(x)$ intersect at some (x^*, y^*), where

$$\hat{x} < y^* < x^* < \Gamma(R + d) < d. \tag{9}$$

This implies that the modified communication equilibrium exists. If

$$\frac{F'(c)}{1 - F(\Gamma(R + d)) + F(c)} < \frac{1}{d - \mu} \tag{10}$$

decisions, a continuation equilibrium consists of a cut-off point for each player such that, conditional of the information revealed by m, each cut-off is a best response to the other. In Section 4.2 we showed that inequality (8) implies a unique communication equilibrium exists. In fact, (8) also implies that a unique continuation equilibrium exists after message m_0. However, after m_1 there may be multiple continuation equilibria. (But only one, namely the fear-spiral, can be part of the unique communication equilibrium.) Since players A and B do not observe each other's types, renegotiation could be prevented by "information leakage": any player who proposes renegotiation is believed to be a dominant strategy hawk. However, the renegotiation might be proposed by a benevolent mediator, with no information leakage.

for all $c \in (\underline{c}, \bar{c})$, then it is unique. We now claim that it is renegotiation-proof, provided that $R > d - \mu$ and F is concave.

Theorem 3 *Suppose player B receives a rent $R > 0$ if he plays H after m_1. If $R+\mu > d$ and F is concave then the (modified) communication equilibrium is renegotiation proof.*

Proof. It suffices to show that following m_1 there is a *unique* continuation equilibrium, where player A playing H and player B playing H unless he is a dominant strategy dove.

A continuation equilibrium consists of a pair of cut-off points, x for player B and y for player A, that are best responses to each other, conditional on m_1 having revealed to player B that $c_A \in (y^*, \Gamma(R+d)]$. If player A uses a cutoff $y \in [y^*, \Gamma(R+d)]$, player B prefers H if and only if

$$R - c_B + \frac{\mu\left(F(\Gamma(R+d)) - F(y)\right)}{F(\Gamma(R+d)) - F(y^*)} \geq \frac{-d\left(F(y) - F(y^*)\right)}{F(\Gamma(R+d)) - F(y^*)}. \qquad (11)$$

Inequality (11) is equivalent to $c_B \leq \Theta(y)$ where

$$\Theta(y) \equiv \frac{(d-\mu)\,F(y)}{F(\Gamma(R+d)) - F(y^*)} + R + \frac{\mu F(\Gamma(R+d))}{F(\Gamma(R+d)) - F(y^*)} - \frac{dF(y^*)}{F(\Gamma(R+d)) - F(y^*)}.$$

Thus, player B's best response is $x = \Theta(y) \in [R+\mu, R+d]$. (Recall that types below $R+\mu$ or above $R+d$ are dominant strategy types.)

Player A's best response to x is given by Γ. If $R+\mu > d$ then $\Gamma(R+\mu) > y^* = \Gamma(x^*)$. To see this, notice that $R+\mu > d$ implies

$$R + \mu > x^* = \frac{\left[1 - F(\Gamma(R+d))\right]\mu + F(y^*)d}{\left[1 - F(\Gamma(R+d))\right] + F(y^*)} \qquad (12)$$

Thus, $\Gamma(R+\mu) > y^*$, and since Γ is increasing, player A's best response to $x \geq R+\mu$ is $y = \Gamma(x) > y^*$.

So far we have shown that in continuation equilibrium, the cut-off points satisfy $x = \Theta(y) \geq R+\mu$ and $y = \Gamma(x) > y^*$. In fact, the curves Γ and Θ intersect at $x = R+d$ and $y = \Gamma(R+d)$, and this yields the strategy played in the modified communication equilibrium: after message m_1, player A plays H for certain (i.e., all types $c_A \in (y^*, \Gamma(R+d)]$ play H) and player B plays H if $c_B \leq R+d$ (i.e., unless he is a conditional dominant strategy

dove). The curves can have no other intersection: if F is concave, both $\Gamma(x)$ and $\Theta(y)$ are concave and can intersect at most once in the relevant region (i.e., $x \in [R+\mu, R+d]$ and $y^* \in [y^*, \Gamma(R+d)]$. ■

Naturally, player B's domestic political concerns make him more aggressive. Theorem 3 shows that the induced reciprocal fear among players A and B can lead to a unique continuation equilibrium. In fact it can be reached by iterated deletion of dominated strategies. The presence of dominant strategy hawks of player B makes coordination types of player A with costs close to y^* prefer H. This then leads to higher and higher types of both players playing H till the equilibrium is reached. Notice that this process is triggered by the presence of dominant strategy hawks on the side of *player* B. That is, it is the aggressive response of player B to a provocation that escalates fear in player A.

4.3.3 Domestic Politics and Uninformed Cheap-Talk

In our basic model, player E knows player A's type. In this subsection, we instead consider provocations by uninformed agents. Al Qaeda operatives may or may not have good information about the population of Yemen, Afghanistan etc. If they don't, they may still be able to trigger the reciprocal fears induced by domestic political concerns of American leaders. Specifically, suppose (in this subsection only) that player E does not know c_A, and consider the marginal impact of provocation on player E's payoff at the communication-free equilibrium. Suppose player i uses a cutoff $\hat{x}_i$. If player A plays H, then player E gets $-c_E + \mu(1 - F(\hat{x}_B))$, and if player A plays D then player E gets $-dF(\hat{x}_B)$. Hence, player E's expected payoff is

$$F(\hat{x}_A)\left[-c_E + \mu(1 - F(\hat{x}_B))\right] - (1 - F(\hat{x}_A))\, dF(\hat{x}_B).$$

Suppose, as in the previous subsection, terrorism makes player B more aggressive because of electoral incentives. The direct effect on player E's payoff from a marginal increase in $\hat{x}_B$ (holding $\hat{x}_A$ constant) is negative, and equals

$$\frac{dF(\hat{x}_B)}{d\hat{x}_B}\left[-\mu F(\hat{x}_A) - (1 - F(\hat{x}_A))\, d\right] < 0. \tag{13}$$

The direct effect is smaller, the more aggressive is player A and the higher is $\hat{x}_A$.

The strategic (indirect) effect is to make player A more aggressive in response to player B's increased aggression. The strategic effect on player

E's payoff from a marginal increase in $\hat{x}_A$ is positive, and equals

$$\frac{dF(\hat{x}_A)}{d\hat{x}_A}\left[-c_E+\mu\left(1-F(\hat{x}_B)\right)+dF(\hat{x}_B)\right]\frac{d\hat{x}_A}{d\hat{x}_B}>0. \tag{14}$$

The strategic effect is greater, the more aggressive is player B and the higher is $\hat{x}_B$.

The expressions (13) and (14) suggest that provocation would be profitable for player E at a communication-free equilibrium where $\hat{x}_A$ and $\hat{x}_B$ are already quite high. More precisely, at a symmetric equilibrium where $\hat{x}_A=\hat{x}_B=\hat{x}$, the net effect of provocation on player E's payoff is

$$\begin{aligned}
&\frac{dF(\hat{x}_A)}{d\hat{x}_A}\left[-c_E+\mu\left(1-F(\hat{x}_B)\right)+dF(\hat{x}_B)\right]\frac{d\hat{x}_A}{d\hat{x}_B}\\
&+\frac{dF(\hat{x}_B)}{d\hat{x}_B}\left[-\mu F(\hat{x}_A)-\left(1-F(\hat{x}_A)\right)d\right]\\
=\;&\frac{dF(\hat{x})}{d\hat{x}}\left[-c_E+\hat{x}\right]\frac{d\hat{x}_A}{d\hat{x}_B}+\frac{dF(\hat{x})}{d\hat{x}}\left[\hat{x}-\mu-d\right]
\end{aligned}$$

(where the equality uses (3)). From (3), $d\hat{x}_A/d\hat{x}_B=(d-\mu)\,f(\hat{x}_B)$, so at the symmetric equilibrium, provocation is profitable if and only if

$$\hat{x}>c_E+\frac{1}{(d-\mu)\,f(\hat{x})}\left(\mu+d\right).$$

Thus, provocation pays if tension (i.e., the probability each player chooses H) is already high at the communication-free equilibrium. With specific distributions, the condition can be characterized more explicitly.[10] Intuitively, provocation should be more profitable where there are already deep-seated suspicions and fears (perhaps Ireland and Sri Lanka might be examples). In contrast, a provocation where tension is low may backfire. First, as the primary home audience, player A, is relatively dovish, there is a bigger chance that a more aggressive secondary audience catches player A unaware, which is costly to the provocateur. The direct negative effect is therefore large. Second, as the secondary audience is relatively dovish, there are smaller costs to dovish behavior by player A and hence less gain from trying to make player A more aggressive. The strategic positive effect is therefore small.

[10] If F is uniform, the condition is

$$\frac{(\bar{c}-\underline{c})\,\mu-\underline{c}\,(d-\mu)}{(\bar{c}-\underline{c})-(d-\mu)}>c_E+\frac{(\bar{c}-\underline{c})}{(d-\mu)}\left(\mu+d\right).$$

5 Cheap-Talk with Strategic Substitutes

In this section, we consider the case of strategic substitutes, $d < \mu$.

5.1 Hawks can't Communicate Effectively

A hawkish extremist cannot communicate effectively when actions are strategic substitutes. From Lemma 1, $M = \{m_0, m_1\}$ with $c_B(m_1) > c_B(m_0)$. The hawkish extremist wants player A (but not player B) to play H, so he would only send m_1 if this message causes player A to play H. But if player A plays H for sure after m_1, then player B's cut-off point is $c_B(m_1) = d$. But, with $d < \mu$, types below d are dominant strategy types who always play H, so we cannot have $c_B(m_0) < d$, a contradiction. Thus, we have:

Proposition 2 *If player E is a hawkish extremist and the game has strategic substitutes, then cheap-talk cannot be effective.*

When actions are strategic substitutes, the message m_1 which makes player B more likely to play H must make player A more likely to play D. But a message which causes player A to back down in this way will never be sent by a hawkish extremist, and this makes the hawkish extremist unable to communicate effectively.

5.2 Dovish Cheap-Talk

Now suppose player E is a dovish extremist and the game has strategic substitutes. We will construct a communication equilibrium where the dovish extremist E sends informative messages. Again, it is surprising that this can be done because c_E is commonly known. To understand the communication equilibrium intuitively, it helps to again recall Lemma 1, but now interpret message m_1 as a "peace rally" and message m_0 as "no peace rally". Intuitively, the peace rally will make player B more aggressive, and player A backs down and chooses D.

Again, say that player A is a susceptible type if his action depends on which message is sent. But now, susceptible types switch from H to D when they hear message m_1. That is, the set of susceptible types is

$$S \equiv (c_A(m_1), c_A(m_0)].$$

The proof of Lemma 1 showed that if $m(c_A) = m_1$ then type c_A must be susceptible. Intuitively, since peace demonstrations make player B more likely to choose H, player E would not engage in them unless player A is a susceptible type. Conversely, whenever player A is a susceptible type, the dovish extremist will engage in peace demonstrations, since he wants player A to choose D. Therefore, $m(c_A) = m_1$ if and only if $c_A \in S$. Accordingly, message m_1 signals that player A will choose D. As argued in the proof of Lemma 1, this implies $c_B(m_1) = \mu$, and player A's best response to this cut-off point is $c_A(m_1) = \Gamma(\mu)$.

It remains only to consider how players A and B behave when there is no peace demonstration (message m_0). Let $y^* = c_A(m_0)$ and $x^* = c_B(m_0)$ denote the cutoff points used in this case. Arguing as for the case of strategic complements, the cut-off points must satisfy $y^* = \Gamma(x^*)$ and $x^* = \tilde{\Omega}(y^*)$, where

$$\tilde{\Omega}(y) \equiv \frac{[1 - F(y)]\,\mu + F(\Gamma(\mu))d}{[1 - F(y)] + F(\Gamma(\mu))}$$

As shown in Figure 3, (x^*, y^*) is an intersection of the two curves $x = \tilde{\Omega}(y)$ and $y = \Gamma(x)$. With strategic substitutes, Assumption 2 implies

$$-1 < \Gamma'(x) < 0$$

Furthermore, $\Gamma(\underline{c}) = \mu < \bar{c}$ and $\Gamma(\bar{c}) = d > \underline{c}$, and

$$\Gamma(\mu) - d = (1 - F(\mu))\,(\mu - d)$$

where

$$0 < (1 - F(\mu))\,(\mu - d) < \mu - d.$$

Therefore,

$$d < \Gamma(\mu) < \mu \tag{15}$$

Let $\hat{x}$ be the unique fixed point of $\Gamma(x)$ in $[\underline{c}, \bar{c}]$. Clearly, $d < \hat{x} < \mu$ (see Figure 3).

Figure 3 shows three curves: $x = \tilde{\Omega}(y)$, $y = \Gamma(x)$ and $x = \Gamma(y)$. The curves $x = \Gamma(y)$ and $y = \Gamma(x)$ intersect on the 45 degree line at the fixed point $\hat{x} = \Gamma(\hat{x})$. It is easy to check that $\tilde{\Omega}(y) > \Gamma(y)$ whenever $y \in (\Gamma(\mu), \bar{c})$. Moreover,

$$\tilde{\Omega}(\bar{c}) = \Gamma(\bar{c}) = d$$

and

$$\tilde{\Omega}(\Gamma(\mu)) = \Gamma(\Gamma(\mu)) > \Gamma(\mu)$$

where the inequality follows from (15) and the fact that Γ is decreasing. Consider now (x^*, y^*) such that $y^* = \Gamma(x^*)$ and $x^* = \tilde{\Omega}(y^*)$, i.e., the intersection of the two curves $x = \tilde{\Omega}(y)$ and $y = \Gamma(x)$. Figure 3 reveals that there exists $(x^*, y^*) \in [\underline{c}, \overline{c}]^2$ such that $y^* = \Gamma(x^*)$ and $x^* = \tilde{\Omega}(y^*)$, and

$$d < \Gamma(\mu) < y^* < \hat{x} < x^* < \mu. \tag{16}$$

Thus, a communication equilibrium exists. What impact do pacifist messages have on the probability of conflict? In the communication-free equilibrium, each player's cutoff is $\hat{x}$. Now (16) reveals that with pacifist communication, player B's cutoff points x^* and μ are strictly greater than $\hat{x}$. Thus, communication makes player B more aggressive, whatever message is actually sent. On the other hand, player A's cutoff points y^* and $\Gamma(\mu)$ are strictly smaller than $\hat{x}$. Thus, communication makes player A less aggressive ("better red than dead"), whatever message is actually sent. Since one player becomes more and the other less aggressive, it is not possible to unambiguously say if communication is good or bad for peace.

The welfare effects are unambiguous, however. As player A is more likely to play D in the communication equilibrium, player B is made better off. Conversely, as player B is more likely to play H, player A is made worse off. The pacifist (dovish extremist) is made better off by the peace rally when it occurs, because it prevents player A from choosing H. On the other hand, the "dog that did not bark" effect makes player B more likely to choose H when there is no peace rally, and this makes player E worse off.

Finally, consider whether the communication equilibrium is unique. The same argument as in Section (4.2) implies that we must impose a "conditional" version of Assumption 2. Specifically,

$$\frac{F'(y)}{1 - F(y) + F(\Gamma(\mu))} < \frac{1}{\mu - d} \tag{17}$$

for all $y \in (\underline{c}, \overline{c})$. It can be checked that (17) implies $-1 < \tilde{\Omega}'(y) < 0$. In this case, since $-1 < \Gamma'(x) < 0$, the two curves $x = \tilde{\Omega}(y)$ and $y = \Gamma(x)$ intersect only once, as indicated in Figure 3. In summary:

Theorem 4 *Suppose player E is a dovish extremist and the game has strategic substitutes. A communication equilibrium exists. All of player A's types prefer the communication-free equilibrium to the communication equilibrium. All of player B's types have the opposite preference. Player E is better off*

in the communication equilibrium if and only if $\Gamma(\mu) \leq c_A < \hat{x}$. *If (17) holds for all* $y \in (\underline{c}, \overline{c})$ *then there is a unique communication equilibrium.*

Theorem 4 is in stark contrast to Theorem 2. With strategic complements, provocations caused both players A and B to become more aggressive, and hence both became worse off. With strategic substitutes, player B benefits from peace rallies in country A, because they make player A back down.

6 Strategic Effects of Ex Ante Investment

Suppose a decision maker can make a publicly observed investment which changes his country's military capability. He might invest in offensive weapons that increase the chances of military victory. Alternatively, he might invest in anti-missile defense systems, or build fortifications that make it less costly to be attacked. To study this, we generalize the model to allow for *ex ante* asymmetries.

The parameters μ and d, and the distribution over cost-types, are now player-dependent. The payoff of player $i \in \{A, B\}$ is given by the following payoff matrix, where the row represents his own choice, and the column represents the choice of player j.

$$\begin{array}{cc} H & D \\ -c_i & \mu_i - c_i \\ -d_i & 0 \end{array} \begin{array}{c} H \\ D \end{array} \tag{18}$$

Player i's type c_i is drawn from a distribution F_i with support $[\underline{c}_i, \overline{c}_i]$. As before, types are independently drawn. In the communication-free equilibrium, equilibrium cutoff points $(\hat{x}_A, \hat{x}_B)$ solve the two equations

$$\hat{x}_A = \mu_A + (d_A - \mu_A) F_B(\hat{x}_B) \tag{19}$$

$$\hat{x}_B = \mu_B + (d_B - \mu_B) F_A(\hat{x}_A) \tag{20}$$

If the obvious analog of Assumption 1 holds and if $F_i'(c_i) < \left| \frac{1}{d_i - \mu_i} \right|$ for $i \in \{A, B\}$ (the analog of Assumption 2), then the communication-free equilibrium is unique by the same argument as in Theorem 1.

Consider the strategic effects of ex ante investment in communication-free equilibrium. Suppose player B, at time 0, can make a publicly observed investment which increases μ_B. This may represent, for example, increased offensive capability. After the investment, the communication-free equilibrium is played (as given by equations (19) and (20)). The investment increases player B's benefit from choosing H, and hence makes player B appear tough (it shifts his best response curve to the right). The strategic effect of the investment is its impact on the behavior of player A. Fudenberg and Tirole [31] classify strategic effects in four categories: Top Dog, Puppy Dog, Fat Cat, and Lean-and-Hungry-Look. These effects differ in whether investment makes a player "soft" or "tough" or whether there is an incentive to "overinvest" or "underinvest". With strategic complements, shifting player B's best response curve to the right causes both $\hat{x}_A$ and $\hat{x}_B$ to increase. Since player B wants player A to choose D, the strategic effect is negative: player B prefers to underinvest in order to appear soft (Puppy Dog strategy). With strategic substitutes, the strategic effect is instead positive: player B then prefers to overinvest in order to appear tough (Top Dog strategy).

Suppose instead the investment reduces d_B. This may represent, for example, better defensive abilities of country B, making it less vulnerable to an attack. This investment will raise player B's benefit from choosing D, and hence make player B appear soft (it shifts his best response curve to the left). With strategic complements, both $\hat{x}_A$ and $\hat{x}_B$ decrease. Thus, the strategic effect is positive: player B prefers to overinvest in order to appear soft (Fat Cat strategy). With strategic substitutes, the strategic effect is instead negative: player B underinvests in order to appear tough (Lean and Hungry Look).

To summarize, in communication-free equilibrium the strategic effects are straightforward. In a game of chicken, there would be an incentive to overinvest in offensive capability in order to intimidate the opponent and force him to back down. In a stag-hunt game, there would be an incentive to overinvest in defensive capability in order to reassure the opponent that one is unlikely to attack out of fear. Now we turn to the case when an extremist observes the investment and can communicate.

Observe that Lemma 1 is still valid in the asymmetric environment. In communication equilibrium, generalized to allow for ex ante asymmetries, player B's publicly observed investment influences player A not only directly but also indirectly, via changes in player E's behavior. Nevertheless, with strategic complements, the strategic effects turn out to be the same as dis-

cussed above: the optimal strategies are still Puppy Dog and Fat Cat, making oneself look less threatening. However, with strategic substitutes, the presence of a dovish extremist dramatically changes the strategic effects. The dovish extremist is, in a sense, an "ally" of player B, because peace protests make player A back down. In this case, Top Dog and Lean and Hungry Look strategies can backfire for player B: by overinvesting in offensive capacity (or underinvesting in defensive capacity), player B alarms the pacifist, who organizes fewer peace protests. The net effect may be to make player B worse off. We now formalize these arguments.

6.1 Strategic Complements

Suppose $d_i > \mu_i$ for $i \in \{A, B\}$ and player E is a hawkish extremist. Define

$$\Gamma_A(x) \equiv \mu_A + F_B(x)\,(d_A - \mu_A) \tag{21}$$

and

$$\Omega_B(y) \equiv \frac{[1 - F_A(\Gamma_A(d_B))]\,\mu_B + F_A(y)d_B}{[1 - F_A(\Gamma_A(d_B))] + F_A(y)} \tag{22}$$

Now let $x_B^* = \Omega_B(y_A^*)$ and $y_A^* = \Gamma_A(x_B^*)$. Arguing as in Section 4.2, if $F_A'(c_A) < (1 - F_A(\Gamma(d)))/(d - \mu)$ for all c_A then there exists a unique pair (x_B^*, y_A^*) such that $y_A^* = \Gamma_A(x_B^*)$ and $x_B^* = \Omega_B(y_A^*)$. Moreover, $\hat{x}_B < x_B^* < d_B$ and $\hat{x}_A < y_A^* < \Gamma_A(d_B) < d_A$.The strategies are the obvious generalizations of the strategies in Section 4.2. Player E sends the message $m(c_A) = m_1$ if and only if $y_A^* < c_A \leq \Gamma_A(d_B)$. Player A's cut-off points are $c_A(m_1) = \Gamma_A(d_B)$ and $c_A(m_0) = y_A^*$. Player B's cut-off points are $c_B(m_0) = x_B^*$ and $c_B(m_1) = d_B$. Notice that, in equilibrium, player A chooses H if and only if $c_A \leq \Gamma_A(d_B)$.

Suppose player B, at time 0, makes a publicly observed investment which increases μ_B. This shifts the Ω_B function to the right: player B becomes "tough". Since $\Gamma_A(d_B)$ does not depend on μ_B, the set of types of player A that choose H does not change. However, the cutoff y_A^* increases when Ω_B shifts, so message m_1 is sent less often. Intuitively, with a higher μ_B, the hawkish extremist has less reason to send m_1, because player A is anyway very inclined to choose H when player B is tough. The message m_1 corresponds to a "barking dog" that reveals that player A will choose H. Because this information is valuable to player B, the strategic effect is negative, and player B will underinvest (Puppy Dog Ploy).

Suppose instead that player B's publicly observed investment reduces d_B. Then $\Gamma_A(d_B)$ falls, so the set of types of player A that choose H shrinks. This

strategic effect is positive for player B. Moreover, the investment shifts the Ω_B function to the left, so y_A^* falls, say to $y_A^* - \varepsilon$. The "bark" m_1 that reveals player A's action now sounds for types in the interval $[y^* - \varepsilon, y^*]$. This is also positive for player B. Thus, both effects make player B better off, so he will overinvest (Fat Cat strategy).

To summarize, with strategic complements, the Puppy Dog Ploy and the Fat Cat strategy are optimal for player B whether or not there is communication by a hawkish extremist. That is, player B has an incentive to underinvest in offensive capability and overinvest in defensive capability, either to make player E's messages more informative or to make player A less aggressive.

6.2 Strategic Substitutes

Suppose $d_i < \mu_i$ for $i \in \{A, B\}$ and player E is a dovish extremist. Define $\Gamma_A(x)$ as in (21), and

$$\Omega_B(y) \equiv \frac{[1 - F_A(y)]\,\mu_B + F_A(\Gamma_A(\mu_B))d_B}{[1 - F_A(y)] + F_A(\Gamma_A(\mu_B))}.$$

Now let $x_B^* = \Omega_B(y_A^*)$ and $y_A^* = \Gamma_A(x_B^*)$. Arguing as in Section 5.2, if $F_A'(c_A) < F_A(\Gamma(\mu))/(\mu - d)$ for all c_A then there exists a unique pair (x_B^*, y_A^*) such that $y_A^* = \Gamma_A(x_B^*)$ and $x_B^* = \Omega_B(y_A^*)$. Moreover, $d_A < \Gamma_A(\mu_B) < y_A^*$ and $\hat{x}_B < x_B^* < \mu_B$. The strategies are the obvious generalizations of the strategies in Section 5.2. Player E sends the message $m(c_A) = m_1$ if and only if $\Gamma_A(\mu_B) < c_A \leq y_A^*$. Player A's cut-off points are $c_A(m_0) = y_A^*$ and $c_A(m_1) = \Gamma_A(\mu_B)$. Player B's cut-off points are $c_B(m_0) = x_B^*$ and $c_B(m_1) = \mu_B$. Notice that, in equilibrium, player A chooses H if and only if $c_A \leq \Gamma_A(\mu_B)$.

Suppose player B, at time 0, makes a publicly observed investment which reduces d_B. This shifts the Ω_B function to the left: player B becomes "soft". Since $\Gamma_A(\mu_B)$ does not depend on d_B, the set of types of player A that choose H does not change. However, the cutoff y_A^* increases when Ω_B shifts, so message m_1 is sent more often. Intuitively, a lower d_B encourages player A to choose H (to take advantage of the not-so-tough player B) but to counter that, the dovish extremist organizes peace protests. Because m_1 is an informative signal that reveals that player A will choose D, the fact that m_1 is sent more often makes player B better off (it becomes easier to exploit player A). This means that the strategic effect is positive, and player B

will overinvest to look soft (Fat Cat). Recall that if the extremist is not present, the Lean and Hungry Look is optimal. Thus, the presence of the dovish extremist flips the strategic effect in the opposite direction. Intuitively, player B and the pacifist have a common interest: to make player A back down. The pacifist becomes more inclined to "help" player B when player B is soft, and this produces the Fat Cat effect.

Suppose instead that player B's publicly observed investment increases μ_B. Then $\Gamma_A(\mu_B)$ falls, so the set of types of player A that choose H shrinks. This strategic effect is positive for player B. However, the investment shifts the Ω_B function to the right, so y_A^* falls, say to $y_A^* - \varepsilon$. Intuitively, with a higher μ_B, the dovish extremist has less reason to organize peace protests, because player A is anyway more inclined to choose D when player B has become tough. Because m_1 is an informative signal that alerts player B that player A is about to choose D, the fact that m_1 is sent less often makes player B worse off. Thus, in this case there are two strategic effects which go in opposite directions. Increasing μ_B has a direct effect on player A, making him more likely to back down, and this benefits player B. But the indirect effect (fewer peace protests) hurts player B. In general, we cannot say if the Top Dog strategy or Puppy Dog Ploy is optimal.

To summarize, with strategic substitutes, the presence of a dovish extremist changes the strategic effects in an interesting way. Player B has less of an incentive to behave aggressively (Top Dog or Lean and Hungry Look) because this would, in effect, make the pacifists in country A less "cooperative" (from player B's perspective). Instead, he has an incentive to overinvest in defensive technology (Fat Cat strategy). The strategic effect of an increased offensive ability cannot be signed.

7 Conclusion

In previous work we argued that when actions are strategic complements, direct face-to-face communication allows "tough" moderates, who would have chosen H in the communication-free equilibrium, to coordinate on D (Baliga and Sjöström [5]). Here, we have instead considered cheap-talk by a third party, interpreted as provocative acts and speech by hawkish extremists, or peace rallies by pacifists. We found that hawkish extremists are either bad for peace (when actions are strategic complements) or irrelevant (when actions are strategic substitutes). Dovish extremists are either irrelevant (strategic

complements) or have an ambiguous impact because they make one player more aggressive while the other backs down (strategic substitutes).

In all cases, informative cheap-talk has a non-convex structure: message m_1 identifies a subset of moderate (intermediate) types of player A. With strategic complements, the hawkish extremist causes "weak" moderates, who would have chosen D in the communication-free equilibrium, to choose H instead, by the indirect route of inflaming the enemy (player B). This does seem to approximate the strategy of at least some real-world extremist groups. For example, I.S.I.-sponsored terrorist attacks seem not to be designed to directly force India to withdraw from Kashmir, which would be an unrealistic goal, but rather to increase tensions between India and Pakistan and perhaps trigger a broader conflict. Of course, other real-world extremists seem to have the more straightforward objective of making the enemy back down by inflicting pain, as when the terrorist group Irgun helped to drive the British out of Palestine (Cohen [20]).

Our theory implies that hawkish extremists engage in rational provocation only when there are real opportunities for peace. Otherwise, it would be counter-productive. The provocations are meant to increase tensions and trigger a spiral of fear between players A and B. But if player A is not responsive, for example if he is a radical dominant-strategy type himself, provocative acts would be counter-productive. According to this theory, the takeover of the American embassy by Iranian radicals would signal that the Iranian leaders are *not* dominant-strategy hawks (but instead moderates capable of turning aggressive in self-defense). Similarly, Hamas's attacks during the Oslo peace accords and before Israeli elections would signal that the leaders of the Palestinian Authority are moderates who, unlike the Hamas, want peace.

Because the hawkish extremist provokes conflict, players A and B have a common interest in suppressing him. If suppression is impossible, why not simply "choose" not to respond to provocation?

> "Terrorism wins only if you respond to it in the way that the terrorists want you to; which means that its fate is in your hands and not in theirs. If you choose not to respond at all, or else to respond in a way different from that which they desire, they will fail to achieve their objectives. The important point is that the choice is yours. That is the ultimate weakness of terrorism

> as a strategy. It means that, though terrorism cannot always be prevented, it can always be defeated. You can always refuse to do what they want you to do." (David Fromkin [30], p. 697)

Unfortunately, our communication equilibrium is logically consistent with rational behavior: no player can gain by a *unilateral* deviation. In reality, this logic is reinforced by the desire of political leaders not to look weak in the face of terrorism. The question of whether players A and B can *jointly* deviate by simply "disregarding" provocations was discussed in Section 4.3.2, where we argued that this is not necessarily the case.

If actions are strategic substitutes, the "peace rally" identifies "tough" moderates, who would have chosen H in the communication-free equilibrium, but now back down and choose D, making the opponent more likely to choose H. The pacifists are better off, because the HH outcome is avoided ("better red than dead"). Player B benefits from the peace rally, but player A would like to suppress it. However, as we saw in Section 6.2, the negative impact of peace rallies on player A is mitigated when player B's ex ante investment is taken into account. For the pacifists in country A to "cooperate" with player B, player B cannot appear too threatening. Hence, player B's (publicly observed) investment will be skewed towards "defensive" measures, and this is good for player A.

In this article, we studied how extremists may deliberately inflame tensions between two antagonistic groups. This does not only happen during international disputes. In the early part of the twentieth century, African-Americans and Irish-Americans moved to the same areas of Chicago, competing for the same jobs. The two groups viewed each other with mutual suspicion. Blacks believed "white men [had] great boxes of guns and ammunition in the cellars of their homes and that white men [were] shooting clubs for the purpose of shooting Negroes in the event of another riot" ([17], p. 21-22). Similarly, whites believed blacks were accumulating weapons. The result was a classic fear spiral, a tinderbox waiting to ignite, with each group ready to use violence in self-defense. The black newspaper *The Whip* warned the white community: "We are not pacifists, therefore we believe in war, but only when all orderly civil procedure has been exhausted and the points in question are justifiable" (Tuttle [57], p. 282). In 1919, provocative acts by so-called "athletic clubs", dominated by extremist Irish-Americans, caused wide-spread rioting ([17], p. 11-17). These "athletic clubs" deliberately ignited the tinderbox, hoping to drive the African-Americans away from their

neighborhoods (Tuttle [57]).

References

[1] Abu Bakr Naji (2004): *The Management of Savagery,* Olin Institute for Strategic Studies, Harvard University.

[2] Anul Aneja (2008): "Mumbai Attacks a Diversion Tactic: Analyst," *The Hindu* (December 17).

[3] Aumann, Robert (1990). "Nash Equilibria are Not Self-Enforcing", in J. J. Gabszewicz, J.-F. Richard and L. A. Wolsey (eds.), *Economic Decision-Making: Games, Econometrics and Optimization* (Amsterdam: Elsevier).

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

[5] Sandeep Baliga and Tomas Sjöström (2004): "Arms Races and Negotiations," *Review of Economic Studies,* Vol. 17, No. 1, pp. 129-163.

[6] Sandeep Baliga and Tomas Sjöström (2008): "Strategic Ambiguity and Arms Proliferation," *Journal of Political Economy* 116: 1023-1057

[7] Sandeep Baliga and Tomas Sjöström (2008): "Bargaining Foundations of Conflict Games," mimeo, Northwestern University.

[8] Heski Bar-Isaac and Mariagiovanna Baccara (2008): "How to organize crime," *Review of Economic Studies*, Volume 75(4), 1039–1067.

[9] Eli Berman (2003): "Hamas, Taliban, and the Jewish Underground: An Economists View of Radical Religious Militias." UC San Diego typescript.

[10] Claude Berrebi and Esteban Klor (2006): "On Terrorism and Electoral Outcomes: Theory and Evidence from the Israeli-Palestinian Conflict," *Journal of Conflict Resolution*, Vol. 50(6): 899-925.

[11] Claude Berrebi and Esteban Klor (2008): "Are Voters Sensitive to Terrorism? Direct Evidence from the Israeli Electorate," *American Political Science Review*, 102(3): 279-301.

[12] Mark Burgess (2003): "A Brief History of Terrorism", Center for Defense Information

[13] Ethan Bueno de Mesquita (2005): "The Quality of Terror." *American Journal of Political Science* 49(3):515–530.

[14] Ethan Bueno de Mesquita (2009) "Regime Change and Revolutionary Entrepreneurs," mimeo, Harris School, Chicago

[15] Ethan Bueno de Mesquita and Eric Dickson (2007): "The Propaganda of the Deed: Terrorism, Counterterrorism, and Mobilization." *American Journal of Political Science* 51(B):364-381.

[16] Ethan Bueno de Mesquita (2008): "The Political Economy of Terrorism: A Selective Overview of Recent Work," *The Political Economist* 10(1):1-12.

[17] Chicago Commission on Race Relations (1919): *The Negro in Chicago: A Study of Race Relations and a Race Riot,* University of Chicago Press: Chicago.

[18] Arthur Conan Doyle (1894): *The Memoirs of Sherlock Holmes*, George Newnes, London, United Kingdom.

[19] Martha Crenshaw (1981): "The Causes of Terrorism," *Comparative Politics* 13, 379-399.

[20] Cohen, M. (1987): *The Origins and Evolution of the Arab-Zionist Conflict*, University of California Press, Berkeley and Los Angeles

[21] Vincent Crawford and Joel Sobel (1982): "Strategic Information Transmission," *Econometrica,* 50: 1431-1451.

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

[23] Sylvain Chassang and Gerard Padro-i-Miguel (2008): "Conflict and Deterrence under Strategic Risk," mimeo, Princeton.

[24] Sylvain Chassang and Gerard Padro-i-Miguel (2009): "Economic Shocks and Civil War," mimeo, Princeton.

[25] Rui de Figueiredo and Barry Weingast (2001): "Vicious Cycles: Endogenous Political Extremism and Political Violence," mimeo, Berkeley and Stanford.

[26] Avinash Dixit and Barry Nalebuff (2008): *The Art of Strategy*, W.W. Norton and Company, New York.

[27] Chris Edmond (2007): "Information Manipulation, Coordination and Regime Change," mimeo, NYU

[28] Joseph Farrell and Robert Gibbons (1989): "Cheap Talk with Two Audiences," *American Economic Review*, 79, 1214-23.

[29] Joseph Farrell and Robert Gibbons (1989): "Cheap Talk Can Matter in Bargaining," *Journal of Economic Theory*, 48, 221-37.

[30] David Fromkin (1975): "The Strategy of Terrorism," *Foreign Affairs*, 53(4): 683-698.

[31] Drew Fudenberg and Jean Tirole (1984): "The Fat-Cat Effect, the Puppy-Dog Ploy, and the Lean and Hungry Look," *American Economic Review*, 74 (2) :361-366.

[32] Maria Goltsman and Gregory Pavlov (2008): "How to Talk to Multiple Audiences?," mimeo, University of Western Ontario.

[33] Sanjeev Goyal and Adrien Vigier (2010): "Robust Networks," mimeo, Cambridge University.

[34] Nir Hefetz and Gadi Bloom (2006): *Ariel Sharon*, Random House, New York.

[35] Bruce Hoffman (2006): *Inside Terrorism*, Columbia University Press: New York.

[36] Lawrence R. Iannaccone and Eli Berman (2006): "Religious Extremists: The Good, the Bad and the Deadly." *Public Choice* 128(1–2):109–129.

[37] David Jaeger and Daniele Paserman (2006): "Israel, the Palestinian Factions and the Cycle of Violence." (with David Jaeger). *American Economic Review*, 96(2), pages 45-49.

[38] David Jaeger and Daniele Paserman (2007): "The Shape of Things to Come? Assessing the Effectiveness of Suicide Bombings and Targeted Killings," mimeo

[39] David Jaeger and Daniele Paserman (2008): "The Cycle of Violence? An Empirical Analysis of Fatalities in the Palestinian-Israeli Conflict," *American Economic Review.*

[40] David Jaeger, Esteban Klor, Sami Miaari and Daniele Paserman (2008): ""The Struggle for Palestinian Hearts and Minds: Violence and Public Opinion in the Second Intifada," mimeo, Boston University.

[41] Robert Jervis (1978): "Cooperation Under the Security Dilemma," *World Politics*, Vol. 30, No. 2., pp. 167-214.

[42] Hanjoon Michael Jung (2007): "Strategic Information Transmission through the Media," Working Paper, Lahore University.

[43] Andrew Kydd and Barbara F. Walter (2002): "Sabotaging the Peace: The Politics of Extremist Violence," *International Organization* 56(B):263–296.

[44] Andrew Kydd and Barbara F. Walter (2006): "The Strategies of Terrorism." *International Security*, 31 (A): 49-79.

[45] Levy, G. and R. Razin (2004): "It takes Two: An Explanation for the Democratic Peace," *Journal of the European economic Association* 2:1-29.

[46] Steven Matthews (1989): "Veto Threats: Rhetoric in a Bargaining Game," *Quarterly Journal of Economics*, 104.

[47] Steven Matthews and Andrew Postlewaite (1989): "Pre-play communication in two-person sealed-bid double auctions," *Journal of Economic Theory* 48, 238-263.

[48] Joseph Nye (2007): *Understanding International Conflict (6th Edition). Longman Classics in Political Science.* Longman: New York City

[49] Peter Ordershook and Thomas Palfrey (1988): "Agendas, Strategic Voting, and Signaling with Incomplete Information," *American Journal of Political Science*, v. 32, #2, pp. 441-66.

[50] Gerard Padro-i-Miguel and Pierre Yared (2009): "The Political Economy of Indirect Control," mimeo, Columbia University.

[51] Angel Rabasa, Robert D. Blackwill, Peter Chalk, Kim Cragin, C. Christine Fair, Brian A. Jackson, Brian Michael Jenkins, Seth G. Jones, Nathaniel Shestak, and Ashley J. Tellis (2009): "The Lessons of Mumbai," RAND.

[52] Bruce Riedel (2008): "How 9/11 is Connected to December 13," *Hindustan Times* (September 11).

[53] Nigel Rees (2002): *Mark my words: Great Quotations and the stories behind them*, Barnes and Noble.

[54] Thomas Schelling (1960): *The Strategy of Conflict.* Cambridge: Harvard University Press.

[55] Jaswant Singh (2006): *In Service of Emergent India: A Call to Honor*, Bloomington: Indiana University Press.

[56] A. Michael Spence (1973): "Job Market Signaling," *Quarterly Journal of Economics* 87(3): 355-374.

[57] William M. Tuttle (1970): "Contested Neighborhoods and Racial Violence: Prelude to the Chicago Riot of 1919," *The Journal of Negro History* 55(4): 266-288.

[58] Paddy Woodsworth (2001): "Why do they Kill? The Basque Conflict in Spain," *World Policy Journal,* 18(1).

Figure 1. Strategic Complements:
Communication-Free Equilibrium

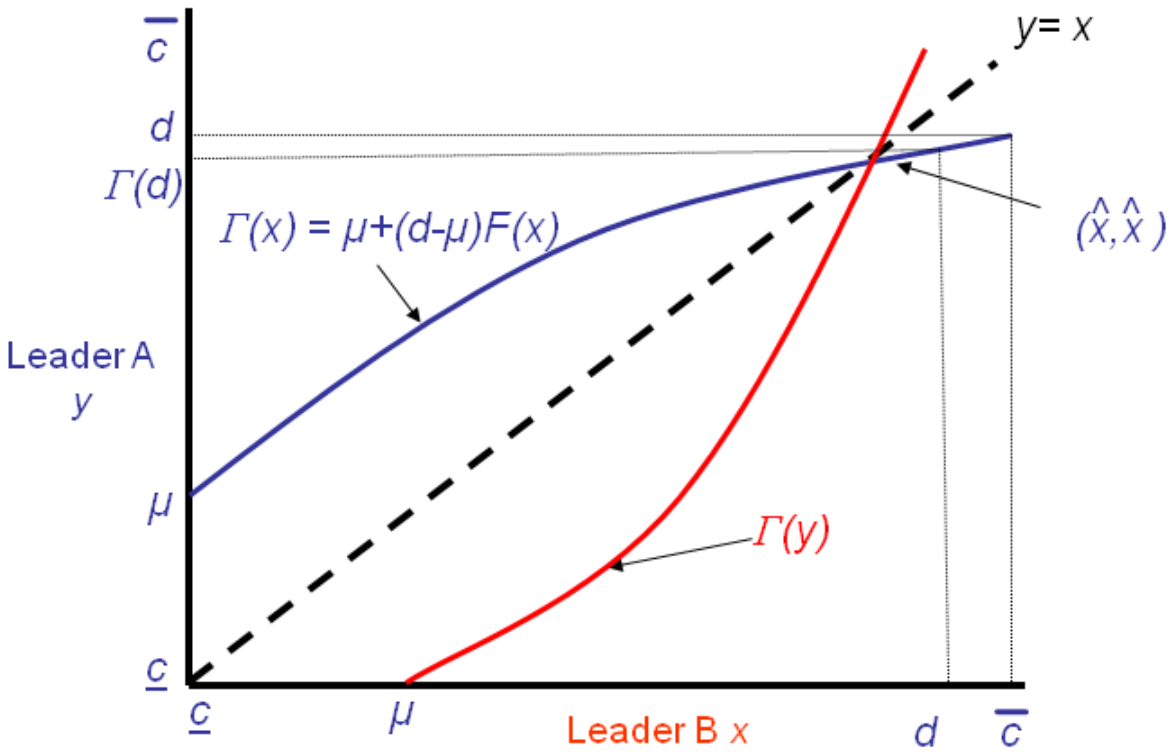

Figure 2. Strategic Complements: Lemma 2

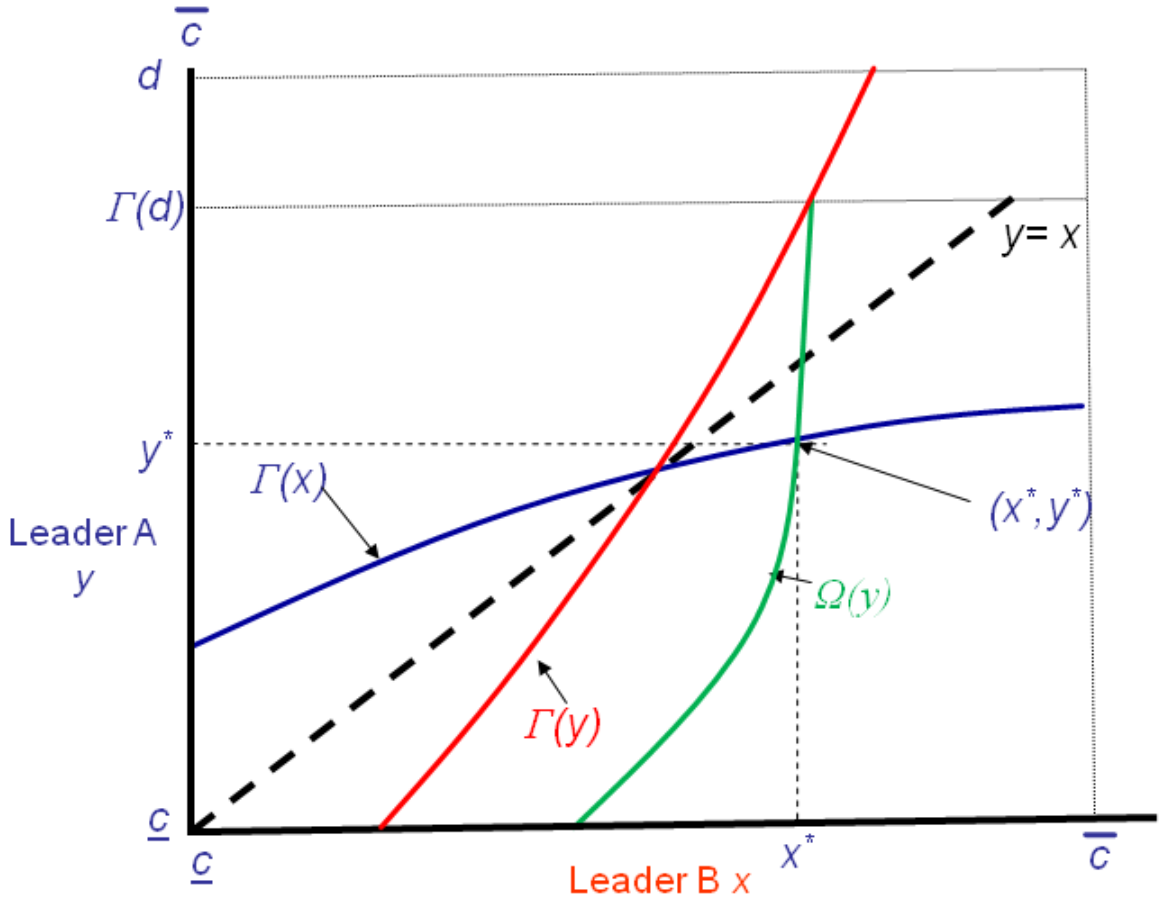

Figure 3. Strategic Substitutes: Lemma 3