The Hobbesian Trap*

Sandeep Baliga
Northwestern University

Tomas Sjöström
Rutgers University

March 24, 2011

1 Introduction

According to Thucydides [29], there are three motives for war: greed, fear and honor. His analysis was elaborated on by Hobbes [14]:

> "So that in the nature of man, we find three principal causes of quarrel. First, competition; secondly, diffidence; thirdly, glory. The first maketh men invade for gain; the second, for safety; and the third, for reputation. The first use violence, to make themselves masters of other men's persons, wives, children, and cattle; the second, to defend them; the third, for trifles..." (Hobbes [14], p. 64).

Game theory helps us understand the third motive, honor, by showing why a reputation is worth fighting for (Milgrom and Roberts [22]). In this article we disregard the dynamic problem of building a reputation, and instead focus on how the other two motives, greed and fear, interact to cause conflicts in a static setting.

Thucydides [29] argued that the Great Peloponnesian War came about because, although neither side wanted war, each side became convinced war was imminent, and each side wanted the first-mover advantage (or at least to prevent the opponent from getting it). Expectations were rational: the war did come. As Thucydides describes it, Corinth, an ally of Sparta, had

*We are grateful to Michelle R. Garfinkel and Stergios Skaperdas for helpful comments.

become involved in a local conflict with Corcyra, a city state with no allies.[1] Corcyraean diplomats persuaded Athens to intervene against Corinth. The clinching argument was that war between Sparta and Athens was inevitable anyway:

> "If any here think that the war wherein we may do you service will not at all be, he is in an error and seeth not how the Lacedaemonians [Spartans], through fear of you, are already in labour of the war; and that the Corinthians, gracious with them and enemies to you, making way for their enterprize, assault us now in the way to the invasion of you hereafter, that we may not stand amongst the rest of their common enemies, but that they may be sure beforehand either to weaken us or to strengthen their own estate. It must therefore be your part, we offering and you accepting the league, to begin with them and to anticipate plotting rather than to counterplot against them" (Thucydides [29], p. 21).

Symmetrically, Corinthian diplomats chided the Spartans for not preempting the inevitable Athenian aggression:

> "And also now you connive at the Athenians who are not as the Medes, far off, but hard at hand, choosing rather to defend yourselves from their invasion than to invade them, and by having to do with them when their strength is greater, to put yourselves upon the chance of fortune" (Thucydides [29], p. 39).

Since the days of Thucydides, many wars have been explained by historians in terms of mutual fear and distrust. According to Hans Morgenthau, "the First World War had its origins exclusively in the fear of a disturbance of the European balance of power" (Morgenthau [23], p. 185):

> "First, the fear of hostile alliances led to the formation of the Triple Alliance. Then, the fear of the latter's dissolution led to the severance by Germany of the friendly relations with Russia.

[1] According to Thucydides, the Corinthians hated the Corcyraeans because they "contemned them and allowed them not their due honour in public meetings" (Thucydides [29], p. 16).

> Finally, the fear of the intentions of the Triple Alliance brought about the Franco-Russian Alliance. It was the mutual fears of these two defensive alliances, and the general insecurity created by the erratic character of the imperialistic utterances of William II, that inspired the diplomatic maneuvers during the two decades before the First World War" (Morgenthau [23], p. 64-65).

Furthermore:

> "It was this fear that motivated Austria in July 1914 to try to settle its accounts with Serbia once and for all, and that induced Germany to support Austria unconditionally. It was the same fear that brought Russia to the support of Serbia, and France to the support of Russia" (Morgenthau [23], p. 186).

The military planning before World War I has been called a Doomsday Machine: a mobilization by a single country, for whatever reason, was more or less guaranteed to trigger a general war (Kissinger [19], Chapter 8). Czar Nicholas II's decision to mobilize in July 1914 is usually attributed to Russian fears that Serbia, their most significant Balkan ally, might become an Austrian protectorate. But a contributing factor was the feeling among the Russian military that a European war was inevitable and that "we were in danger of losing it before we had time to unsheath our sword" (Kissinger [19], p. 215). Germany's military plan required that France must be defeated within six weeks, faster than Russia could become fully mobilized, in order to avoid the risk of a two-front war. Therefore, the Russian mobilization triggered a German attack on France, despite Russian assurances that its mobilization was not directed against Germany (Kissinger [19] p. 215) and even though "[t]here was not a single specific Russian demand on Germany or a single German demand on Russia, which merited a local war, much less a general one" (Kissinger [19] p. 206).

Arguments about the origins of wars, based on subjective readings of history, lead to inevitable disagreements. Did Gustavus Adolphus enter the Thirty Years War to uphold the honor of the Protestant religion, or because he feared that the Habsburg empire might come to dominate the Baltic region, or was it just greed (an attempt to grab all the commerce and customs revenues of the Baltic region)? Napoleon III was snubbed by Czar Nicholas I, who refused to call him "brother" (Kissinger [19], p. 106), but in 1852

Napoleon managed to persuade the Turkish Sultan to make him "Protector of the Christians in the Ottoman Empire". Czar Nicholas reacted by breaking off diplomatic relations with the Sultan, and the Crimean War soon followed. Did Nicholas want to restore his honor as protector of the orthodox religion, or was he just pursuing the traditional Russian goal of controlling the Straits ? Russia's territorial expansion from Peter the Great to the fall of Communism is often attributed to a sense of insecurity (Kissinger [19], Ch. 6), so a desire to control the Straits could be attributed to either fear or greed. Probably all three of Thucydides's motives were intermingled in the minds of King Gustavus and Czar Nicholas.

Economists typically try to understand the logic of complex events by breaking them up into small pieces and studying each piece separately, using the simplest possible model. In this article, we hope to contribute one piece to the puzzle of why conflicts occur. Specifically, we study how uncertainty generates fear and triggers conflict when actions are strategic complements. In passing, we point out that when actions are strategic substitutes, uncertainty can instead promote peace by keeping greed in check.[2]

In Section 2, we use a stag-hunt game with payoff-uncertainty to understand how fear and greed interact. Section 3 shows how the players can use cheap-talk messages to create a peaceful outcome, but Section 4 explains how a third-party provocateur can use cheap-talk to create conflict. Section 5 argues that more democratic states are not necessarily more peaceful. Section 6 discusses a simple model of arms inspections, and Section 7 concludes.

2 The Conflict Game

In the basic conflict game, two players, A and B, simultaneously choose a *hawkish* (aggressive) action H or a *dovish* (peaceful) action D. Players A and B are the pivotal decision-makers in countries A and B, for simplicity referred to as the "leaders" of their countries. Different interpretations of the actions are possible. Action H might represent entering a disputed territory (while D represents not doing so). Alternatively, H might represent buying or developing new weapons in an arms race. The game is played only once.

If both players choose D, then they coexist peacefully, and payoffs are

[2]For theories of conflict without uncertainty about preferences or capabilities, see Fearon [8], Fearon [9], Garfinkel and Skaperdas [11], Powell [26] and Jackson and Morelli [15].

normalized to zero. If player $i \in \{A, B\}$ chooses H, he incurs a cost $c_i \geq 0$. In addition, if player j chooses D when player i chooses H, player i gets $\mu > 0$, which can be interpreted as a gain from being on the offensive, i.e., a first-mover advantage. Symmetrically, if player i chooses D while player j chooses H, then player i incurs a cost $d > 0$ of being on the defensive. The parameters d and μ can be said to represent the motives of fear and greed, respectively. The game is not zero-sum: a conflict destroys value. Accordingly, the gain of being on the offensive is smaller that the cost of being on the defensive: $\mu < d$. Player i's payoffs are summarized by the payoff matrix (1), where the row represents the choice of player i and the column the choice of player j.

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

For simplicity, d and μ are the same for each player, but it is possible that $c_A \neq c_B$. Player i is a *coordination type* if $\mu < c_i < d$, a *dominant strategy hawk* if $c_i < \mu$, and a *dominant strategy dove* if $c_i > d$.

According to Hobbes [14], reason dictates that people "seek peace, and follow it" but also "by all means we can, to defend ourselves":

> "every man ought to endevor peace, as far as he has hope of obtaining it; and when he cannot obtain it, that he may seek, and use, all helps, and advantages of war" (Hobbes [14], p. 66).

In effect, Hobbes describes a stag-hunt game, with D the best response to D ("seek peace, and follow it") and H the best response to H ("by all means we can, to defend ourselves"). Formally, our game is a stag-hunt game if payoffs are common knowledge and both players are coordination types. The stag-hunt has two Nash equilibria, HH and DD. Notice that DD Pareto dominates HH. But Hobbes singled out HH, "war of every man against every man", as the more likely outcome in a "state of nature". In fact, Hobbes gave a striking argument in favor of this outcome, namely, that there may be some types who actually desire conflict, and this causes everyone to be aggressive in self-defense:

> "Also because there be some, that taking pleasure in contemplating their own power in the acts of conquest, which they

> pursue farther than their security requires; if others, that otherwise would be glad to be at ease within modest bounds, should not by invasion increase their power, they would not be able, long time, by standing only on their defense, to subsist" (Hobbes [14], p. 64).

Since the opponent's true type would normally be impossible to know for sure, we introduce payoff uncertainty. For simplicity, suppose c_A and c_B are independently drawn from a distribution F which is uniform on $[0, \bar{c}]$.[3] Thus, for $0 \leq c \leq \bar{c}$, the probability that $c_i \leq c$ is $F(c) = c/\bar{c}$. Each player i knows his own type, c_i, but not the opponent's type, c_j. Player i is naturally more aggressive, the lower is his cost of aggression c_i. A *strategy* for player i is a *cutoff point* x_i such that player i chooses D if $c_i \geq x_i$ and H if $c_i < x_i$.

Dominant strategy hawks take "pleasure in contemplating their own power in the acts of conquest" so H is their dominant strategy. Their behavior may be said to be completely ruled by *greed*. To eliminate the uninteresting case where each player is sure to be a dominant strategy hawk, assume $\bar{c} > \mu$. The probability that a player is a dominant strategy hawk is $F(\mu) = \mu/\bar{c} < 1$.

The conflict game with payoff uncertainty has a unique Bayesian Nash equilibrium. If $\bar{c} < d$, so dominant strategy doves are ruled out, then each player chooses H with probability one (Baliga and Sjöström [2]). In particular, the coordination types must all choose H in equilibrium. They are unable to coexist peacefully even if μ is small so each player is very likely a coordination type. Their behavior is completely ruled by *fear*.[4] This is the *Hobbesian trap* or *Schelling's dilemma* (see Schelling [27], Jervis [16] and Kydd [20]).[5]

[3] See Baliga and Sjöström [4] for more general distributions, including correlated types. In contrast to the theory of "global games", we do not focus on the case where information is highly correlated. The theory of global games is applied to conflicts by Chassang and Padró i Miquel [6].

[4] Recall that the parameter d represents the fear of being taken advantage of. If $d > \bar{c}$ then this fear dominates the cost of being aggressive even for the most peaceful type $\bar{c}$.

[5] An early statement of the dilemma is due to Rousseau, quoted by Jervis ([16], p. 63):

> "It is quite true that it would be much better for all men to remain always at peace. But so long as there is no security for this, everyone, having no guarantee that he can avoid war, is anxious to begin it at the moment which suits his own interest and so forestall a neighbor, who would not fail to forestall the attack in turn at any moment favorable to himself, so that many

To understand the "fear-spiral" underlying the Hobbesian trap, suppose player i is "almost dominant strategy hawk" in the sense that

$$\mu < c_i < \mu + F(\mu)(d - \mu).$$

Since $c_i > \mu$, he would play D if he were convinced the opponent plays D. Unfortunately, he cannot be so convinced, because with probability $F(\mu) > 0$ the opponent is a dominant strategy hawk. Since $(1 - F(\mu))\mu - c_i > -F(\mu)d$, the fear of dominant strategy hawks is sufficient to cause the "almost dominant strategy hawk" to choose H. By a similar argument, the fear that the opponent is either a dominant strategy hawk or an "almost dominant strategy hawk" (both of which we know choose H) forces even types with c_i slightly above $\mu + F(\mu)(d - \mu)$ to choose H. Continuing this argument, the contagion causes higher and higher types to choose H. If $\bar{c} < d$ then there is no barrier (no dominant strategy doves) to stop the fear-spiral from infecting the whole population.

If $\bar{c} > d$ then the unique Bayesian Nash equilibrium is interior. Suppose player j uses cutoff x. Then, player i's expected payoff from playing D is $-dF(x)$, and his payoff from playing H is $\mu(1 - F(x)) - c_i$. Player i is indifferent between playing H and D if and only if

$$c_i = \Gamma(x) \equiv \mu + F(x)(d - \mu) = \mu + \frac{x}{\bar{c}}(d - \mu). \tag{2}$$

If player j uses cutoff x, then player i's best response is to use cutoff $\Gamma(x)$. As $\mu < d < \bar{c}$, we get

$$0 < \Gamma'(x) < 1. \tag{3}$$

Interpreting Γ as a best-response curve, condition (3) is a well-known condition guaranteeing a unique equilibrium (Baliga and Sjöström [5]). The unique equilibrium must be symmetric: the cutoff point $\hat{x}$ satisfies $\hat{x} = \Gamma(\hat{x})$ and, using equation (2), it can be computed explicitly:

$$\hat{x} = \frac{\mu\bar{c}}{\bar{c} - (d - \mu)}. \tag{4}$$

---

wars, even offensive wars, are rather in the nature of unjust precautions for the protection of the assailant's own possessions than a device for seizing those of others."

The comparative statics are as expected. For example, an increase in d makes the players more fearful, which raises the equilibrium cutoff $\hat{x}$ and leads to more conflict.

Since $d > \mu$, actions are strategic complements: a player is more inclined to choose H if he thinks his opponent is likely to choose H. Formally, the best-response curve $\Gamma(x)$ is upward sloping: $\Gamma'(x) > 0$. Strategic complementarity drives the fear-spiral which causes aggression to escalate into conflict, as Hobbes imagined it would in a "state of nature". In Section 6, we consider the possibility that actions may be strategic substitutes, as in a game of chicken.

3 Cheap Talk

Players A and B do not know each others' true types. This causes mutual fear and distrust, which leads to the Hobbesian trap. It is natural to ask if it can be mitigated by communication.

Consider the conflict game with payoff uncertainty described in Section 2. Suppose $\bar{c} < d$ so that without communication all types choose H in the unique Bayesian Nash equilibrium. In the cheap-talk extension of the game, players A and B exchange costless messages before choosing their actions (H or D). Messages have no direct effect on payoffs: each player i's payoff matrix is still the matrix (1). However, the messages might change the players' beliefs about each other.

A naive intuition suggests that coordination types should announce their true preferences and then go on to play DD, thus escaping the trap. This intuitive argument encounters the following objection. As $\mu > 0$ and $d > 0$, all types want the opponent to choose D, whatever they themselves plan to do. So wouldn't all types, including dominant strategy hawks, send whatever message is most likely to convince the opponent to choose D? But then communication would not change the players' beliefs about each other and could not prevent the Hobbesian trap.

It turns out that the objection can be overcome: if μ is small, then there exist informative cheap-talk equilibria where the probability of the outcome DD is close to one (Baliga and Sjöström [2]). To understand this result, observe that although all types want to increase the probability that the opponent chooses D, coordination types also want to avoid a coordination failure: they want to know the opponent's action in order to know how to

respond. In contrast, dominant strategy types have no interest in finding out what the opponent will do. Since different types trade off these two objectives at different rates, it is possible to induce different types to send different messages.

Based on two cutoffs c^L and c^H, where $\mu < c^L < c^H < \bar{c}$, Baliga and Sjöström [2] partition the type space $[0, \bar{c}]$ into three sets: *very tough*, *fairly tough* and *peaceable*.[6] The "very tough" types have costs below c^L; this includes all dominant strategy hawks. The "fairly tough" types are coordination types with costs between c^L and c^H. The "peaceable" type are the most peaceful ones, with costs of aggression above c^H. The very tough types and the peaceable types mainly want to minimize the probability that the opponent plays H. But the fairly tough types put a lot of value on obtaining information about the opponent's action in order to coordinate with him.

At the cheap-talk stage, the two players simultaneously say either "Hawk" (an aggressive message) or "Dove" (a conciliatory message). Saying Dove will minimize the probability that the opponent plays H in the next (action) stage, but the uncertainty about the opponent's action is not resolved. Saying Hawk yields a higher probability that the opponent plays H, but after the talk there will be no ambiguity about the opponent's action. In equilibrium, the very tough types and the peaceable types say Dove in order to minimize the probability that the opponent plays H, while the fairly tough types say Hawk in order to minimize the probability of a coordination failure.

The equilibrium actions are as follows. If *both* players say Hawk, then *neither* plays H in stage two. (This is continuation equilibrium because the messages imply that neither player is a dominant strategy hawk.) If one player says Hawk and the other says Dove, then both players choose H in the action stage. (This is continuation equilibrium because there are no dominant strategy doves by hypothesis). Finally, if both players say Dove, then a player who is very tough will choose H in the action stage, while peaceable types choose D.

An exchange of dovish messages convinces each player that the opponent is *either* peaceable *or* very tough. Why is it now continuation equilibrium for peaceable types to choose D? Intuitively, as fairly tough types are ruled out, the contagion described in Section 2 is blocked, and the conditional distribution becomes more conducive to cooperation among the peaceable types. Of course, with some probability a peaceable type encounters a very

[6] Baliga and Sjöström [2] call the peaceable types "normal".

tough type and gets $-d$. Still, the peaceable types are willing to trust an opponent who sends a conciliatory message, as long as very tough types are sufficiently rare (which requires μ, and hence $F(\mu)$, to be small).

In the cheap-talk stage, peaceable types and tough types prefer to say Dove. The peaceable types prefer their "sincere dovish strategy", because saying Dove allows them to coexist peacefully with other peaceable types. The very tough types prefer their "insincere dovish strategy", because saying Dove allows them to take advantage of unsuspecting peaceable types. That is, by masquerading as doves, the very tough types get to play H unilaterally against the peaceable types, who cannot tell a very tough opponent from a peaceable one.

The key to the equilibrium construction is the incentive of fairly tough types to separate themselves out by saying Hawk. By saying Hawk, they will always get to coordinate with the opponent: they coordinate on DD with other fairly tough types, and on HH with everyone else. Hence, the expected payoff of a fairly tough type with cost c is

$$-\left(1-F(c^H)\right)c - F(c^L)c. \tag{5}$$

Suppose a fairly tough type deviates and says Dove. If the other player says Hawk, it is certainly optimal to go on to choose H. If the other player says Dove, the fairly tough type can either go on to behave like a peaceable type and choose D ("the first option"), or like a very tough type and choose H ("the second option").

The first option is most attractive to type c^H, as he has the highest cost among fairly tough types. His payoff would be

$$-(F(c^H)-F(c^L))c^H - F(c^L)d. \tag{6}$$

Being the highest type who says Hawk, type c^H must be indifferent between his equilibrium strategy and the first option. Hence, c^H is defined by the equality of expressions (5) and (6), which yields

$$\left[1-2\left(F(c^H)-F(c^L)\right)\right]c^H = F(c^L)d. \tag{7}$$

The second option is most attractive to type c^L as he has the lowest cost among fairly tough types. His payoff would be

$$-c^L + \left(1-F(c^H)\right)\mu. \tag{8}$$

Being the lowest type who says Hawk, type c^L must be indifferent between his equilibrium strategy and the second option. Hence, c^L is defined by the equality of expressions (5) and (8), which yields

$$\left[F(c^H) - F(c^L)\right] c^L = \left(1 - F(c^H)\right) \mu. \tag{9}$$

The equilibrium requires that c^L and c^H simultaneously solve equations (7) and (9). Baliga and Sjöström [2] showed that as long as μ is small, such a solution exists, and almost all types are peaceable. Therefore, in equilibrium the probability will be close to one that both players say Dove and then play DD.

Theorem 1 *(Baliga and Sjöström [2]). Fix any $\delta > 0$. There is $\bar{\mu} > 0$ such that if $0 < \mu < \bar{\mu}$ then there exists a perfect Bayesian equilibrium with informative cheap-talk, where the outcome is DD with probability greater than $1 - \delta$.*

Recall that we are assuming $\bar{c} < d$, so each player chooses H with probability one in equilibrium without communication, even if μ is very small. But the theorem implies that informative cheap-talk can reduce the probability of choosing H to almost zero when μ is small. That is, the Hobbesian trap can be almost completely escaped. Again, the key assumption is that μ is small, i.e., each player is very unlikely to be a dominant strategy hawk. It is easy to see why this assumption is needed, because if the opponent is likely to be a dominant strategy hawk, nothing can persuade a coordination type to trust the opponent and choose D.

4 Provocation

Section 3 showed how cheap-talk between players A and B can break the fear-spiral which underlies the Hobbesian trap. In this section we will argue that cheap-talk by a third party "provocateur" can inflame the fear-spiral and deepen the trap. Real-world examples of provocations are not hard to find. For example, Ariel Sharon's symbolic visit to the Temple Mount in September 2000 helped spark the Second Intifada and derail the Israeli-Palestinian peace process (Hefetz and Bloom [13]).

Following Baliga and Sjöström [5], suppose before players A and B play the conflict game with payoff uncertainty, a third party, player E, publicly

announces either "Hawk" (an aggressive message) or "Dove" (a conciliatory message).[7] We think of player E as an "extremist" from country A, and interpret the hawkish message as a "provocation" (e.g., a visit to the Temple Mount). Player E takes no action except sending a message, and players A and B do not send any messages.

The payoff matrix for each player $i \in \{A, B\}$ is again the matrix (1). Player E's payoff matrix 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.

We assume $c_E < 0$, i.e., aggression is inherently beneficial to player E. Therefore, player E is guaranteed a strictly positive payoff if player A chooses H (he gets either $-c_E > 0$ or $\mu - c_E > 0$), but he gets a non-positive payoff if player A chooses D (either $-d < 0$ or 0). Accordingly, player E surely wants player A to choose H.

As before, players A and B do not know each others' true types, but c_E is commonly known. Also, assume player E knows c_A (but not c_B). This greatly simplifies the analysis, because in equilibrium, player E will know player A's reaction to player E's message. (Consider that Ariel Sharon is very familiar with Israeli public opinion and the intentions of the Israeli government.)

Suppose $\bar{c} > d$ so, as shown in Section 2, without communication the unique Bayesian Nash equilibrium is interior: types above an equilibrium cutoff point $\hat{x}$, given by equation (4), choose D. We will argue that player E can use cheap-talk to increase the risk of conflict above the level of the communication-free equilibrium. It is surprising that player E can do this. After all, it is commonly known that player E is a provocateur who takes pleasure in aggression (as $c_E < 0$), and if his cheap-talk triggers conflict it will make players A and B worse off. So why do they allow themselves to be manipulated by player E?

In equilibrium each player $j \in \{A, B\}$ uses a "conditional" cutoff strategy: for any message $m \in \{Hawk, Dove\}$, there is a cutoff $c_j(m)$ such that if player j hears message m, then he chooses H if and only if $c_j \leq c_j(m)$. Following the provocative message "Hawk", the equilibrium cutoffs are $c_A(Hawk) = \Gamma(d)$ and $c_B(Hawk) = d$, where Γ is defined by equation (2). Notice that this means that after a provocation player B will choose H with probability $F(d)$;

[7] A related model of provocation is provided by Jung [17].

therefore, 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's cutoff point $c_A(Hawk) = \Gamma(d)$ is a best-response.

Now let $y^* = c_A(Dove)$ and $x^* = c_B(Dove)$ denote the cutoff points if there is no provocation. In equilibrium, it will be the case that $x^* < d$, so if there is no provocation player B will choose H with probability strictly less than $F(d)$. For player A to use a best response, we must have $y^* = \Gamma(c_B(Dove)) < \Gamma(c_B(Hawk)) = \Gamma(d)$. That is, both players use lower cutoff points, meaning the fear-spiral is less severe, in the absence of a provocation.

Player E's equilibrium strategy is to send the hawkish message if and only if $c_A \in (y^*, \Gamma(d)]$. The provocation makes player B more likely to play H. Player E does not want player B to choose H. However, when $c_A \in (y^*, \Gamma(d)]$, the provocation also causes player A to switch from D to H, and this is what player E wants. In contrast, when $c_A \notin (y^*, \Gamma(d)]$, player A's action does not depend on player E's cheap-talk message, and player E prefers say "Dove" in order to minimize the probability that player B chooses H. Indeed, if $c_A \leq y^*$ then player A is inherently so hawkish as to choose H even in the absence of a provocation; conversely, if $c_A > \Gamma(d)$ then player A is dovish enough to choose D even following a provocation. In both cases, a provocation would backfire, as it would simply inflame player B with no benefit to player E.

In equilibrium, a provocation only occurs when player A is a coordination type 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). Thus, the provocation exacerbates the fear-spiral: each player behaves aggressively because he expects the other will (just as in a "bad" HH equilibrium of a complete-information stag-hunt game).

Curiously, the *absence* of a provocation *also* inflames player B. In *the curious incident of the dog in the night-time* (Conan Doyle [7]), the dog did not bark at an intruder because the dog knew him well. Similarly, when player A is inherently a very hawkish type, the extremist does not behave provocatively. Hence, "an extremist who does not bark" (i.e., who says Dove) alerts player B that player A might be a very hawkish type. This triggers a fear-spiral, and both players A and B are more likely to play H than in the communication-free equilibrium. (The cutoff points x^* and y^* are strictly higher than the equilibrium cutoff point $\hat{x}$ given by equation (4)). Accordingly, the presence of the extremist is bad for peace *no matter which message*

he sends. Any type that would have chosen H in the communication-free equilibrium of Section 2 necessarily chooses H in the equilibrium described here. But, in the equilibrium described here, whether or not a provocation occurs, there are types who choose H 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 the presence of the extremist, because each wants the opponent to choose D.

There is an interesting contrast between the "benevolent" cheap-talk of Section 3, which prevented conflict, and the "malevolent" cheap-talk of the current section, which triggers conflict. In both cases, the cheap-talk has a non-convex structure, in the sense that intermediate types get separated out from the rest. In Section 3, it was "tough" coordination types who separated themselves out, bringing peace by preventing the contagion from infecting the whole population with fearfulness. The intermediate types themselves coexisted peacefully! In contrast, extremist cheap-talk 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. Even "an extremist who does not bark" is bad for peace, because the absence of "weak" coordination types leads to a less favorable type-distribution.

Baliga and Sjöström [5] show that *all* equilibria with extremist communication have the same structure and always make both players A and B worse off. Why can't players A and B simply disregard the extremist ? This question is often asked 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 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 [10], p. 697.)

In our model, the question has an obvious answer: since players A and B expect each other to respond to the extremist's message, unilaterally disregarding the message is not optimal. The more difficult question of whether

players A and B can *jointly* deviate, by some sort of self-enforcing agreement to disregard the extremist, is discussed by Baliga and Sjöström [5], who argue that this is not necessarily the case.

5 Democratic Peace

The idea that democracy promotes peace is associated with Immanuel Kant:

> "If...the consent of the subjects is required to determine whether there shall be war or not, nothing is more natural than that they should weigh the matter well, before undertaking such a bad business" (Immanuel Kant [18], p. 122).

However, if conflict is due to the Hobbesian trap, and if the representative citizen is a fearful type, then it is not obvious that democratic reforms will create peace.

Baliga, Lucca and Sjöström [1] extended the conflict game with payoff uncertainty by assuming each player is the leader of a country with a continuum of citizens. After the two leaders have chosen H or D, each citizen will support his leader if and only if the leader's action was a best-response, *according to the citizen's own preferences.* To stay in power, leader $i \in \{A, B\}$ needs a critical level of support σ_i^* among his citizens. The value of staying in power is $R > 0$.

For example, suppose leader A chooses H and leader B chooses D. Leader A is then supported by those citizens of country A who think H is a best response to D, i.e., the dominant strategy hawks, while leader B is supported by the dominant strategy doves in country B. If the distribution of cost types in each population is F, then leaders A and B are supported by fractions $F(\mu)$ and $1 - F(d)$ of their populations, respectively. If $F(\mu) \geq \sigma_A^*$, then leader A stays in power, and his payoff is $\mu - c_A + R$ (where c_A is his private cost type). But if $F(\mu) < \sigma_A^*$, then leader A loses power, and his payoff is only $\mu - c_A$. (Similarly, leader B remains in power if and only if $1 - F(d) \geq \sigma_B^*$).

Baliga, Lucca and Sjöström [1] assume there are more dominant strategy hawks than doves, and the median type is a coordination type:

$$1 - F(d) < F(\mu) < 1/2.$$

Under this assumption, each country i falls in one of three categories, depending on σ_i^*.

First, if σ_i^* is small enough that leader i never loses power, then country i is a *dictatorship*. The dictator's payoff matrix is the matrix (1), with R added to each entry. Of course, the dictator is unconcerned with the opinions of his citizens.

Second, if σ_i^* is large enough that leader i needs the support of the *median* type to stay in power, then country i is *full democracy*. Recall that the median type is a coordination type by assumption. Therefore, if country i is a full democracy then leader i stays in power if and only if he matches the action of the opponent, giving leader i the payoff matrix (10):

$$\begin{array}{ccc} & H & D \\ H & R-c_i & \mu-c_i \\ D & -d & R \end{array} \tag{10}$$

To stay in power and collect R, the leader of a full democracy is inclined to choose D against a peaceful opponent ("dovish bias"), but he is inclined to choose H against an aggressive opponent ("hawkish bias"). The "dovish bias" produces a "Kantian peace" between two full democracies.[8] However, in a more hostile environment, the median voter supports aggression out of fear, and will replace a leader who is not aggressive enough, producing a hawkish bias. In contrast, a dictator is not responsive to the preferences of his citizens, so there is neither a hawkish nor a dovish bias. Accordingly, as shown by Baliga, Lucca and Sjöström [1], a dyad of two dictators is less peaceful than a fully democratic dyad, but a dictator responds less aggressively than a democratically elected leader to increased threats from abroad.

Third, if σ_i^* lies between $1-F(d)$ and $F(\mu)$, then leader i loses power if and only if he chooses D while the opponent chooses H. This category is intermediate between dictatorship and full democracy; Baliga, Lucca and Sjöström [1] label it *limited democracy*. Here, leader i's payoff matrix is the matrix (11):

$$\begin{array}{ccc} & H & D \\ H & R-c_i & R+\mu-c_i \\ D & -d & R \end{array} \tag{11}$$

A limited democracy *always* has a hawkish bias, since to stay in power the leader must avoid the outcome DH. Baliga, Lucca and Sjöström [1] show

[8] Other theories of the Kantian peace based on incomplete information are provided by Levy and Razin [21] and Tangeras [28].

that replacing any other regime type in country i with a limited democracy increases the equilibrium probability of conflict.

In this model, the leader of a limited democracy risks losing power if he appears too dovish. Therefore, limited democracies behave more hawkishly than all other regime types (including dictatorships). By triggering the fear spiral, limited democracy is bad for peace. In full democracies, if the citizens feel safe they want a dovish leader, but if they feel threatened they want a hawkish leader. Thus, if the environment is perceived as hostile, full democracies have a hawkish bias. In short, if conflict is due to the Hobbesian trap, the relationship between democracy and peace is not straightforward.

6 Arms Inspections

Military capabilities are often kept secret. For example, Saddam Hussein possessed weapons of mass destruction (WMD) in the early 1990's, but not in the late 1990's. In neither situation did he reveal the truth. In the first situation, he may have wanted to avoid sanctions or preemptive strikes by not revealing his WMD; in the second situation, he may have wanted to create "deterrence by doubt" by not revealing that he lacked WMD. Whatever the motive, intuition suggests that ambiguity about military capabilities creates fear and mistrust and hence, by the familiar Hobbesian argument, fuels conflict. Conversely, arms inspections might promote peace by eliminating ambiguity. This conventional wisdom is embodied in the *Treaty on the Non-proliferation of Nuclear Weapons*, which requires that nations submit to inspections of nuclear facilities by the IAEA.

Consider a simple model of arms inspections based on Baliga and Sjöström [3]. Player A is the leader of a major power who has to decide whether or not to attack the smaller country B. Player B has to decide how much country B should invest in a weapons program. Let x denote the investment. For country B to acquire WMD, the weapons program must be successful, which is more likely the greater is x. If country B acquires WMD, we say country B is *armed*. Thus, there are two possible states that country B can be in: *armed* or *unarmed*. The probability that country B is armed is $\sigma(x)$, where $d\sigma/dx > 0$.

The time line is as follows.

1. Player B chooses x, which is observed by player A.

2. The state of country B is realized ("armed" with probability $\sigma(x)$, "unarmed" with probability $1-\sigma(x)$). Player B privately observes the true state.

3. Player B may reveal country B's true state ("armed" or "unarmed") to player A.

4. Player A may decide to attack country B.

The information revealed at stage 3 is "hard", i.e., impossible to falsify. If B is truly unarmed, weapons inspectors can verify this; if B is armed, player B can simply reveal the weapons. But an unarmed country cannot reveal any weapons, and - in our idealized model - weapons inspectors will never certify that an armed country is unarmed.[9] The assumption that player A can directly observe player B's stage 1 investment level is made for simplicity.[10]

If, in equilibrium, player B reveals the true state when he is armed, but not when he is unarmed, there is never any real ambiguity; for when player B doesn't reveal the true state, player A can deduce that B must be unarmed, knowing B's equilibrium strategy. Similarly, there is no real ambiguity if player B reveals the true state only when he is unarmed. For real ambiguity to exist in the mind of player A, player B must (like Saddam Hussein) refrain from revealing the true state *both when he is armed and when he is unarmed.*

Consider now the preferences of player A. Suppose player A is a *fearful type*: he would like to live in peace, but he fears that player B's WMD will end up in the hands of terrorists. In a version of the Hobbesian trap, such a fearful player A may feel compelled to attack country B in order to eliminate the threat.[11] Indeed, player A would be more fearful, hence more inclined to attack, the more likely it is that country B is armed with WMD. If player B knows that A is fearful in this sense, then if B is unarmed, he should allow weapons inspectors to verify this. Player A's fear would then be reduced, and an attack less likely to occur. Conversely, if player B does *not* allow weapons inspections, player A must conclude that player B is armed. Therefore, if

[9] In contrast, the information revealed in Section 3 was "soft" information about preferences, which can be falsified (e.g., a dominant strategy hawk can say "Dove").

[10] Baliga and Sjöström [3] assume player B's investment is unobserved, which adds another level of uncertainty in the mind of player A.

[11] Recall that the stated purpose of the invasion of Iraq in 2003 was to disarm Iraq of weapons of mass destruction.

player A is commonly known to be fearful, there will never be any ambiguity about country B's true state when player A makes his decision at stage 4.

Suppose instead that player A is a *greedy type* who would like to control country B's natural resources. But an attack on country B would be less tempting if country B has WMD to defend itself. Therefore, a greedy player A would be *less* inclined to attack, the more likely it is that country B is armed with WMD. If player B knows that A is greedy in this sense, then if B is armed, he should reveal this in order to deter an attack.[12] If player B does *not* reveal that he has WMD, player A must conclude that player B is unarmed. Therefore, if player A is commonly known to be greedy, there will never be any ambiguity about country B's true state when player A makes his decision at stage 4.

The above reasoning implies that player B can only create ambiguity in the mind of player A if player B is *unsure* about whether player A is greedy or fearful.[13] Formally, suppose player B thinks player A is a greedy type with probability $p > 0$ and a fearful type with probability $1 - p > 0$. If player B reveals that he is unarmed, then player A prefers to attack if he is the greedy type, but not if he is the fearful type. If player B reveals that he is armed, then player A prefers to attack if he is the fearful type, but not if he is the greedy type. More generally, there will exists $x_g \in (0, 1)$ such that the greedy type prefers to attack iff the probability that player B is armed is less than x_g, and $x_f \in (0, 1)$ such that the fearful type prefers *not* to attack iff the probability that player B is armed is less than x_f.

To see how ambiguity can exist in equilibrium, suppose $x_f > x_g$. If player B invests less than x_g at stage 1, then he is attacked with positive probability at stage 4. Indeed, if he never reveals the state at stage 3, the greedy type attacks at stage 4 (because $x < x_g$ does not deter). If, on the other hand, he reveals the true state at stage 3, he is either attacked by the fearful type (if he reveals that he is armed) or by the greedy type (if he reveals that he is unarmed). In contrast, if player B invests the amount $x = x_g$, and never reveals the state at stage 3, then he is never attacked. Indeed, the greedy type is "deterred by doubt" as $x \geq x_g$, but the fearful type is not sufficiently fearful to attack as $x < x_f$. It is clear that, if the cost of investing is not

[12] To maintain secrecy about weapons that could deter an attack would be irrational. As *Dr. Strangelove* put it, "the whole point of a Doomsday Machine is lost if you keep it a secret! Why didn't you tell the world, EH?"

[13] Alternatively, ambiguity can exist if player B faces several opponents, some greedy and some fearful.

too high, and if acquiring WMD does not have a very large intrinsic value to player B, then investing $x = x_g$ and maintaining complete ambiguity about his arsenal is his optimal strategy. Notice that if player B's weapons program is successful, he prefers not to reveal it, in order not to risk an attack from a fearful player A. If player B's weapons program is unsuccessful, he prefers not to reveal it, in order not to risk an attack from a greedy player A. So the strategy is sequentially rational. Moreover, the policy of ambiguity means player A never attacks.[14]

Suppose the game is changed so that at stage 3, the state of country B is automatically revealed. For example, new technology might allow player A to monitor country B's weapons program; or, more fancifully, a regime of mandatory arms inspections may be imposed by a "world government". In any case, there can be no ambiguity about country B's weapons capabilities. Now player B faces the following dilemma. If he acquires WMD, he might trigger an attack from the fearful type. If instead he remains unarmed, he might suffer an attack from the greedy type. Depending on player B's beliefs about player A's type, player B may prefer to acquire WMD, by increasing the size of his weapons program, in order to deter the greedy type. This would of course be bad for player A. Again, without ambiguity there will necessarily be attacks in equilibrium: if player B's weapons program succeeds, he is attacked by the fearful type; otherwise he is attacked by the greedy type. Eliminating the ambiguity about player B's weapons capabilities can be bad for peace, a result which contradicts the conventional wisdom discussed above.

Equilibrium ambiguity requires doubt about whether player A will respond to toughness (i.e., weapons acquisition) with escalation (as in a game with strategic complements) or by backing down (as in a game with strategic substitutes). Without this doubt, there is at least one state of the world where player B wants player A to know the truth and will reveal it; and this allows player A to deduce the truth also in the other state. But if doubt exists about whether actions are strategic substitutes or complements, then strategic ambiguity about weapons capabilities might make *all* players better off.

[14] If instead $x_f < x_g$, ambiguity can still be part of an optimal policy for player B. But now the probability of an attack must be strictly positive in equilbrium, for any investment level that is large enough to deter the greedy type will necessarily trigger an attack from the fearful type.

7 Conclusion

In the basic conflict game of Section 2, each player thinks there is a non-zero probability that the opponent is a dominant strategy hawk. Since actions are strategic complements, a contagion of fear causes even peaceful types to behave aggressively. Face-to-face communication can prevent the contagion. Cheap-talk cannot be fully revealing, but Section 3 showed that a subset of coordination types can be separated out from the rest. The remaining coordination types are willing to behave peacefully as long as μ is small. When μ is large, however, the informative cheap-talk equilibrium breaks down. The Revelation Principle (Myerson [24]) suggests that *mediation* might help. Future research might show if, for some range of μ, a mediator can produce peace when cheap-talk fails.

As discussed in Section 4, a third-party provocateur can use cheap-talk to aggravate fear and create conflict. Inciting a dominant strategy hawk is redundant, and inciting a dominant strategy dove is impossible. But the provocateur can cause coordination types to become more aggressive. The extent to which players A and B can defeat the provocation by exchanging their own messages is not known.

Section 5 argued that a leader may behave aggressively in order to appease a hawkish constituency, and democratization is not necessarily good for peace. However, the model did not allow any communication. When strategies are strategic complements and conflict is caused by mutual fear, leaders may be able to create peace by exchanging messages, as discussed in Section 3. The problem is to make sure the messages are credible. Levy and Razin [21] argue that a democratic leader may be able to communicate more credibly than a dictator, because the democratically elected leader faces two audiences: his domestic constituency as well as the other leader.

In Section 5, political institutions were exogenously given. But war can trigger a change in political institutions. The Falklands War led to a change in the political leadership of Argentina and helped clear the path towards democracy. The Second Gulf War created a more democratic regime in Iraq. Some political institutions may be more fragile than others and more susceptible to civil war and regime change. A model where both political institutions and decisions to go to war are endogenously determined is left for future research.

In the spirit of Hobbes, we have emphasized models where actions are strategic complements, and fear of the opponent triggers conflict. In contrast,

if actions are strategic substitutes, fear of the opponent can deter aggression and prevent conflict. In practise, it can be difficult to distinguish problems of escalation from problems of deterrence. During the Cold War, some believed the main problem was to deter Soviet aggression, while others argued that the main problem was to prevent a fear-spiral. In Section 6, we considered a model where player B can show toughness by arming himself, but he does not know if toughness deters player A from attacking or makes player A more likely to attack out of fear. If arms inspections reveal that player B is not armed, then player A is less likely to attack out of fear, but more likely to attack out of greed as there is no deterrence. Player B's best option may be to deliberately create uncertainty in the mind of player A, i.e., to maintain "strategic ambiguity". Strategic ambiguity provides deterrence when player B is unarmed, reducing player B's incentives to accumulate weapons. Baliga and Sjöström [3] construct cheap-talk equilibria where arms inspections are triggered by messages sent by player A. As in Sections 3 and 4, the cheap-talk equilibria have a *non-convex* structure.

Arms inspections remove ambiguity about weapons capabilities. But nations sometimes maintain strategic or "constructive" ambiguity about *intentions* rather than capabilities. For example, the U.S. has maintained ambiguity about how it would respond to an attack on Taiwan from mainland China. Future research may clarify the logic of this type of ambiguity.

In this article, we have emphasized models where two opponents choose either hawkish or dovish actions. A different paradigm is a bargaining game, where players make demands, and a war may occur if bargaining breaks down (see Fearon [8]). The connections between the two strands are well worth studying.

References

[1] Sandeep Baliga, David Lucca and Tomas Sjöström (2011): "Domestic Political Survival and International Conflict: Is Democracy Good for Peace?" *Review of Economic Studies* **78**:458-486.

[2] Sandeep Baliga and Tomas Sjöström (2004): "Arms Races and Negotiations," *Review of Economic Studies* **17**:129-163.

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

[4] Sandeep Baliga and Tomas Sjöström (2009): "Conflict Games with Payoff Uncertainty," mimeo, Northwestern University.

[5] Sandeep Baliga and Tomas Sjöström (2010): "The Strategy of Manipulating Conflict," mimeo, Northwestern University.

[6] Sylvain Chassang and Gerard Padró i Miquel (2008): "Conflict and Deterrence under Strategic Risk," mimeo, Princeton University.

[7] Arthur Conan Doyle (1894): *The Memoirs of Sherlock Holmes* (London: George Newnes).

[8] James Fearon (1995): "Rationalist Explanations for War," *International Organization* **49**:379-414.

[9] James Fearon (1996): "Bargaining over Objects that Influence Future Bargaining Power," mimeo, Stanford University.

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

[11] Michelle Garfinkel and Stergios Skaperdas (2000): "Conflict without Misperceptions or Incomplete Information: How the Future Matters," *Journal of Conflict Resolution* **44**:793-807.

[12] Michael Gordon and Bernard Trainor (2006): *Cobra II: The Inside Story of the Invasion and Occupation of Iraq* (New York: Pantheon Books).

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

[14] Thomas Hobbes (1886): *Leviathan*, Second Edition (London: Ballantyne Press).

[15] Matthew Jackson and Massimo Morelli (2007): "Political Bias and War," *American Economic Review* **97**:1353-1373.

[16] Robert Jervis (1976): *Perception and Misperception in International Politics* (Princeton: Princeton University Press).

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

[18] Immanuel Kant. (1795, 1903): *Perpetual Peace: A Philosophical Essay*, translated by M. Campbell Smith (London: Swan Sonnenschein & Co.).

[19] Henry Kissinger (1994): *Diplomacy* (New York: Touchstone).

[20] Andrew Kydd (1997): "Game Theory and the Spiral Model," *World Politics* **49**:371–400.

[21] Gilat Levy and Ronny Razin (2004): "It takes Two: An Explanation for the Democratic Peace," Journal of the European Economic Association **2**:1-29.

[22] Paul Milgrom and John Roberts (1982): "Predation, Reputation, and Entry Deterrence," *Journal of Economic Theory* **27**:280-312.

[23] Hans Morgenthau (1967): *Politics Among Nations: The Struggle for Power and Peace*, Fourth Edition (New York: Alfred A. Knopf).

[24] Roger Myerson (2008): "Revelation Principle," in *The New Palgrave Dictionary of Economics* (London: Palgrave Macmillan).

[25] Thomas Paine (1985): *Rights of Man* (London: Viking Penguin Inc.).

[26] Robert Powell (2006): "War as a Commitment Problem," *International Organization* **60**:169-203.

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

[28] Thomas Tangerås (2009): "Democracy, Autocracy and the Likelihood of International Conflict," *Economics of Governance* **10**:99-117.

[29] Thucydides (1989): *The Peloponnesian War: the Complete Hobbes Translation*, Edition 1989 (Chicago: University of Chicago Press).

[30] Kevin Woods, Michael Pease, Mark Stout, Williamson Murray and James Lacey (2006): "Iraqi Perspectives Project," mimeo, Joint Center for Operational Analysis.