Agents with extreme agendas sometimes take provocative actions that inflame conflicts. For example, Ariel Sharon’s symbolic visit to the Temple Mount in September 2000 helped spark the Second Intifada and derailed the Israeli-Palestinian peace process (Hefetz and Bloom 2006). How can extremists manipulate conflicts and when is it rational to respond aggressively to provocations?

Provocations play a key role in the conflict between the two nuclear powers India and Pakistan. After September 11, 2001, Pakistani President Musharraf sent troops to the Afghanistan border, and tried to suppress militant groups within Pakistan. In December 2001, militants sponsored by the Pakistani intelligence agency ISI attacked the Indian Parliament. India mobilized for war, and Musharraf shifted his troops from the Afghanistan border to the Indian border. Similarly, in November 2008 a terrorist attack in Mumbai raised tensions at a time when Pakistani President Zardari wanted improved relations with India. ISI-sponsored militants seem to deliberately inflame the conflict between Pakistan and India, partly because India is seen as an implacable foe, but also because the conflict relieves the pressure on extremists supported by the ISI. For Pakistani and Indian leaders, a hawkish stance may be the best response, given the (correct) belief that their opponent will become more aggressive.

The Strategy of Manipulating Conflict

By Sandeep Baliga and Tomas Sjöström

Two players choose hawkish or dovish actions in a conflict game with incomplete information. An “extremist,” who can either be a hawk or a dove, attempts to manipulate decision making. If actions are strategic complements, a hawkish extremist increases the likelihood of conflict, and reduces welfare, by sending a public message which triggers hawkish behavior from both players. If actions are strategic substitutes, a dovish extremist instead sends a public message which causes one player to become more dovish and the other more hawkish. A hawkish (dovish) extremist is unable to manipulate decision making if actions are strategic substitutes (complements). (JEL D74, D82)
Our model is based on the conflict game of Baliga and Sjöström (2004). There are two countries, A and B. In country \( i \in \{A, B\} \), a decision maker, player \( i \), chooses a dovish action \( D \) or a hawkish action \( H \). Player \( i \) may be interpreted as the median voter, a political leader, or some other pivotal decision maker in country \( i \). The hawkish action might represent accumulation of weapons, sending soldiers to a contested territory, or an act of war. Alternatively, it could represent aggressive bargaining tactics. (For example, in 2000, Ehud Barak and Yasser Arafat had to decide whether to adopt a tough stance \( H \) or a conciliatory stance \( D \) in peace negotiations.) Finally, \( H \) might represent choosing a hawkish agent who will take aggressive actions on the decision maker’s behalf. (For example, the median voters in Israel and Palestine decide whether to support Likud or Kadima, or Hamas or Fatah, respectively.)

Each decision maker can be a dominant strategy dove, a dominant strategy hawk, or a “moderate” whose best response depends on his beliefs about the opponent’s action. Neither decision maker knows the other’s true type. In Baliga and Sjöström (2004), we studied how fear of dominant strategy hawks makes moderates choose \( H \) when actions are strategic complements. Now our main purpose is to understand how a third party can manipulate the conflict. In addition, we generalize the conflict game by allowing actions to be strategic substitutes as well as complements.\(^3\) 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 communication-free equilibrium.

To study how decision makers can be manipulated by third parties, such as Sharon or the ISI, we add a third player called “the extremist” (player \( E \)). 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 his true preferences are commonly known. We consider two cases: a hawkish extremist (“provocateur”) who wants player \( A \) to choose \( H \), and a dovish extremist (“pacificist”) who wants player \( A \) to choose \( D \). Both kinds of extremists prefer that the opposing player \( B \) chooses \( D \). Political insiders, like Ariel Sharon or the ISI, 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. We simplify by assuming the extremist has \textit{perfect} information about the true preferences of the pivotal decision maker in country \( A \).

\(^3\)Baliga and Sjöström (2011) show how actions can be either strategic complements or substitutes in a bargaining game with limited commitment to costly conflict. Several empirical articles have tried to establish whether actions are strategic complements or substitutes in the Israel-Palestine conflict. Berrebi and Klor (2006, 2008) find that terrorism increases support for Israel’s right-wing Likud party, and that there is more terrorism when the left-wing Labor party is in power. Jaeger and Paserman (2008, 2009) find that Palestinian violence leads to increased Israeli violence, but Israeli violence either has no effect or possibly a deterrent effect.
To isolate the pure logic of manipulation of conflict, we assume the extremist can do nothing except communicate. Before players A and B make their decisions, player E sends a publicly observed cheap-talk message. A visit to the Temple Mount might be a real-world example. Our main interest is in communication equilibria, defined as equilibria where the extremist’s cheap-talk influences the decisions of players A and B. It may be surprising that such equilibria exist. Models of signaling and cheap-talk usually assume the sender’s preferences depend directly on his private information. In contrast, we assume it is commonly known exactly what player E wants players A and B to do. Player A knows what player E knows, but player A will pay attention to player E’s message if he thinks it might influence player B, as it will in equilibrium. We show that a communication equilibrium always exists, and find assumptions under which it is unique. Importantly, even if multiple communication equilibria exist, they always have the same structure and the same welfare implications.

In communication equilibrium, some message $m_1$ will make player B more likely to choose $H$. A provocateur is willing to send $m_1$ only if player A also becomes more likely to choose $H$. Such covarying actions must be strategic complements. On the other hand, a pacifist 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 provocateur can communicate effectively. By sending $m_1$, the provocateur triggers an unwanted (by players A and B) cascade 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 pacifist can communicate effectively. By sending $m_1$, the pacifist causes player A to back down and choose $D$.

With strategic complements, message $m_1$ can be interpreted as a provocation which increases the tension between players A and B. In equilibrium, the provocateur sends $m_1$ only when player A is a “weak moderate,” i.e., a type who would have chosen $D$ in the communication-free equilibrium, but who will choose $H$ out of fear if a provocation makes it more likely that player B chooses $H$. In response to $m_1$, player B indeed chooses $H$ with a very high probability. The absence of a provocation reveals that player A is not a weak moderate. Eliminating these types makes player B more inclined to choose $H$ than in the communication-free equilibrium, which makes player A more inclined to choose $H$ as well. Thus, with strategic complements, communication increases the probability players A and B choose $H$, whether or not a provocation actually occurs. Because each decision maker always wants the other to choose $D$, eliminating the provocateur 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 aligned with the provocateur. In view of this, one may ask why players A and B do not jointly agree to ignore the provocation and behave more peacefully. One answer may be that they do

---

4 In some situations, only costly messages (e.g., acts of violence) might be noticed above the background noise and daily concerns of media and politicians. We will show that our results are robust to messages being costly to send and receive.

5 Fromkin (1975) and others have made similar arguments about terrorism: 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
not trust each other to follow through. In Section IIB, we offer another answer: a player might appear weak if he does not react aggressively to a provocation, and appearing weak is costly.

With strategic substitutes, message $m_1$ can be interpreted as a “peace rally” in country $A$, organized by a pacifist who wants his key audience to renounce violence. For example, the Campaign for Nuclear Disarmament, formed by Bertrand Russell during the Cold War, proposed unilateral disarmament even at the cost of giving in to communism. In our model, a peace rally occurs only when player $A$ is a “tough moderate” who would have chosen $H$ in the communication-free equilibrium, but who is deterred from doing so if he fears a hawkish opponent. Following a peace rally, player $B$ indeed becomes more hawkish, and the tough moderate type of player $A$ backs down and chooses $D$. Since peace protests in country $A$ make player $B$ more hawkish, player $A$ would like to ban them if he could. On the other hand, because player $A$ becomes more dovish, the peace rally makes player $B$ better off.

We consider several extensions of the basic model. The structure of the communication equilibrium carries over to the case of provocateurs in both countries, although, surprisingly, the probability of peace may increase when the second provocateur is added. The basic results also go through with a small amount of uncertainty about whether actions are strategic substitutes or complements (in which case provocation can result in a player backing off, and a peace rally can result in mutual de-escalation), and when player $E$ may not know player $A$’s true type (in which case provocation can backfire: player $A$ might stick to $D$ while player $B$ switches to $H$).

In related work, Levy and Razin (2004) also consider cheap-talk with multiple audiences: a democratic leader sends a message to his own citizens and to another country. The citizens have the same state-contingent preferences as their leader, and the leader would prefer to send them a private message but this is assumed to be impossible. In our model, the preferences of the sender (the extremist) differ from both receivers (the decision makers), and private messages would not be useful, because the extremist seeks to indirectly influence player $A$ by publicly provoking player $B$.

In Baliga and Sjöström (2004), we show that communication between players $A$ and $B$ can be good for peace when actions are strategic complements. Although neither player wants to provoke the other to choose $H$, some types are more conflict-averse than others. This allows the construction of a “peaceful” cheap-talk equilibrium where moderate types who exchange “peaceful messages” coordinate on $D$. This construction relies on the fact that both players send messages and their preferences depend directly on their privately known types. In our current model, the provocateur’s preferences are commonly known, and his messages are bad for peace. The logic behind his manipulation of the conflict is quite different from the role played by communication in our earlier work.

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 (Fromkin 1975, p. 697).

In our model, a unilateral deviation along the lines suggested by Fromkin can never be profitable (by definition of equilibrium), but renegotiating an equilibrium at some point of the game tree might make both decision makers better off.

“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 by Rees 2002).
Jung (2007) shows how communication by a hawkish Ministry of Propaganda can refine the set of equilibria in a version of the Baliga and Sjöström (2004) model. For this purpose it is crucial that messages are not cheap-talk. In contrast, we study cheap-talk equilibria which do not replicate the outcome of any communication-free equilibrium. Edmond (2008) considers a global game where 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 (2010) studies a related model where the level of violence inflicted by uninformed extremists generates information for the population. In contrast to the global games literature, we do not assume highly correlated types (in fact, types are uncorrelated).

I. The Model

A. The Conflict Game without Cheap-Talk

The conflict game without cheap-talk is similar to the game studied in Baliga and Sjöström (2004). Two decision makers, players A and B, simultaneously choose either a hawkish (aggressive) action $H$ or a dovish (peaceful) action $D$. 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 $j \neq i$.

\[
\begin{array}{cc}
H & D \\
H & -c_i & \mu - c_i \\
D & -d & 0 \\
\end{array}
\]

We assume $d > 0$ and $\mu > 0$, so player $j$’s aggression reduces player $i$’s payoff. Notice that $d$ captures the cost of being caught out when the opponent is aggressive, while $\mu$ represents a benefit from being more aggressive than the opponent. If $d > \mu$, player $i$’s incentive to choose $H$ over $D$ increases with the probability that player $j$ chooses $H$, so the game has strategic complements. If $d < \mu$, player $i$’s incentive to choose $H$ decreases with the probability player $j$ chooses $H$ and the game has strategic substitutes.

Player $i$ has a privately known cost $c_i$ of taking the hawkish action, referred to as his “type.” Types are independently drawn from the same distribution. Let $F$ denote the continuous cumulative distribution function, with support $[\underline{c}, \bar{c}]$, and where $F'(c) > 0$ for all $c \in (\underline{c}, \bar{c})$. When taking an action, player $i$ knows $c_i$ but not $c_j, j \neq i$.

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 $D$ ($d \leq c_i \leq \mu$). 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 \( c < \mu < d < \bar{c} \). If the game has strategic substitutes then \( c < d < \mu < \bar{c} \).

Suppose player \( j \) chooses \( H \) 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 player \( i \) chooses \( H \) instead of \( D \), his net gain is

\[
\mu - c_i + (d - \mu)p_j.
\]

A strategy for player \( i \) is a function \( \sigma_i : [c, \bar{c}] \rightarrow \{H, D\} \) which specifies an action \( \sigma_i(\xi_i) \in \{H, D\} \) for each cost type \( c_i \in [c, \bar{c}] \). 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 \( x \in [c, \bar{c}] \) such that \( \sigma_i(c_i) = H \) if and only if \( c_i \leq x \). Because (2) is monotone in \( c_i \), all BNE must be in cutoff strategies. Any such strategy can be identified with its cutoff point \( x \in [c, \bar{c}] \). By Assumption 1, dominant strategy doves and hawks have positive probability, so 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 the cutoff \( x_i = \Gamma(x_j) \), where

\[
\Gamma(x) \equiv \mu + (d - \mu)F(x).
\]

The function \( \Gamma \) is the best response function for cutoff strategies. Notice that \( \Gamma'(x) = (d - \mu)F'(x) \), so the best response function is upward (downward) sloping if actions are strategic complements (substitutes). Moreover, \( \Gamma(c) = \mu > c \) and \( \Gamma(\bar{c}) = d < \bar{c} \). Since \( \Gamma \) is continuous, a fixed-point \( \hat{x} \in (c, \bar{c}) \) exists. Thus, a BNE exists (where by the symmetry of the game each player uses cutoff \( \hat{x} \)).

Assumption 2 states that the density of \( F \) is not too large anywhere, i.e., that there is significant uncertainty about types.\(^7\)

ASSUMPTION 2: \( F'(c) < \frac{1}{d - \mu} \) for all \( c \in (c, \bar{c}) \).

Assumption 2 implies that \( 0 < \Gamma'(x) < 1 \) if \( d > \mu \) and \( -1 < \Gamma'(x) < 0 \) if \( d < \mu \). Hence, in both cases a well-known sufficient condition for uniqueness is satisfied: the best response functions have slope strictly less than one in absolute value (see Vives 2001). Thus, we have:

THEOREM 1: The conflict game without cheap-talk has a unique Bayesian Nash equilibrium.

\(^7\)As long as Assumption 1 is satisfied, the uniform distribution on \( [c, \bar{c}] \) satisfies Assumption 2. But Assumption 2 is much weaker than uniformity. What it rules out is having probability mass highly concentrated around one particular type. This guarantees that the BNE is unique. See Morris and Shin (2005) for a detailed discussion of uniqueness in this type of game.
Theorem 1 says that without cheap-talk there is a unique BNE, which we refer to as the communication-free equilibrium, whether actions are strategic substitutes or complements. In equilibrium, player $i$ chooses $H$ if $c_i \leq \hat{x}$, where $\hat{x}$ is the unique fixed point of $\Gamma(x)$ in $[c, \bar{c}]$. See Figure 1 for the case of strategic complements (the equilibrium is the intersection of the best response curves $x_B = \Gamma(x_A)$ and $x_A = \Gamma(x_B)$).

The unique communication-free equilibrium can be reached via iterated deletion of dominated strategies. With strategic complements, the fear of dominant strategy hawks causes coordination types who are “almost dominant strategy hawks” (i.e., types slightly above $\mu$) to play $H$, which in turn causes “almost-almost dominant strategy hawks” to play $H$, etc. The “hawkish cascade” causes higher and higher types to choose $H$. Meanwhile, since dominant strategy doves play $D$, “almost dominant strategy doves” (i.e., types slightly below $d$) also play $D$, knowing that the opponent may be a dominant strategy dove. The “dovish cascade” causes lower and lower types to choose $D$. With sufficient uncertainty about types, these two cascades completely resolve the ambiguity about what coordination types will do.\footnote{It is obvious from Figure 1 that the equilibrium is also unique if Assumption 2 is replaced by the assumption that $F$ is concave. Sometimes concavity of $F$ is convenient to work with (c.f. Section IIB) but it is hard to justify intuitively. In contrast, Assumption 2 formalizes the intuitive notion of sufficient uncertainty about types.}

\footnote{Strategic substitutes generate a different kind of spiral. Fearing dominant strategy hawks, “almost dominant strategy doves” back down and play $D$. This emboldens “almost dominant strategy hawks” to play $H$, and so on.}
B. Cheap-Talk

We now introduce a third player, player $E$, the extremist. 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 $c_B$.

We consider two possibilities. First, if player $E$ is a hawkish extremist (a “provocateur”), then $c_E < 0$. Thus, the provocateur enjoys a benefit $(−c_E) > 0$ if player $A$ is aggressive. The provocateur is guaranteed a strictly positive payoff if player $A$ chooses $H$, but he gets at most zero when player $A$ chooses $D$, so he certainly wants player $A$ to choose $H$. Second, if player $E$ is a dovish extremist (a “pacifist”), then $c_E > μ + d$. The most the pacifist can get if player $A$ chooses $H$ is $μ − c_E$, while the worst he can get when player $A$ chooses $D$ is $−d > μ − c_E$, so he certainly 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 IA, player $E$ sends a publicly observed cheap-talk message $m ∈ M$, where $M$ is his message space. The time line is as follows:

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

(ii) Player $E$ sends a (publicly observed) cheap-talk message $m ∈ M$.

(iii) Players $A$ and $B$ simultaneously choose $H$ or $D$.

In a “babbling” equilibrium, messages are disregarded and at time (iii) players $A$ and $B$ behave just as in the unique communication-free equilibrium of Section IA. Cheap-talk is effective if there is a positive measure of types that choose different actions at time iii than they would have done in the communication-free equilibrium. For cheap-talk to be effective, player $E$’s message must reveal some information about player $A$’s type. A Perfect Bayesian Equilibrium (PBE) with effective cheap-talk is a communication equilibrium. We will show that communication equilibria have a very specific structure, allowing us to unambiguously compare communication equilibrium payoffs with the payoffs in the babbling (communication-free) equilibrium.

A strategy for player $E$ is a function $m : [c, \bar{c}] → M$, where $m(c_A)$ is the message sent by player $E$ when player $A$’s type is $c_A$. Without loss of generality, each player $j ∈ \{A, B\}$ uses a “conditional” cutoff strategy: for any message $m ∈ M$, there is a cutoff $c_j(m)$ such that if player $j$ hears message $m$, he chooses $H$ if and only if $c_j ≤ c_j(m)$. The next lemma shows that any communication equilibrium can be taken to involve just two messages, say $m_0$ and $m_1$. One message, say $m_1$, must make player $B$ behave more hawkishly than the other message, $m_0$. 
LEMMA 1: In communication equilibrium, it is without loss of generality to assume that \( M \) contains only two messages, \( M = \{m_0, m_1\} \). The probability that player B plays \( H \) is higher after \( m_1 \) than after \( m_0 \). That is, \( c_B(m_1) > c_B(m_0) \).

All omitted proofs are in the Appendix. Lemma 1 applies for both strategic substitutes and strategic complements, and for both pacifists and provocateurs. The proof of the lemma does not use Assumption 2.

II. Cheap-Talk with Strategic Complements

In this section, we consider the case of strategic complements, \( d > \mu > 0 \).

A. Main Results

From Lemma 1, we can assume only two messages, \( m_0 \) and \( m_1 \), are sent in equilibrium. Player B is more likely to choose \( H \) after \( m_1 \) than after \( m_0 \). If player A’s action does not depend on the message, then the extremist certainly prefers to send \( m_0 \). If player A’s action depends on the message, then player A must be a coordination type who (by strategic complements) plays \( H \) in response to \( m_1 \) and \( D \) in response to \( m_0 \).

If player E is a pacifist, then he wants both players A and B to choose \( D \), so he must always send \( m_0 \) in equilibrium. But a constant message is not informative, and the outcome must be equivalent to the unique communication-free equilibrium of Section I A. Thus, we have the following result.

THEOREM 2: If player E is a pacifist and the game has strategic complements, then cheap-talk cannot be effective.

Now suppose player E is a provocateur. We will show there exists a communication equilibrium where the provocateur uses cheap-talk to increase the risk of conflict above the level of the communication-free equilibrium. The communication equilibrium has the following structure. If \( c_A \) is either very high or very low, then player A’s action will not depend on the message, and sending \( m_0 \) is optimal as it reduces the probability that player B will choose \( H \). The provocateur can only benefit from message \( m_1 \) if it causes player A to switch from \( D \) to \( H \). Thus, the provocateur’s strategy must be non-monotonic: he sends message \( m_1 \) if and only if player A belongs to an intermediate range of coordination types who play \( D \) following \( m_0 \) but \( H \) following \( m_1 \).

By this logic, if message \( m_1 \) is sent then player B knows that player A will play \( H \). Therefore, player B plays \( H \) unless he is a dominant strategy dove. That is, his optimal cutoff point is \( c_B(m_1) = d \), and the probability that he plays \( H \) is \( F(d) \). Accordingly, player A’s best response is to choose \( H \) if and only if \( c_A \leq \Gamma(d) \), where \( \Gamma \) is defined by equation (3). That is, \( c_A(m_1) = \Gamma(d) \). Thus, conditional on message \( m_1 \), players A and B must use cutoffs \( c_A(m_1) = \Gamma(d) \) and \( c_B(m_1) = d \), respectively.

Since player B is less likely to play \( H \) after \( m_0 \) than after \( m_1 \), by strategic complements, so is player A. Thus, \( c_A(m_0) < c_A(m_1) = \Gamma(d) \). If player A is of type \( c_A \leq c_A(m_0) \) then he plays \( H \) following any message; if his type is \( c_A > c_A(m_1) = \Gamma(d) \) then he plays \( D \) following any message. But if \( c_A \in (c_A(m_0), \Gamma(d)) \), then
player A chooses D after $m_0$ and H after $m_1$. As the provocateur wants player A to be hawkish, he sends $m_1$ if and only if $c_A \in (c_A(m_0), \Gamma(d)]$.

It remains to determine the cutoffs used by players A and B conditional on message $m_0$, denoted $y^* = c_A(m_0)$ and $x^* = c_B(m_0)$. These cutoffs, and the associated strategy profiles, are indicated in Figure 2. As always, optimal cutoffs are determined by the probability that the opponent plays H. Player B uses cutoff $x^*$ after $m_0$ so he plays H with probability $F(x^*)$. Therefore, player A’s optimal cutoff is $y^* = \Gamma(x^*)$, where $\Gamma$ is defined by equation (3). Now, the message $m_0$ is sent when $c_A$ is either below $y^*$ or above $\Gamma(d)$, and player A chooses H in the former case and D in the latter case. Therefore, conditional on $m_0$, player A chooses H with probability

$$F(y^*) \over 1 - F(\Gamma(d)) + F(y^*).$$

Player B’s optimal cutoff $x^*$ is the best response to the belief that player A chooses H with probability given by (4). Since $y^* = \Gamma(x^*)$, to prove existence of communication equilibrium we use a fixed-point argument to show that $x^*$ and $y^*$ exist. This is given in the proof of part (i) of Theorem 3 (in the Appendix).

**THEOREM 3**: Suppose player E is a provocateur and the game has strategic complements. (i) A communication equilibrium exists. (ii) 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 $\hat{x} < c_A \leq \Gamma(d)$ (where $\hat{x}$ is the cutoff in the communication-free equilibrium). (iii) If

$$F(y) \left\{ \frac{1}{1 - F(\Gamma(d))} + \frac{1}{F(y)} \right\} < \frac{1}{d - \mu}$$

for all $y \in (c, \bar{c})$, then the communication equilibrium is unique.

Proving part (ii) of Theorem 3 involves showing that players $A$ and $B$ behave more hawkishly than in the communication-free equilibrium, no matter which message is sent. Intuitively, we interpret $m_1$ as a “provocation” which occurs when player $A$ is a “weak” coordination type $c_A \in (y^*, \Gamma(d)]$. Following a provocation, player $B$ chooses $H$ (except if he is a dominant strategy dove) and this causes player $A$ to toughen up and play $H$. It is as if the provocation makes players $A$ and $B$ coordinate on a “bad” equilibrium of a stag-hunt game: they behave aggressively because they believe (correctly) that the other will be aggressive.

The cutoffs conditional on $m_0$ are lower than the cutoffs conditional on $m_1$, so the decision makers behave less aggressively following $m_0$ than following $m_1$, which justifies interpreting $m_0$ as the absence of a provocation. This absence is informative, just as Sherlock Holmes, in the story *Silver Blaze*, found it informative that a dog did not bark (Conan Doyle 1894). Specifically, message $m_0$ reveals that player $A$ is not a weak coordination type $\left( c_A \notin (y^*, \Gamma(d)) \right)$. The weak coordination types would have chosen $D$ in communication-free equilibrium, so eliminating these types is bad for peace. Thus, message $m_0$ actually triggers more aggression than the communication-free equilibrium (although not as much as $m_1$ does). Formally, the proof of Theorem 3 shows that the cutoffs after message $m_0$ are higher than the cutoffs in the communication-free equilibrium: $x^* > \hat{x}$ and $y^* > \hat{y}$.

It follows from these arguments that if a type would have chosen $H$ in the communication-free equilibrium, then 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. Since all types of players $A$ and $B$ want their opponent to choose $D$, they are all harmed by the third party’s cheap-talk.

For the provocateur, the benefits of cheap-talk are ambiguous. 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. On the other hand, if $\hat{x} < c_A \leq \Gamma(d)$, then player $A$ would have chosen $D$ in the communication-free equilibrium, but in the communication equilibrium he plays $H$, making player $E$ better off.

Part (iii) of Theorem 3 shows that the communication equilibrium is unique if a “conditional” version of Assumption 2 holds.\(^{10}\) 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

\(^{10}\) That is, except for trivial relabeling of messages, there is only one PBE with effective cheap-talk. Of course, the “babbling” PBE always exists as well.
equilibrium must be the equilibrium of a “conditional” game where player A’s type distribution $G$ has support $[c, y^*] \cup (\Gamma(d), \bar{c}]$ and density

$$G'(c) = \frac{F'(c)}{1 - F(\Gamma(d)) + F(y^*)}.$$  

Furthermore, following $m_0$, player A’s cutoff type $y^* = c_A(m_0)$ is indifferent between $H$ and $D$. Therefore, in the “conditional” game, the only possible cutoff type is $y^*$. Theorem 1 showed that equilibrium in the communication-free game is unique if Assumption 2 holds, i.e., if the distribution is sufficiently diffuse. The analogous “conditional” diffuseness condition for communication equilibrium turns out to be $G'(y^*) < 1/(d - \mu)$ for all $y^*$.\(^{11}\) Note that even if this condition is violated, the only possible non-uniqueness comes from the possibility of multiple fixed points $(x^*, y^*)$, but the structure of the communication equilibrium is always the same (i.e., provocations occurring for weak coordination types, welfare effects given by part (ii) of Theorem 3, etc.).

The current model assumes a third party extremist communicates while players $A$ and $B$ are silent. In Baliga and Sjöström (2004), we found that (in the absence of an extremist) the two decision makers could reduce conflict by sending their own messages. These messages separated out “tough” coordination types who would have played $H$ in the communication-free equilibrium, which cut the “hawkish cascade” and allowed the intermediate types to coexist peacefully. In the current model, a provocation separates out “weak” coordination types, who would have played $D$ in the communication-free equilibrium but now switch to $H$. This brings conflict when peace could have prevailed. Even when no provocation occurs, the situation is still worse than the communication-free equilibrium, because the absence of weak coordination types leads to a less favorable type-distribution (the “dovish cascade” is cut off).

**B. Extensions**

*Provocateurs in Both Countries.*—Extremists may not be confined to just one country. Suppose each country $i \in \{A, B\}$ has its own provocateur, player $E_i$, who knows the type of player $i$ (the decision maker in country $i$). The two provocateurs simultaneously send (publicly observed) messages. We obtain the following symmetric version of the communication equilibrium of Section IIA. There are two cutoffs $\bar{x}$ and $\bar{y}$, with $\mu < \bar{x} < \bar{y} < d$. In each country $i \in \{A, B\}$, player $E_i$ sends $m_1$ (a “provocation”) if $c_i \in (\bar{x}, \bar{y}]$, and $m_0$ otherwise. Player $i \in \{A, B\}$ behaves as follows. If player $E_j$ sends $m_1$, where $i \neq j$, then player $i$ chooses $H$ if and only if $c_i \leq d$. If player $E_i$ sends $m_1$, where $i \neq j$, then player $i$ chooses $H$ if and only if $c_i \leq \bar{y}$. Finally, if both extremists send $m_0$, then player $i$ chooses $H$ if and only if $c_i \leq \bar{x}$. The existence proof (in the Appendix) uses a fixed-point argument to find $\bar{x}$ and $\bar{y}$.

\(^{11}\)For example, suppose $F$ is uniform on $[0, \bar{c}]$. Then inequality (5) holds if $\bar{c}$ is big enough, more precisely if $(\bar{c} - d)\bar{c} > (d - \mu)d$.  

THEOREM 4: With a provocateur in each country and strategic complements, a symmetric communication equilibrium exists.

The logic of this equilibrium is just as in Section IIA. Extreme cost-types with $c_i \leq \tilde{x}$ or $c_i > \tilde{y}$ are not responsive to provocation so player $E^j$ sends $m_0$ to minimize the probability that player $j$ chooses $H$. If instead $c_i \in (\tilde{x}, \tilde{y})$, the message sent by player $E^i$ is pivotal if and only if player $E^j$ sends $m_0$. Then, if player $E^i$ sends $m_1$ instead of $m_0$ he changes player $i$'s action from $D$ to $H$, which he prefers. Therefore, each extremist is provocative only in the intermediate range. When player $E^j$ sends $m_1$, player $i \neq j$ knows player $j$ will play $H$, so player $i$ chooses $H$ unless he is a dominant strategy dove. When player $E^j$ sends $m_0$, player $i$'s incentive to choose $H$ depends on the message sent by player $E^j$. Player $j$ is more hawkish when player $E^j$ sends $m_1$ rather than $m_0$ and hence by strategic complementarities so is player $i$ (that is, $\tilde{x} < \tilde{y}$).

It might seem as if two provocateurs will create more conflict than one, but this is not necessarily the case. If no information is revealed about player $j$, then player $i$'s type $\Gamma(d)$ is the highest type that could conceivably be convinced to play $H$ (because player $j$'s types above $d$ play $D$ for sure). In the communication equilibrium of Section IIA, the provocateur in country $A$ actually achieves this upper bound: for $c_A = \Gamma(d)$, as well as for lower types, a provocation occurs which causes player $A$ to choose $H$ and player $B$ to choose $H$ with probability $F(d)$. Thus, a single provocateur has a remarkable ability to provoke aggression. But now player $E^B$ reveals information about player $B$. If player $E^B$ sends $m_0$, then player $A$ knows that $c_B \notin (\tilde{x}, \tilde{y})$. Since the removed types in $(\tilde{x}, \tilde{y})$ are not dominant strategy doves, player $A$ knows that the probability that player $B$ will choose $H$ must be strictly less than $F(d)$, so if $c_A = \Gamma(d)$ then player $A$ strictly prefers $D$, and the same is true for types slightly below $\Gamma(d)$. Thus, the information revealed about player $B$ actually makes it harder to convince player $A$ to choose $H$. Formally, in the communication equilibrium of Section IIA the peaceful outcome $DD$ occurred when $c_A > \Gamma(d)$ and $c_B > x^*$. Here, with two extremists, the outcome $DD$ occurs when $c_A > \tilde{y}$ and $c_B > \tilde{y}$. It can be shown that $x^* < \tilde{y} < \Gamma(d)$, so it is not possible to say if the peaceful outcome is more or less likely.

In what follows, we assume there is an extremist only in country $A$.

Costly Messages.—What happens if it is costly for the provocateur to ensure that his message is heard? In our model, the provocateur is willing to incur a cost to manipulate the conflict game, so such costs do not change the nature of our arguments. Suppose the “provocative” message $m_1$ imposes a cost $\tau_j > 0$ on player $j \in \{A, B, E\}$. The other message, $m_0$, involves no costs. The extremist does not internalize $\tau_A$ and $\tau_B$, and as these costs are already incurred when players $A$ and $B$ move, they do not affect strategic behavior. We now argue that if $\tau_E$ is not prohibitively big, then the communication equilibrium exists as before. Player $E$’s expected payoff from $m_1$ when $c_A \in (\tilde{y}, \Gamma(d)]$ is $-c_E + (1 - F(d)) \mu - \tau_E$, as player $A$ plays $H$ and player $B$ plays $H$ unless he is a dominant strategy dove. If player $E$ instead chooses $m_0$, then player $A$ plays $D$ and player $E$’s expected payoff is $-dF(x^*)$. Player $E$ prefers $m_1$ if

$$dF(x^*) - c_E + (1 - F(d)) \mu > \tau_E.$$
The left hand side is strictly positive, so if \( \tau_E \) is not too big, the communication equilibrium of Section IIA still exists. In what follows, we return to the case of pure cheap-talk.

**Credibility, Renegotiation, and Domestic Politics.**—The provocateur’s messages create conflict, which is bad for players \( A \) and \( B \). Given that the messages are publicly observed, the two decision makers cannot simply agree to disregard the messages and behave as in the communication-free equilibrium, because the messages convey information about player \( A \)’s type. Neither can they convince the provocateur to voluntarily refrain from provoking conflict, because he benefits from it. The question is whether, conditional on the information revealed by the extremist’s message, players \( A \) and \( B \) can “renegotiate” their strategies. In the communication equilibrium, message \( m_1 \) triggers a hawkish continuation equilibrium. But since the message in fact reveals that player \( A \) is a weak coordination type, there also exists a dovish continuation equilibrium, where player \( A \) chooses \( D \) and player \( B \) chooses \( D \) unless he is a dominant strategy hawk. However, renegotiation would face several problems. The first is information leakage: if renegotiation is not anticipated, but player \( B \) wants to renegotiate, player \( A \) might fear that player \( B \) is a dominant strategy hawk out to trick him. Second, even if there is no information leakage, there is a credibility problem. Each player, regardless of type, has an incentive to try to convince the opponent to become more dovish, even if he doubts that this will work so that he himself plans to stick to the original hawkish equilibrium. Therefore, an appeal to renegotiate and behave more peacefully is not informative of the player’s own intentions, and may therefore not convince the opponent to deviate from the original equilibrium (c.f. Aumann 1990).

A third problem is that a leader who does not react hawkishly to a provocation may look weak, and 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, suppose player \( B \) gets an extra payoff \( R > 0 \) if he plays \( H \) after \( m_1 \), interpreted as rents from increased popularity. Assume for convenience \( \bar{c} > R + d \) to rule out corner solutions. The communication equilibrium of Section IIA is modified as follows to take \( R \) into account. Player \( A \)’s cutoff points are \( c_A(m_0) = y^{**} \) and \( c_A(m_1) = \Gamma(R + d) \). Player \( B \)'s cutoff points are \( c_B(m_0) = x^{**} \) and \( c_B(m_1) = R + d \). Player \( E \) sets \( m(c_A) = m_1 \) if and only if \( c_A \in (y^{**}, \Gamma(R + d)) \). As before, a fixed-point argument is used to find \( x^{**} \) and \( y^{**} \). But now messages are not cheap-talk, and we can obtain a stronger result than before. Specifically, if \( R + \mu > d, F \) is concave, and a condition analogous to (5) holds, namely

\[
\frac{F'(y)}{1 - F(\Gamma(R + d)) + F(y)} < \frac{1}{d - \mu},
\]

then the unique (modified) communication equilibrium is renegotiation-proof in the following strong sense: following any message there is a unique continuation equilibrium. Thus, even abstracting from the information leakage and credibility problems, there is no self-enforcing agreement where players \( A \) and \( B \) behave more dovishly following \( m_1 \). Intuitively, player \( B \) is sufficiently aggressive following \( m_1 \).
that the iterated deletion of dominated strategies (the hawkish cascade) generates a unique continuation equilibrium.

Moreover, there can be no “babbling” PBE. To see this, notice that if $c_B \leq R + \mu$, then following $m_1$, $H$ dominates $D$ for player $B$. Thus, in any PBE, $c_B(m_1) \geq R + \mu$. If $c_B \geq d$, then following $m_0$, $D$ dominates $H$ for player $B$. Thus, in any PBE, $c_B(m_0) \leq d$. If $R + \mu > d$ then $c_B(m_1) > c_B(m_0)$, and $c_A(m_1) > c_A(m_0)$ by strategic complements. The provocateur therefore prefers to send $m_1$ if $c_A(m_0) < c_A \leq c_A(m_1)$ (since this makes player $A$ choose $H$) but $m_0$ otherwise (since this minimizes the probability that player $B$ chooses $H$). Thus, a provocation necessarily occurs if and only if player $A$ is an intermediate type.

**Theorem 5:** If $R > d - \mu$, $F$ is concave and inequality (6) holds for all $y \in (\kappa, \tilde{c})$, then the (modified) communication equilibrium is the unique PBE, and it is renegotiation-proof.

*Partially Uninformed Cheap-Talk.*—If $c_A$ is either very high or very low, then the fact that the provocateur knows $c_A$ makes him worse off because of the “dog that did not bark” effect (part (ii) of Theorem 3). He cannot escape this logic by staying silent, because it will simply be equated with sending $m_0$ (and hence informative).

However, suppose the provocateur is known to be informed only with probability $p$, where $0 < p < 1$. His “silence” is less informative and players $A$ and $B$ are more peaceful. But this means there will be more scope for provocation to create conflict for intermediate $c_A$.

First, we informally discuss the provocateur’s incentive to be provocative when he does not know $c_A$. That is, he does not know how player $A$ will react to his message. If each player $i \in \{A, B\}$ plays $H$ with probability $p_i$, then player $E$’s expected payoff is

$$p_A[-c_E + (1 - p_B)\mu] - (1 - p_A)p_B d.$$

Suppose a provocation increases each decision maker $i$’s probability of playing $H$ from $p_i$ to $p'_i = p_i + \delta_i > p_i$. After some manipulations, the change in player $E$’s expected payoff can be expressed as the following weighted sum of $\delta_A$ and $\delta_B$:

$$(-c_E + (1 - p'_B)\mu + p'_B d)\delta_A - (p_A\mu + (1 - p_A)d)\delta_B.$$

This expression confirms that the increase in $p_A$ makes the provocateur better off (the first term is positive), but the increase in $p_B$ makes him worse off (the second term is negative). Depending on the relative sizes of $\delta_A$ and $\delta_B$, either term might dominate, so in general, we cannot say whether provocations would pay for the uninformed extremist. However, the weight on $\delta_A$ is bigger, the bigger is $p'_B$ (as $d > \mu$). Intuitively, if tensions are high, so player $B$ is likely to choose $H$, increasing $p_A$ is very valuable to the provocateur, because he reduces the chance of incurring the cost $d$. On the other hand, the weight on $\delta_B$ is smaller (in absolute value) the bigger is $p_A$. Intuitively, if player $A$ is likely to choose $H$, increasing $p_B$ is not so costly to the provocateur, because he is unlikely to incur the cost $d$. Thus, provocations are
more likely to benefit an uninformed extremist in situations where tensions are high
and hawkish behavior not unlikely. In contrast, a provocation where tension is low
may backfire by causing the outcome DH. Suppose, in fact, the uninformed provo-
cateur prefers to send \( m_0 \) to reduce the risk of the outcome DH.

The informed provocateur will, following the logic of Section IIA (where in effect
\( p = 1 \)), send \( m_1 \) to provoke conflict when \( c_A \) is in some intermediate range. But the
“dog that did not bark” effect is diluted since message \( m_0 \) may come from some-
one who has no information about \( c_A \). Therefore, player B is more likely to play \( D \)
after message \( m_0 \) if \( p < 1 \) than if \( p = 1 \). By strategic comelments, so is player A.
This causes the informed provocateur to send \( m_1 \) even when \( c_A \) is fairly low, to pre-
vent player A from choosing \( D \). Because the absence of a provocation may simply
mean that the extremist is uninformed, there is less conflict in this case, and so the
informed extremist resorts to provocations more frequently to prevent peace.

III. Cheap-Talk with Strategic Substitutes

In this section, we consider the case of strategic substitutes, 0 < \( d < \mu \). Lemma
1 still applies, but now the message \( m_1 \) which makes player B more likely to play \( H \)
must make player A more likely to play \( D \). Since \( \mu > 0 \) and \( d > 0 \), player E always
prefers player B to play \( D \). Also, a hawkish extremist (provocateur) wants player A
to choose \( H \), so he clearly would always send \( m_0 \). This gives us the following result.

**THEOREM 6:** If player E is a provocateur and the game has strategic substitutes,
then cheap-talk cannot be effective.

If player E is a pacifist, however, a communication equilibrium exists. Since \( m_1 \)
makes player B more hawkish (\( c_B(m_1) > c_B(m_0) \)), by strategic substitutes it makes
player A more dovish (\( c_A(m_1) < c_A(m_0) \)). The pacifist will send \( m_1 \) if and only if
player A is an opportunistic type who is induced by \( m_1 \) to switch from \( H \) to \( D \) (i.e.,
when \( c_A(m_1) < c_A \leq c_A(m_0) \)). Intuitively, we can interpret message \( m_1 \) as a “peace
rally” which signals that player A will back down and choose \( D \) for sure. This causes
player B to choose \( H \), unless he is a dominant strategy dove (\( c_B(m_1) = \mu \)). Player
A’s optimal cutoff point is \( c_A(m_1) = \Gamma(\mu) \). The cutoff points following \( m_0 \), denoted
\( y^* = c_A(m_0) \) and \( x^* = c_B(m_0) \), are constructed in the Appendix. The same argument
as in Section IIA implies that for uniqueness, we must impose a “conditional” ver-
sion of Assumption 2, specifically,

\[
\frac{F'(y)}{1 - F(y) + F(\Gamma(\mu))} < \frac{1}{\mu - d}.
\]

\[\text{\[7\]}\]

\[\text{\[12\]}\]

If we had assumed \( 0 > \mu > d \), then player E would prefer that player B plays \( H \) in the strategic substitutes
case. In this case, a relabeling of player B’s strategies, \( H \rightarrow d \) and \( D \rightarrow h \), would restore strategic complementarity;
again, only hawkish extremists would be able to communicate effectively. We in fact assume, however, that the
provocateur always wants player A to choose \( H \) and player B to choose \( D \), while the pacifist always wants both to
choose \( D \). Maintaining \( \mu > 0 \) and \( d > 0 \), the strategic substitutes and complements cases are not isomorphic; a
relabeling of strategies cannot turn one case into the other.
THEOREM 7: Suppose player E is a pacifist and the game has strategic substitutes. (i) A communication equilibrium exists. (ii) All of player A’s types prefer the communication-free equilibrium to any 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 \( \Gamma(\mu) < c_A \leq \hat{x} \) (where \( \hat{x} \) is the unique fixed point of \( \Gamma(x) \) in \([c, \overline{c}]\)). (iii) If condition (7) holds for all \( y \in (\overline{c}, \hat{x}) \) then the communication equilibrium is unique.

The communication equilibrium has a “better red than dead” flavor, in the sense that the pacifist sends \( m_1 \) to make player A back down, even at the cost of making player B more hawkish. Evidently, player B benefits from message \( m_1 \). In fact, player B benefits from message \( m_0 \) as well, as it eliminates types of player A who would have played \( H \) in communication-free equilibrium. This makes player B more likely to choose \( H \), and hence player A more likely to choose \( D \), than in the communication-free equilibrium. In summary, whichever message is sent, player B is more hawkish and player A more dovish—hence player A worse off—than in communication-free equilibrium. This makes player B would have played \( m_1 \) in communication-free equilibrium. In fact, player B benefits from message \( m_0 \) as well, as it eliminates types of player A who would have played \( H \) in communication-free equilibrium. This makes player B more likely to choose \( H \), and hence player A more likely to choose \( D \), than in the communication-free equilibrium. (Formally, player B’s cutoffs \( x' \) and \( \mu \) are both strictly greater than \( \hat{x} \), while player A’s cutoffs \( y' \) and \( \Gamma(\mu) \) are both strictly smaller than \( \hat{x} \).) It is not possible to unambiguously say if the pacifist is good for peace, since he makes one player more dovish but the other more hawkish.

An interesting generalization is that the slope of a best response function may be uncertain. We will argue that the communication equilibria of Theorems 3 and 7 are robust to a small amount of uncertainty of this kind, but they fail to exist if there is too much uncertainty. Specifically, suppose the parameter \( \mu \) in the payoff matrix (1) is \( \mu_A \) for player A and \( \mu_B \) for player B. Player i’s best response function is \( \Gamma_i(x) \equiv \mu_i + (d - \mu_i)F(x) \). For simplicity, \( \mu_A \) is fixed, but \( \mu_B \) can take two values, \( \mu_B \in \{\mu, \mu'\} \), where \( \mu < d < \mu' \). The probability that \( \mu_B = \mu' \) is \( \eta \), where \( 0 < \eta < 1 \). Only player B knows the true \( \mu_B \). Notice that with probability \( \eta \), player B’s best response function slopes down (\( \Gamma'_B(x) < 0 \)), as with strategic substitutes, but with probability \( 1 - \eta \) it slopes up (\( \Gamma'_B(x) > 0 \)), as with strategic complements.

Suppose that following message \( m \), player j chooses \( H \) with probability \( p_j(m) \). From (2), player B’s optimal cutoff following \( m \) is \( \mu_B + (d - \mu_B)p_A(m) \). Thus,

\[
(8) \quad p_B(m) = (1 - \eta)F(\mu + (d - \mu)p_A(m)) + \eta F(\mu' + (d - \mu')p_A(m)).
\]

Suppose \( m_0 \) minimizes \( p_B(m) \). If player E is a provocateur and \( \mu_A > d \) (so player A’s best response function slopes down), or if player E is a pacifist and \( \mu_A < d \) (so player A’s best response function slopes up), then player E would always send \( m_0 \), so communication is ineffective in these two cases (mimicking our earlier results).

Now suppose player E is a provocateur and \( \mu_A < d \), so player A’s best response function slopes up. If \( \eta > 0 \) is small enough, there exists a communication equilibrium similar to the one described in Theorem 3. Player E will send \( m_0 \) if player A’s action is not responsive to the message, but he will send \( m_1 \) if it changes player A’s action from \( D \) to \( H \). Therefore, in equilibrium, following \( m_1 \) player A must choose \( H \) for sure: \( p_A(m_1) = 1 \). From (8), we get \( p_B(m_1) = F(d) \). In contrast, \( p_A(m_0) < 1 \). If \( \eta \) is small, then \( p_B(m_0) < F(d) = p_B(m_1) \),
because $d > \mu$. Therefore, since he considers actions to be strategic complements, there will indeed be a set of types of player $A$ who want to play $D$ following $m_0$ but $H$ following $m_1$. This allows the equilibrium construction to go through as in the proof of Theorem 3. Thus, the communication equilibrium is robust to a small amount of uncertainty about whether player $B$’s best response function has positive or negative slope. Indeed, if there is a (small) chance that a provocation causes “the enemy” (player $B$) to back down, this actually strengthens the extremist’s incentive to be provocative. However, if $\eta$ is sufficiently big then equation (8) implies $p_B(m_0) > F(d) = p_B(m_1)$. In this case, if the provocation causes player $A$ to become more hawkish, then the probability that player $B$ becomes more dovish is so large that player $A$ would also want to be more dovish (since his best response function slopes up), a contradiction. Thus, when $\eta$ is too big, the communication equilibrium construction fails. Intuitively, the provocateur cannot create a hawkish cascade if player $B$ is very likely to react to aggression by backing down.

A similar reasoning reveals that if player $E$ is a pacifist and $\mu_A > d$, a communication equilibrium similar to the one described in Theorem 7 exists if $\eta$ is sufficiently big, so it is likely that players $A$ and $B$ agree that actions are strategic substitutes. With $\eta < 1$ there is even a chance that a peace rally will make player $B$ more peaceful, which strengthens the pacifist’s incentive to stage the rally: it might bring about the outcome $DD$. However, if $\eta$ is too small, then if the peace rally causes player $A$ to become more dovish, the probability that player $B$ also becomes more dovish is so large that player $A$ would actually want to be more hawkish (as his best response function slopes down), a contradiction. Thus, in this case the communication equilibrium construction fails when $\eta$ is too small.

IV. Conclusion

The International Relations literature distinguishes fear-spirals, like the one preceding World War I, from conflicts like World War II where lack of deterrence emboldened Hitler (Nye 2007, p. 111). Games with strategic complements or substitutes are stylized representations of these two kinds of strategic interactions. We have studied how a hawkish extremist can trigger conflicts when actions are strategic complements. When actions are strategic substitutes, the hawkish extremist is powerless, but a dovish extremist can convince one side to back down.

Provocateurs gain extra power if, unlike in our model, their actions cannot be clearly distinguished from those of the country’s highest leaders. For example, Ellsberg (2002) describes how elements within the US government wanted to provoke North Vietnam. On January 28, 1965, US naval patrols “with the mission of provoking an attack, were ordered back into the Tonkin Gulf” (Ellsberg 2002, p. 66). The mission succeeded, and paved the way for heavy American involvement in Vietnam. It was probably unclear to the Vietnamese whether these provocative patrols had been approved at the highest levels of the US government. In contrast, our model illuminates how a provocative act can trigger conflict even if it is commonly known to be the act of a third party. For example, after the 2008 Mumbai terrorist attack, Indian government officials clearly distinguished between Pakistan’s civilian government, which India believed was not involved in the attacks, and the ISI, which is believed to be outside the control of Pakistan’s political leaders (Walsh 2010).
It is sometimes argued that the ISI wants to force India to relinquish Kashmir by making India’s presence in Kashmir costly. Our model suggests, however, that the ISI’s optimal strategy may depend on the preferences of Pakistan’s highest military and political leadership, because without their cooperation, the ISI will find it very difficult to drive India out of Kashmir. If Pakistan’s leaders are sufficiently hawkish, the ISI’s best option might be to develop a network of insurgents and lay the groundwork for a surprise attack by the Pakistani military, corresponding to the outcome HD (as in the 1999 Kargil war, for example). Since the ISI would not be aiming to provoke India, it would correspond to message $m_0$. Of course, if India understands this strategy, the absence of provocations will not be very reassuring. In contrast, if the ISI thinks Pakistan’s leaders are indecisive, the ISI’s best option might be to use provocations to raise tensions between the two countries (corresponding to message $m_1$). Recent provocations by ISI-sponsored militants occurred when Pakistan’s leaders were preoccupied with the “war on terror” rather than the struggle over Kashmir. According to our theory, these provocations were actually (moderately) good news, in the sense that they indicated the ISI believed Pakistan’s political leaders were not dominant strategy hawks on Kashmir. However, if the ISI thinks Pakistan’s current leaders are too weak to ever turn hawkish, the ISI’s best option may again be to lay the groundwork for a future conflict, anticipating the arrival of a more hawkish Pakistani leader. Since provoking India would not be the objective, it would again correspond to $m_0$. In this way, a non-monotonic strategy could come about naturally, perhaps without being explicitly formulated in advance.

In Section III we showed that the communication equilibrium is robust to a small amount of uncertainty about whether actions are truly strategic substitutes or complements. In reality there may be significant uncertainty on this point. For example, the Cold War was characterized by disagreements about whether toughness would make the Soviet Union back down or become more aggressive. The model of Baliga and Sjöström (2008) emphasized this kind of uncertainty, but there was no third party who manipulated the conflict. Third party manipulation in such environments is an interesting topic for future research.

**Appendix**

**Proof of Lemma 1:**

Suppose strategy $\mu$ 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 $m'$, so 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

---

13 Many other examples of this logic could be given. For example, the provocative takeover of the American embassy by Iranian radicals would signal that Iranian leaders were not dominant strategy hawks (i.e., not necessarily implacable foes of the United States). Hamas’ attacks during the Oslo peace accords and before Israeli elections would signal that the leaders of the Palestinian Authority were moderates who, unlike Hamas, wanted peace.
send whatever message minimizes the probability that player $B$ plays $H$. Call this message $m_0$. Thus,

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

Message $m_0$ is the unique minimizer of $c_B(m)$, since $c_B(m) \neq c_B(m_0)$ 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 $M = \{m_0, m_1\}$ and this completes the proof.

**PROOF OF THEOREM 3:**

The argument in the text showed that any communication equilibrium must have the following form. Player $E$ sends message $m_1$ if and only if $c_A \in (y^*, \Gamma(d)]$. Player $A$’s cutoff points are $c_A(m_0) = y^*$ and $c_A(m_1) = \Gamma(d)$. Player $B$’s cutoff points are $c_B(m_0) = x^*$ and $c_B(m_1) = d$. Moreover, $y^* = \Gamma(x^*)$ and $x^*$ is a best response to player $A$’s playing $H$ with probability $F(y^*)/[1 - F(\Gamma(d)) + F(y^*)]$. To show part (i) of the theorem, we need to show that such $x^*$ and $y^*$ exist.

Conditional on message $m_0$, player $A$ will choose $H$ with probability $F(y^*)/[1 - F(\Gamma(d)) + F(y^*)]$, so player $B$ prefers $H$ if and only if

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

Inequality (10) 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, $x^* = \Omega(y^*)$. We now show graphically that we can find $x^*$ and $y^*$ such that $x^* = \Omega(y^*)$ and $y^* = \Gamma(x^*)$. 
By Assumption 2, $\Gamma$ is increasing with a slope less than one. Since $F(c) = 0$ and $F(\bar{c}) = 1$, we have $\Gamma(c) = \mu > c$ and $\Gamma(\bar{c}) = d < \bar{c}$. Furthermore,

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

Therefore,

$$\Gamma(d) < d.$$ 

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 $[c, \bar{c}]$. Clearly, $\mu < \hat{x} < \Gamma(d)$ (see Figure 1).
Figure 3 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

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

so $\Omega$ is increasing. It is easy to check that $\Omega(y) > \Gamma(y)$ whenever $y \in (c, \Gamma(d))$. Moreover, $\Omega(c) = \Gamma(c) = \mu$ and

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

where the inequality follows from (11) and the fact that $\Gamma$ is increasing. These properties are shown in Figure 3. Notice that the curve $x = \Omega(y)$ lies to the right of the curve $x = \Gamma(y)$ for all $y$ such that $c < y < \Gamma(d)$ (because $\Omega(y) > \Gamma(y)$ for such $y$), but the two curves intersect when $y = c$ and $y = \Gamma(d)$.

As shown in Figure 3, the two curves $x = \Omega(y)$ and $y = \Gamma(x)$ must intersect at some $(x^*, y^*)$, and it must be true that

$$(12) \quad \hat{x} < y^* < x^* < \Gamma(d) < d.$$ 

By construction, $y^* = \Gamma(x^*)$ and $x^* = \Omega(y^*)$. Thus, a communication equilibrium exists. The welfare comparisons in part (ii) follow from the fact that $\hat{x} < y^* < x^*$ and the argument in the text.

Finally, part (iii) is equivalent to showing uniqueness of $(x^*, y^*)$. It can be verified that (5) 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 3.

**PROOF OF THEOREM 4:**

Consider the continuous function $F : [\mu, d]^2 \to [\mu, d]^2$, defined by

$$F(x, y) = \begin{bmatrix} F^x(x, y) \\ F^y(x, y) \end{bmatrix},$$

where

$$F^x(x, y) \equiv \frac{(1 - F(y))\mu + F(x)d}{1 - F(y) + F(x)}$$

and

$$F^y(x, y) \equiv \frac{(1 - F(d))\mu + (F(x) + F(d) - F(y))d}{1 - F(y) + F(x)}.$$

There exists a fixed point $(\tilde{x}, \tilde{y}) = F(\tilde{x}, \tilde{y})$. It is easy to check that $\mu < \tilde{x} < \tilde{y} < d$. 
Consider the strategy profile described in the text. Player $E_i$ maximizes his payoff by sending $m_1$ if and only if $c_i \in (\tilde{x}, \tilde{y}]$. Now consider player $A$. If player $E_B$ sends $m_1$, then player $B$ is expected to choose $H$. Therefore, player $A$ plays $H$ unless $D$ is his dominant strategy. Suppose instead that player $E_B$ sends $m_0$ and player $E_A$ sends $m_1$. Then either $c_B \leq \tilde{x}$ or $c_B > \tilde{y}$, and player $B$ chooses $H$ if and only if $c_B \leq d$. Therefore, the probability that player $B$ chooses $H$ is

$$\frac{F(\tilde{x}) + F(d) - F(\tilde{y})}{1 - F(\tilde{y}) + F(\tilde{x})}.$$ 

It can be checked that $\tilde{y} = F^y(\tilde{x}, \tilde{y})$ implies that player $A$’s type $\tilde{y}$ is indifferent between $H$ and $D$. Thus, the best response is to choose $H$ when $c_A \leq \tilde{y}$.

Finally, suppose both extremists send $m_0$. Again, either $c_B \leq \tilde{x}$ or $c_B > \tilde{y}$. Player $B$ chooses $H$ in the former case and $D$ in the latter case. Thus, the probability that player $B$ chooses $H$ is

$$\frac{F(\tilde{x})}{1 - F(\tilde{y}) + F(\tilde{x})}.$$ 

It can be checked that $\tilde{x} = F^x(\tilde{x}, \tilde{y})$ implies that player $A$’s type $\tilde{x}$ is indifferent between $H$ and $D$. Thus, the best response is to choose $H$ when $c_A \leq \tilde{x}$. Hence, player $A$ maximizes his payoff. The situation for player $B$ is symmetric.

PROOF OF THEOREM 5:

The argument in the text proves that there can be no “babbling” (uninformative) PBE. Communication equilibria (with informative messages) must have the familiar form. Arguing as in Section IIA, $y^{**} = \Gamma(x^{**})$ where $\Gamma$ is defined by equation (3), and $x^{**}$ is a best response to player $A$ playing $H$ with probability

$$\frac{F(y^{**})}{1 - F(\Gamma(R + d)) + F(y^{**})}.$$

The function $\Omega$ is modified to take $R$ into account:

$$\hat{\Omega}(y) \equiv \left[1 - F(\Gamma(R + d))\right] \mu + \frac{F(y)d}{\left[1 - F(\Gamma(R + d))\right] + F(y)}.$$

As before, it can be shown that the two curves $x = \hat{\Omega}(y)$ and $y = \Gamma(x)$ intersect at some point $(x^{**}, y^{**})$, where

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

There is only one intersection if (6) holds, so a unique communication equilibrium exists as before. Moreover, (6) guarantees that there is a unique continuation equilibrium following $m_0$. We need to show that there is also a unique continuation equilibrium following $m_1$. Specifically, following $m_1$ player $B$ must expect that
player $A$ will play $H$ and thus player $B$ plays $H$ if $c_B \leq R + d$ (i.e., unless $D$ is his dominant action following $m_1$).

Any continuation equilibrium must consist of a pair of cutoff 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 $c_A \in [y^{\ast\ast}, \Gamma(R + d)]$. If player $A$ uses a cutoff $y \in [y^{\ast\ast}, \Gamma(R + d)]$, player $B$ prefers $H$ if and only if

$$R - c_B + \frac{\mu(F(\Gamma(R + d)) - F(y))}{F(\Gamma(R + d)) - F(y^{\ast\ast})} \geq \frac{-d(F(y) - F(y^{\ast\ast}))}{F(\Gamma(R + d)) - F(y^{\ast\ast})}.$$  \hfill (14)

Inequality (14) is equivalent to $c_B \leq \Theta(y)$ where

$$\Theta(y) \equiv \frac{(d - \mu)F(y)}{F(\Gamma(R + d)) - F(y^{\ast\ast})} + R + \frac{\mu F(\Gamma(R + d))}{F(\Gamma(R + d)) - F(y^{\ast\ast})} - \frac{dF(y^{\ast\ast})}{F(\Gamma(R + d)) - F(y^{\ast\ast})}.$$  \hfill (15)

Thus, player $B$’s best response is $x = \Theta(y) \in [R + \mu, R + d]$. (Types below $R + \mu$ or above $R + d$ have dominant actions following $m_1$.)

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

$$R + \mu > x^{\ast\ast} = \frac{[1 - F(\Gamma(R + d))] \mu + F(y^{\ast\ast})d}{[1 - F(\Gamma(R + d))] + F(y^{\ast\ast})}.$$  \hfill (15)

Thus, $\Gamma(R + \mu) > y^{\ast\ast}$, and since $\Gamma$ is increasing, player $A$’s best response to $x \geq R + \mu$ is $y = \Gamma(x) > y^{\ast\ast}$.

So far we have shown that the cutoffs conditional on $m_1$ satisfy $x = \Theta(y) \geq R + \mu$ and $y = \Gamma(x) > y^{\ast\ast}$. In fact, the curves $y = \Gamma(x)$ and $x = \Theta(y)$ intersect at $(x, y) = (R + d, \Gamma(R + d))$ which yields the strategy played in the unique communication equilibrium: after message $m_1$, player $A$ is expected to play $H$ (all types $c_A \in (y^{\ast\ast}, \Gamma(R + d))$ play $H$) and player $B$ plays $H$ if $c_B \leq R + d$. The curves can have no other intersection if $F$ is concave, since both $\Gamma$ and $\Theta$ are concave and can intersect at most once in the relevant region where $x \in [R + \mu, R + d]$ and $y \in [y^{\ast\ast}, \Gamma(R + d)]$. Thus, the continuation equilibrium following $m_1$ is unique.

**PROOF OF THEOREM 7:**

Arguing as in Section IIA, $y^* = \Gamma(x^*)$, and $x^*$ is a best response to player $A$ playing $H$ with probability $F(\Gamma(\mu))/[1 - F(y^*) + F(\Gamma(\mu))]$. To show the existence of $x^*$ and $y^*$ is again a fixed-point argument. Let

$$\tilde{\Omega}(y) \equiv \frac{[1 - F(y)] \mu + F(\Gamma(\mu))d}{[1 - F(y)] + F(\Gamma(\mu))}.$$
The cutoffs \((x^*, y^*)\) are 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, \(\Gamma(c) = \mu < \bar{c}\) and \(\Gamma(\bar{c}) = d > c\). Also, 
\[
\Gamma(\mu) - d = (1 - F(\mu))(\mu - d),
\]
where 
\[
0 < (1 - F(\mu))(\mu - d) < \mu - d.
\]
Therefore,
\[
(16) \quad d < \Gamma(\mu) < \mu.
\]
Let \(\hat{x}\) be the unique fixed point of \(\Gamma(x)\) in \([c, \bar{c}]\). It is easy to check that \(d < \hat{x} < \mu\).

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 (16) and the fact that \(\Gamma\) is decreasing. Consider now the intersection of the two curves \(x = \tilde{\Omega}(y)\) and \(y = \Gamma(x)\). A figure analogous to Figure 3 reveals that there exists \((x^*, y^*) \in [c, \bar{c}]^2\) such that \(y^* = \Gamma(x^*)\) and \(x^* = \tilde{\Omega}(y^*)\), and
\[
(17) \quad d < \Gamma(\mu) < y^* < \hat{x} < x^* < \mu.
\]
This proves parts (i) and (ii) of Theorem 7. For part (iii), it can be checked that \((7)\) implies \(-1 < \tilde{\Omega}'(y) < 0\). Since \(-1 < \Gamma'(x) < 0\), the two curves \(x = \tilde{\Omega}(y)\) and \(y = \Gamma(x)\) intersect only once.

REFERENCES


