Review of Economic Studies (2004) **71**, 351–369 0034-6527/04/00150351$02.00

Arms Races and Negotiations

SANDEEP BALIGA
Northwestern University

and

TOMAS SJÖSTRÖM
Pennsylvania State University

First version received October 2001; *final version accepted January* 2003 (*Eds.*)

Two players simultaneously decide whether or not to acquire new weapons in an arms race game. Each player's type determines his propensity to arm. Types are private information, and are independently drawn from a continuous distribution. With probability close to one, the best outcome for each player is for neither to acquire new weapons (although each prefers to acquire new weapons if he thinks the opponent will). There is a small probability that a player is a dominant strategy type who always prefers to acquire new weapons. We find conditions under which the unique Bayesian–Nash equilibrium involves an arms race with probability one. However, if the probability that a player is a dominant strategy type is sufficiently small, then there is an equilibrium of the *cheap-talk extension* of the game where the probability of an arms race is close to zero.

1. INTRODUCTION

Many interactions contain a fundamental element of coordination, so multiple Nash equilibria exist. For example, a government may not desire an arms race, but prefers to acquire new weapons if it believes the government of a neighbouring state will acquire them. Then there may be two pure strategy Nash equilibria: an "arms race" equilibrium in which both governments acquire new weapons, and a "detente" equilibrium in which neither government acquires any. Similarly, a depositor may not need his money right away, but prefers to withdraw his money from the bank if he believes other depositors will do so. Again, two pure strategy Nash equilibria may exist: a "bank run" equilibrium in which all depositors try to withdraw their money, and a "liquid" equilibrium with no bank run. In both these examples, one Nash equilibrium is likely to Pareto dominate the other. Introducing uncertainty can allow us to select one of these equilibria. In the bank run example, each depositor may assign some very small probability to the event that the other depositor has suffered a liquidity shock and has a dominant strategy to withdraw money. In the arms race model, each government may assign a small but positive probability to the event that the opponent is a truly aggressive type for whom acquiring arms is a dominant strategy. This kind of uncertainty raises the benefit of taking a defensive action (withdrawing money or acquiring new weapons). But, since each player knows that his opponent is thinking this way, a *multiplier effect* appears, creating an escalating cycle of pessimistic expectations that spiral toward the Pareto inferior equilibrium. This type of argument was first stated clearly by Schelling (1960) in his classic book.

Schelling's insight has been applied to macroeconomic phenomena such as currency attacks and liquidity crises (see Morris and Shin, 2003). Our article differs from this literature in two ways. First, in our model there is no common shock to preferences. A player's type is his private benefit from taking a certain action, which is independent across players. Second, we study the role of cheap-talk. In particular, our model sheds some light on the strategic considerations that

underlie information transmission and coordination in arms negotiations. While it is convenient to present our model in terms of an arms race, the arguments would apply to any model where an underlying coordination game has been augmented with incomplete information about independently drawn types.

Suppose a player's true propensity to arm, *i.e.* his type, is his private information. Types are drawn independently from a continuous distribution. All types prefer the opponent to remain unarmed, whatever action they themselves take. Also, for all types, the worst possible outcome is to be unarmed while the opponent arms. However, the type determines whether or not the player prefers to arm when he is unsure about the actions of the opponent. At one end of the distribution are the aggressive *dominant strategy types* who prefer to arm regardless of the opponent's actions. At the other end of the distribution are the peaceful types who prefer to arm only if they are virtually sure that the opponent will arm. Let the fraction of dominant strategy types be some small $\varepsilon > 0$. These types will certainly arm, but this triggers a multiplier effect. Some fraction $\delta > 0$ of all types are not dominant strategy types but prefer to arm when the opponent arms with at least probability ε. These "almost dominant strategy types" will arm if they know that the dominant strategy types will do so. But then, all types that prefer to arm when the opponent arms with at least probability $\varepsilon + \delta$ will also arm, etc. The contagion takes hold. Even though each player thinks it is *extremely unlikely* that the opponent is a dominant strategy type, the *unique* Bayesian–Nash equilibrium may involve an arms race with probability one.

What can be done to escape this logic? One possibility is an "arms control treaty". However, in the absence of an enforcement authority (a "World Court") such treaties are pure cheap-talk. It is not obvious that cheap-talk can be effective in this kind of situation. Suppose, before making the decision to arm, each player can send a hawkish (aggressive) or a dovish (conciliatory) message to the opponent. Since messages are cheap-talk, a player is always free to arm himself regardless of what messages were sent. Since all types are better off if the opponent does not arm (whatever they themselves decide to do), one might suspect that all types would send the same message, namely, whatever message is most likely to persuade the opponent not to arm. Clearly, if all types send the same message then the talk is not informative and cannot prevent an arms race spiral. However, we show that in fact there are cheap-talk equilibria where the talk is informative, and the information that is generated is used to reduce the likelihood of an arms race. If the dominant strategy types are sufficiently rare, then the equilibrium probability of an arms race is close to zero. Thus, communication can have the dramatic effect of reducing the probability of an arms race from one to almost zero.[1]

The reasoning behind the informative cheap-talk equilibrium is subtle. Although all types want to reduce the probability that the opponent arms, players who are not dominant strategy types also want to resolve the uncertainty about the opponent's action in order to avoid a coordination failure. Since different types trade off these two objectives at different rates, it is possible to induce different types to send different messages. Consider a simplified example with only three types. There is a "very tough" type with a high propensity to arm. He will always arm himself, regardless of what messages are sent or received. There is "normal" type with a low propensity to arm. He will arm himself only if he is almost sure that the opponent will arm (and even in that case, his preference in favour of arming is not very strong). Finally, there is an intermediate "fairly tough" type with a medium propensity to arm. He prefers to take whatever action he thinks the opponent is most likely to take. Clearly, it is the fairly tough type who puts the highest value on resolving the uncertainty about the opponent's action. The very tough type and the normal type are mainly interested in reducing the probability that the opponent arms.

1. In any cheap-talk game, there are "babbling" equilibria where the parties simply disregard the messages. We focus on informative equilibria.

We construct a separating equilibrium as follows. Sending the conciliatory message "I am a dove" minimizes the probability that the opponent arms, but the uncertainty about the opponent's action is not resolved. Sending the aggressive message "I am a hawk" yields a higher probability that the opponent arms, but after the talk there is no ambiguity about the opponent's action. In equilibrium, the very tough type and the normal type send the dovish message in order to minimize the probability that the opponent arms, while the fairly tough type sends the hawkish message in order to minimize the probability of a coordination failure. The equilibrium is "non-monotonic" in the sense that the types with the highest and lowest cost pool, and the intermediate type separates out. In order to study contagions, we will work with a continuum of types, but the intuition for why cheap-talk works in our model is the same as in the discrete example.

Green and Stokey (1980) and Crawford and Sobel (1982) studied sender–receiver games where only player 1 (the sender) has private information and sends a message, and only player 2 (the receiver) takes an action. Let $t \in [0, 1]$ denote player 1's type, and let $p \in \mathbf{R}$ denote player 2's action. Crawford and Sobel assumed preferences satisfy a sorting condition: the best value of p, from either player's standpoint, is strictly increasing in t. Under this assumption, they showed that all equilibria are partition equilibria of a particular kind. Specifically, the set of types that send a particular message is always convex. Clearly, the sorting condition is required for this result. With general preferences, different kinds of equilibria may well exist (see Green and Stokey, 1980). In our model, types are ordered in a natural way according to their propensity to arm, but "non-monotonic" equilibria exist because *intermediate* types have the strongest preference for *coordinating on the same action as the opponent*. This reasoning depends on *both* players being allowed to talk and act (otherwise the issue of coordination is moot), so it has no counterpart in sender–receiver games of the kind studied by Green and Stokey (1980) and Crawford and Sobel (1982).

Rubinstein (1989) and Carlsson and van Damme (1993) show how introducing a small number of dominant strategy types into a coordination game can select a unique Bayesian–Nash equilibrium. In the Carlsson and van Damme (1993) model, the players make noisy observations of the true pay-off matrix. Signals (types) consist of a common shock plus a noise term. When the noise is small, the players' types are very highly correlated, and there is a unique equilibrium. In contrast, we assume types are independent. This is a reasonable assumption in many applications. In the bank-run model, liquidity shocks may be uncorrelated. In the arms race model, the costs and benefits from acquiring new weapons may depend on psychological, moral, economic and political considerations that are specific to a certain country or a certain leader. The enjoyment of the "prestige" of being the leader of a nuclear state, or the pay-off from appeasing an important constituency by building new weapons systems, are examples of benefits that would seem to be largely independent across countries. The cost of developing nuclear weapons will be much lower if a nation has access to technical expertise and fissile material from the former Soviet Union, but such access will depend on numerous idiosyncratic events. Finally, even if the *monetary* cost of building weapons is known to be the same in the two countries, what matters is the economic (opportunity) cost. The opportunity cost may have a large idiosyncratic component, since it depends on how much the leader values the pursuit of *other* goals (for example, alleviating poverty), as well as on his beliefs about the needs and resources of his country.

Harsanyi (1973) has shown that every Nash equilibrium of a complete information game is the limit of (pure strategy) Bayesian–Nash equilibria of any sequence of close-by incomplete information games with independent types. In contrast, we find a condition under which (in the absence of communication) only the Pareto-inefficient arms race is a Bayesian–Nash equilibrium outcome of the game with incomplete information, even as the fraction of dominant strategy types goes to zero. However, our uniqueness condition is satisfied only if there is sufficient uncertainty about the actual numerical pay-offs (as opposed to the ordinal ranking of the outcomes). Thus, a

unique equilibrium can be the result of either sufficient uncertainty about independent types, or sufficient correlation of types (for a discussion, see Morris and Shin, 2002*b*).

A large literature on sender–receiver games followed Green and Stokey (1980) and Crawford and Sobel (1982). Unlike most articles in that literature, in our model *both* players have private information, talk and act. There is a small set of models with a related structure. Baliga and Morris (2002) discuss an example of cheap-talk with two-sided incomplete information, Matthews and Postlewaite (1989) consider double auctions, and Austen-Smith (1990) considers legislators with private information taking part in a debate before they vote. Most closely related is Banks and Calvert's (1992) model of a battle-of-the-sexes game with two-sided incomplete information. In their model, each player has his own favourite outcome, but the intensity of his preference is determined by his type ("high" or "low"). There is no dominant strategy type and no multiplier effect. Without communication there are efficient asymmetric Bayesian–Nash equilibria where one player always gets his favourite outcome. However, the only *symmetric* equilibrium involves a randomization which is inefficient. Banks and Calvert show how communication can produce a symmetric equilibrium where coordination occurs more frequently, and player i's favourite action is more likely to be chosen when the intensity of his preference is high. In the battle-of-the-sexes game, player i always wants player j to take the same action as player i, while in our arms race game player i always wants player j not to arm. Thus, the nature of communication is different in the two cases. Indeed, the true pay-off matrix in our arms race game is with high probability a *stag hunt game*. Aumann (1990) argues that communication is particularly difficult in such a game, as each player wants his opponent to take the same action whatever action he himself takes (in an arms race, each player always wants the opponent to stop building weapons).

The article by Banks and Calvert (1992) is particularly interesting because they show that a *mediator* may be necessary for efficiency. In our model, we can approximate first-best efficiency without a mediator when the dominant strategy types are rare. The issue of whether a mediator is useful when the dominant strategy types are not rare is an interesting topic for future work.

This paper is organized as follows. Section 2 discusses previous models of arms races and some historical events. Section 3 presents the basic model without communication. A *multiplier condition* on the distribution of types is shown to be necessary and sufficient for an arms race to occur with probability one even if the dominant strategy types are very rare. Section 4 shows how cheap-talk reduces the probability of an arms race to almost zero when the dominant strategy types are very rare. Section 5 concludes. Technical calculations are contained in the Appendix.

2. ARMS RACES IN THEORY AND PRACTICE

There is in practice not much distinction between offensive and defensive weapons (Schelling (1960), Jervis (1976)). Even if a country arms for defensive purposes, these armaments will make other countries feel less secure. Therefore, the suspicion that a country may arm, *for whatever purpose*, makes other countries more likely to arm in self-defence, creating an arms race spiral which makes everyone worse off. This is the well-known *security dilemma* or *spiralling model* (Jervis, 1976, 1978). Jervis (1978) based a formal discussion of this dilemma on a stag hunt game, where each state thinks the best possible outcome is for nobody to arm, but each prefers to arm if they think the opponent will. Jervis argued that it is *irrationality* that drives the arms race spiral: "if the spiral theory is correct, it is so partly because the actors do not understand it or follow its prescriptions" (Jervis, 1978, p. 81). Indeed, if it is common knowledge that the pay-offs are those of a stag hunt game, it is not clear why rational players could not refrain from an arms build-up. Kydd (1997) added incomplete information about the opponent's preferences. In his model there are "greedy" states who want war and "security seekers" who want peace. The discrete

type space means there is no contagion in our sense (if greedy types are sufficiently rare then there is no reason for security seekers to arm). Kydd (1997) discussed how initial armaments can signal a player's type in a multi-period game. We show how, with a continuous type space, arms races can be triggered by an arbitrarily small probability that a player is "greedy", and we show how cheap-talk can signal a player's type. Schelling (1960) developed a formal model where players may attack each other inadvertently because of a "false alarm". Knowing that the opponent may inadvertently attack, each will be more likely to attack, triggering a contagion. If the underlying problem is an imperfect warning system, then cheap-talk cannot be the solution. We show that cheap-talk can be useful when the underlying problem is incomplete information about the opponent's preferences.

History provides many examples where fear and distrust, sparked by uncertainty about the opponent's motives, appear to have triggered arms races. For example, it is often argued that the arms race spiral that preceded World War I was caused by Britain's and Germany's mutual distrust of each other, rather than any nation's desire to fight a war (Sontag (1933), Wainstein (1971)). Similar mechanisms may have operated during the cold war (Leffler, 1992). The India–Pakistan arms race is a contemporary example of escalation fuelled by mutual distrust.[2]

Analogous mechanisms may underlie *war initiations* as well. Thucydides (1972, Book I, 23) claimed that "the growth of Athenian power and the fear which this caused in Sparta" made the Peloponnesian War inevitable.[3] Rousseau (quoted by Jervis, 1976, p. 63) argued that "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 neighbour, who would not fail to forestall the attack in turn at any moment favourable to himself, so that many 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". Jervis (1976, p. 94) argues that war broke out in 1914 because "each of the continental powers believed that the side that struck first would gain a major military advantage. Since to wait for the other side to clarify its intentions could mean defeat, even a country that preferred the status-quo to a war would feel great pressures to attack". This suggests that one can interpret our model as a model of war initiations, with each nation simultaneously deciding whether to attack or to hold back. However, if the perceived first-strike advantage is large, then the true pay-off matrix is more likely to be a prisoners' dilemma than a stag hunt game. Spiral theorists have argued that the stag hunt game and the prisoner's dilemma tend to generate equally bad outcomes (Jervis, 1976, p. 67). In our model, the true pay-off matrix is a stag hunt game with high probability, and we show that the outcome will be bad for everyone if a "multiplier condition" is satisfied. However, there is an important difference between our version of the security dilemma, where it is highly unlikely that a player has a dominant strategy to attack, and a prisoner's dilemma, where it is known that each player has a dominant strategy to attack. The distinction is that pre-play communication is useful in the former but not the latter case. Our results on the value of communication are irrelevant to the case of war initiations if

2. "Pakistan does not intend to aggress…[W]e are the victim of (Indian) aggressions" (Foreign Minister Gohar Ayub Khan quoted by the Pakistan News Service, June 1999). "In India, one often hears that 'Pakistan understands' that India has no hostile designs on it.… In Pakistan, however, there is strong sense that the nation's survival is potentially at risk in the event of a major Indian attack. Without a clearer understanding of India's defence doctrine, this could generate a catastrophic miscalculation" (CSIS South Asia Monitor, 1 February, 1999).

3. A famous passage describes how the Spartans are spurred on by the Corinthians: "You Spartans are the only people in Hellas who wait calmly on events, relying on your defence not on action but on making people think you will act. You alone do nothing in the early stages to prevent an enemy's expansion; you wait till the enemy has doubled his strength. Certainly you used to have the reputation of being safe and sure enough; now one wonders if this reputation was deserved.… The Athenians…live close to you, yet you still do not appear to notice them; instead of going out to meet them, you prefer to stand still and wait till you are attacked, thus hazarding everything by fighting with opponents who have grown far stronger than they were originally" (Thucydides, 1972, Book I, 69).

the perceived first-mover advantage is large. If a player is inclined to strike even if he thinks his opponent will not, then the problem becomes one of *deterrence* rather than a security dilemma.[4]

History shows that negotiations have a mixed record at preventing arms races. The British attempt to prevent an arms race in Dreadnought warships with Germany in 1912 may be described as a *sincere dovish strategy*, that is, a dovish attitude with an intent of arming only if the opponent does *not* appear dovish. The British felt "it might be possible by friendly, sincere and intimate conversation to avert this perilous development" and that "surely something could be done to break the chain of blind causation" (Churchill, 1931, p. 75). Churchill proposed a "naval holiday" for 1913, but the Germans took a hawkish stance, thereby triggering an arms race. In 1935, the British strategy was not very different from the one used in 1912, but this time the Germans seemed more accommodating. A naval accord limited the German fleet to 35% of the British. The British worried about the fact that Hitler's true military strength was difficult to assess, but they felt that trusting Hitler was a chance worth taking (Kissinger, 1994, pp. 295–296). In fact, Hitler used an *insincere dovish strategy*, signing treaties he did not intend to respect.[5] In 1940, Goebbels explained how Hitler had lured the western powers into a false sense of security: "up to now we have succeeded in leaving the enemy in the dark concerning Germany's real goals.... They left us alone and let us slip through the risky zone, and we were able to sail around all dangerous reefs. *And when we were done, and well armed, better than they, then they started the war!*" (Kissinger, 1994, p. 295). On the other hand, the successful test ban treaty signed by the United States and the Soviet Union in 1963 may have been part of a sincere dovish strategy on *both* parts. Finally, the *hawkish strategy* of Ronald Reagan, who called the Soviet Union an "evil empire" prepared "to commit any crime, to lie, to steal" to achieve its goals (Reagan, 1983), did not prevent him from eventually concluding very successful arms control talks, once he had become convinced that the U.S. would not be taken advantage of. These historical examples illustrate the possibility that different "types" use different negotiating tactics. In our cheap-talk equilibrium, very tough, normal and fairly tough types use different strategies which may be described as insincere doves, sincere doves, and hawks, respectively.

3. THE ARMS RACE GAME

Two players must simultaneously and independently decide whether or not to invest in a new weapons programme. The possible choices are *Build new weapons* (B) or *No new weapons* (N). We normalize the pay-off to zero for each player if both choose N. A player who chooses N while the other player chooses B suffers a loss of $d > 0$, which represents the disutility of having a less advanced weapons system than the opponent. Presumably, d could be quite large. A player who acquires new weapons never has to suffer this cost, as he will always be at least as strong as his opponent. However, he has to pay the cost of the new weapons. Let player i's cost of acquiring new weapons be denoted $c_i \geq 0$. This could be a psychological or a monetary cost. A player who builds the new weapons system while his opponent does not receives a gain of $\mu > 0$, which represents the value of having a more advanced weapons system than the opponent.

4. During the cold war, two theories were put forward. Some argued that wars are caused by misunderstandings and mutual distrust ("the security dilemma"). Therefore, the U.S. ought to be friendly and kind to the Soviet Union in order to prevent any misunderstandings about American motives. But others argued that wars occur because an aggressor has not been sufficiently deterred ("deterrence theory"). This led to the opposite recommendation, *i.e.* a policy of toughness ("deterrence") against the Soviet Union. Our paper cannot contribute to this debate as we are *assuming* the first hypothesis (wars are likely to be caused by mutual distrust). Deterrence theorists may in fact argue that arms races can *prevent* wars, by making a war so costly that the opponent is deterred from starting it. A dynamic model along the lines of Kydd (1997), where players first decide whether or not to arm, and then whether or not to go to war, might address these issues, but it is beyond the scope of this paper.

5. Hitler's abrogation of the naval accord is described by Craig (1978, pp. 686–710). Other dovish messages sent by Hitler included the signing of the German–Polish non-aggression pact.

We shall be mainly interested in the case where μ is small, so that the temptation to build new weapons is not too big (our interest is in coordination problems, not prisoner's dilemmas). Player i's pay-offs can be represented in a pay-off matrix as follows (player i chooses a row, player j a column):

	B	N
B	$-c_i$	$\mu - c_i$
N	$-d$	0

(1)

If $d > c_i > \mu$ for each $i \in \{1, 2\}$, then each player thinks the *best* possible outcome is for *neither* player to build new weapons, but the *worst* possible situation is to refrain from building while the opponent builds. If these preferences are common knowledge, then there are two pure strategy Nash equilibria: (B, B) and (N, N). The game is a stag hunt game, as discussed by Jervis (1978), not a prisoner's dilemma. Knowing this, rational players should be able to coordinate on the Pareto dominant equilibrium (N, N), perhaps after communicating with each other (O'Neill, 1999). However, from now on we will assume c_i is player i's *private information*.[6]

We refer to c_i as player i's *type*. Each player i knows his own type c_i, but not the other player's type c_j. The types c_1 and c_2 are independently drawn from the same distribution, with continuous cumulative distribution function F. F has support $[0, \bar{c}]$ with $F(0) = 0$, $F'(c) > 0$ whenever $0 < c < \bar{c}$, and $F(\bar{c}) = 1$. We assume $\bar{c} < d$. Everything except the true c_1 and c_2 is common knowledge.

Since $-c_i \geq -\bar{c} > -d$, B is always a (strict) best response against B. Therefore, there is a Bayesian–Nash equilibrium where all types choose B with probability one. Is there any other Bayesian–Nash equilibrium? Notice that N is a best response against N for player i if and only if $c_i \geq \mu$. If $c_i < \mu$ then player i is a *dominant strategy type*: B is a strictly dominant strategy for him. The probability that player i is a dominant strategy type is $F(\mu)$, which is close to zero if μ is small. However, the existence of dominant strategy types may have a large effect on the set of equilibria even as $\mu \to 0$.

Any equilibrium will have a cut-off property: if type c_i builds new weapons, then any type $c_i' < c_i$ will also build. The dominant strategy types certainly play B. Knowing that the opponent (who might be a dominant strategy type) plays B with strictly positive probability, a type that is "almost" a dominant strategy type ($\mu - c_i$ negative but close to zero) will also play B. This "infects" other types with slightly higher cost, who also decide to play B, and so on. Now, if this contagion stops before all types have been infected, then there must be some cut-off type $c_i^* > 0$ such that all types with a lower cost than him play B and all types with a higher cost play N. Type c_i^* himself must be indifferent. Suppose for the moment that the equilibrium is symmetric, so $c_1^* = c_2^* = c^*$. The condition that player i's type c^* is indifferent between B and N when player j is expected to choose B with probability $F(c^*)$ is $S(c^*) = 0$, where

$$S(c) \equiv F(c)(d - c) + (1 - F(c))(\mu - c). \tag{2}$$

This leads to the following definition.

Definition 1. The distribution satisfies the *multiplier condition* if $F(c)d \geq c$ for all $c \in [0, \bar{c}]$.

6. Carlsson and van Damme (1993) assume each player observes the true pay-off matrix with noise, and find a condition under which each player will choose the action which is *risk dominant in the game that he observes*. In contrast, we assume each player receives no signal about the opponent's pay-off function, and risk dominance plays no role. (In our game, (B, B) is risk dominant, in the sense of Carlsson and van Damme, in the true pay-off matrix if and only if $d + \mu > c_1 + c_2$.)

If the multiplier condition is satisfied then $S(c) > 0$ for any $\mu > 0$ and any $c \geq 0$, so there can be no symmetric Bayesian–Nash equilibrium where N is chosen with positive probability. The multiplier condition guarantees that each type will strictly prefer to choose B whenever he thinks all types with lower cost than him will choose B, so the contagion to play B will infect the whole population. The proof of Theorem 1 shows that when the multiplier condition holds there are also no asymmetric equilibria where N is chosen with positive probability. Thus, there is an arms race with probability one, confirming Schelling's vicious cycle argument.

Notice that $F(0)d = 0$ and $F(\bar{c})d = d > \bar{c}$ by assumption. Diagrammatically, the multiplier condition says that the graph of F lies on or above a ray through the origin with slope $1/d$. The uniform distribution, $F(c) = c/\bar{c}$, satisfies the multiplier condition because $cd/\bar{c} \geq c$ for all $c \geq 0$.[7] More generally, the multiplier condition is satisfied if F is concave, because concavity implies $F(c) \geq c/\bar{c} \geq c/d$ for all $c \geq 0$.

If the multiplier condition is violated, then for sufficiently small μ there exists $c^* < \bar{c}$ such that $S(c^*) = 0$. That means type c^* is just indifferent between building and not building if he thinks his opponent plays B with probability $F(c^*)$. Therefore, there exists a symmetric Bayesian–Nash equilibrium where N is chosen with positive probability. Thus, for small μ the multiplier condition is necessary as well as sufficient for the contagion to play B to infect the whole population. The multiplier condition is violated if F is sufficiently *convex*. Intuitively, convexity implies that low-cost types are relatively rare. In the convex case, a player with a relatively high cost of arming may not want to arm even if he thinks all types with a lower cost than him will arm, simply because meeting such a low-cost type is unlikely. This stops the contagion to play B from infecting the whole population.

We now state the formal result.

Theorem 1. (i) *If the multiplier condition is satisfied, then for any $\mu > 0$ there is a unique Bayesian–Nash equilibrium. In this equilibrium all players choose B, regardless of type.* (ii) *If the multiplier condition is violated, then for sufficiently small $\mu > 0$ there exists a Bayesian–Nash equilibrium where N is chosen with strictly positive probability.*

Proof. If $\mu \geq d$ then B is a dominant strategy for all types, so the analysis is trivial. Suppose instead that $0 < \mu < d$. First, we establish the cut-off property: if B is a weak best response for type c_i then it is a strict best response for type $c_i' < c_i$. Indeed, if player i thinks player j will choose B with probability p_j, the pay-off to player i from B is

$$p_j(-c_i) + (1 - p_j)(\mu - c_i) = (1 - p_j)\mu - c_i$$

while the pay-off from N is $p_j(-d) + (1 - p_j) \times 0$. Type c_i weakly prefers B if and only if

$$c_i \leq (1 - p_j)\mu + p_j d. \tag{3}$$

Notice that all of player i's types have the same beliefs about player j, since types are assumed to be uncorrelated. If type c_i weakly prefers B, then inequality (3) is strict for type $c_i' < c_i$ so type c_i' strictly prefers B. Now we can prove the two parts of the theorem.

(i) If player j chooses B with probability one, then all of player i's types will choose B since $\bar{c} < d$. Therefore, there is always an equilibrium where all players choose B, regardless of type. Suppose in addition there is an equilibrium where N is played with positive probability. We claim the multiplier condition is violated. If a player chooses N

7. A sufficient condition for the multiplier condition to hold is for $c/F(c)$ to be non-decreasing. This is true in the uniform case.

then he must expect the opponent to choose N with strictly positive probability, hence if N is chosen with positive probability by one player then *both* players must choose N with positive probability. For $i \in \{1, 2\}$ let c_i^* be such that B is a weak best response for player i at the equilibrium if and only if his type satisfies $c_i \leq c_i^*$. By hypothesis, $c_i^* < \bar{c}$, for otherwise player i chooses N with probability zero. The probability that player $i \in \{1, 2\}$ chooses B is $p_i = F(c_i^*)$. Since type c_i^* must be indifferent between B and N,

$$c_i^* = (1 - p_j)\mu + p_j d = (1 - F(c_j^*))\mu + F(c_j^*)d.$$

Without loss of generality, suppose $c_1^* \leq c_2^*$. Then

$$c_1^* = (1 - F(c_2^*))\mu + F(c_2^*)d \geq (1 - F(c_1^*))\mu + F(c_1^*)d > F(c_1^*)d$$

since $0 < \mu < d$ and $F(c_1^*) \leq F(c_2^*) < 1$. Thus, the multiplier condition is violated.

(ii) Suppose the multiplier condition is violated. Then, there exists c' such that $c' > dF(c')$. For sufficiently small $\mu > 0$, we have $S(c') \leq 0$ where S is defined by (2). Also, $S(\bar{c}) = d - \bar{c} > 0$. By continuity, there is $c^* < \bar{c}$ such that $S(c^*) = 0$. Let each player i choose B if and only if $c_i \leq c^*$. Since $S(c^*) = 0$, by the construction of S it follows that type c^* is indifferent between B and N. Type $c_i < c^*$ strictly prefers B and type $c_i > c_i^*$ strictly prefers N. Thus, these strategies form a Bayesian–Nash equilibrium. ‖

A few remarks can be made. First, the proof of Theorem 1 shows that when the multiplier condition holds, the game is interim dominance solvable. After iterated elimination of strongly (interim) dominated strategies only the "arms race" outcome remains. Second, the arms race outcome is inefficient because all types prefer NN to BB. Third, the arms race is caused by *mutual* distrust. No contagion can occur if one player's preferences are common knowledge.[8]

Harsanyi (1973) has shown that every Nash equilibrium of a complete information game is the limit of (pure strategy) Bayesian–Nash equilibria of any sequence of close-by incomplete information games with independent types. Consider a complete information version of the arms race game with $c_1 = c_2 = c^*$. If $\mu < c^*$ then there are two pure strategy Nash equilibria, (B, B) and (N, N). Now consider a nearby game of incomplete information, where the probability that a player's type belongs to the neighbourhood $(c^* - \delta, c^* + \delta)$ is $1 - \varepsilon$. Suppose $\delta > 0$ and $\varepsilon > 0$ are small enough so that $d\varepsilon < c^* - \delta$. Then, for $c' \in (d\varepsilon, c^* - \delta)$ we have $F(c') \leq \varepsilon < c'/d$. Thus, the multiplier condition (which guarantees uniqueness of equilibrium) is violated with a small amount of uncertainty and independent types. Conversely, the uncertainty is maximized when types are uniformly distributed, in which case the multiplier condition is satisfied. Thus, sufficient uncertainty generates uniqueness *even if* types are independent, and even though it is common knowledge that the true pay-off function is that of a stag hunt game with a very high probability. Morris and Shin (2002*b*) have recently clarified the relationship between common shocks, independent types, noise and multiplicity. They show that in a general model of correlated types, uniqueness can be obtained by assuming *either* sufficient uncertainty (as in our model) *or* sufficient correlation of types (as in Carlsson and van Damme, 1993). Their results suggest that adding a common shock will not change our uniqueness result, as long as a sufficient amount of heterogeneity remains after the common shock is accounted for.

Our assumption that the pay-off from choosing N is independent of type is without loss of generality, because all that matters is the difference between the pay-off from choosing B and the

8. Suppose that c_2 (but not c_1) becomes common knowledge as soon as the types are determined by nature. Then there would exist a Bayesian–Nash equilibrium where player 2 as well as all the non-dominant strategy types of player 1 choose N whenever $c_2 \geq (1 - F(\mu))\mu + F(\mu)d$, which happens with probability close to 1 for μ small enough. If both c_1 and c_2 become common knowledge as soon as they are determined by nature, then there is an equilibrium where both players choose N whenever $c_1 \geq \mu$ and $c_2 \geq \mu$, which again happens with probability close to 1 for μ small enough.

pay-off from choosing N. It is also without loss of generality to assume the pay-offs are linear in the type: if the true cost was some increasing function $h(c_i)$ we would simply define the type to be $h(c_i)$. The assumption that the type matters equally much when the opponent chooses N as when he chooses B does involve a loss of generality. An interesting topic for future research is to study more general pay-off matrices.

What is important for us is that some privately observed parameter influences a player's propensity to arm. Although we have talked about the *cost* of acquiring new weapons, the private information could equally well relate to a private *benefit* from arming. Suppose the monetary cost of acquiring weapons is commonly known to be $\bar{c} > 0$ for all types. Player i's true cost of arming is $c_i = \bar{c} - b_i$ where b_i is the private benefit that player i derives from acquiring weapons (there can be a private cost component, which is subtracted from b_i). The true value of b_i is known only to player i. If we assume b_1 and b_2 are independent random variables with support $[0, \bar{c}]$, then the model is formally equivalent to the one described above.

4. CHEAP-TALK

From now on we will assume the multiplier condition holds. Without cheap-talk, there is a discontinuity in the equilibrium correspondence: if μ were zero then there would be an equilibrium where all types choose N, but $\mu > 0$ implies that a contagion is started by the fraction $F(\mu) > 0$ of dominant strategy types. The game unravels and everybody plays B, as shown in Theorem 1. In this section, we will show that adding cheap-talk restores continuity in the sense that for small enough $\mu > 0$ there exists an equilibrium where almost all types choose N.

In the cheap-talk extension of the arms race game there are three stages. In stage zero, nature determines c_1 and c_2, and c_i becomes player i's private information. In stage one, messages are announced simultaneously and publicly. The two messages that are sent in equilibrium will be labelled Dove and Hawk. "Dove" is interpreted as a conciliatory message and "Hawk" is interpreted as an aggressive message. In stage two, the players simultaneously choose either B or N, and player i's pay-off is determined by his pay-off matrix, as in (1). The messages sent in stage one do not influence the pay-offs directly, but they may convey information about what the players plan to do in the future. Our main theorem states that arms races can be avoided almost surely if the fraction of dominant strategy types $F(\mu)$ is sufficiently small, that is, if $\mu > 0$ is sufficiently small.

Credible communication is difficult to achieve because, no matter what player i's true type is and whatever he himself plans to do, he always strictly prefers the opponent to choose N (because $\mu > 0$ and $d > 0$). To see the difficulty, let us try to construct the simplest possible non-trivial cheap-talk equilibrium: there is $\hat{c} \in (0, \bar{c})$ such that player i says Dove if $c_i > \hat{c}$ and Hawk if $c_i < \hat{c}$. But the hawkish message will reveal that player i has a relatively high propensity to arm, and the opponent will be inclined to act on this information by arming in self-defence. But then, types with a high propensity to arm will want to announce Dove unless this message also does not prevent an arms race. In fact, as we show in the Appendix (Theorem A1), an arms race must follow with probability one regardless of what player i says. Thus, this kind of simple equilibrium is not the solution to our problem. What we need is a slightly more complicated kind of equilibrium, where players with a very high propensity to arm never reveal their true nature, but instead behave just like types with a very low propensity to arm. Such "non-monotonic" equilibria can be constructed by exploiting the fact that it is the types with intermediate propensity to arm who care most about coordination.

Informally, our equilibrium works as follows. The type space $[0, \bar{c}]$ is partitioned into three sets: "very tough", "fairly tough", and "normal". The very tough types have the lowest cost of arming. In particular, the dominant strategy types are very tough. The fairly tough types have a

slightly higher cost. They are not dominant strategy types, so they are willing to play N if they think the opponent will. The normal types have the highest cost, and are the ones least willing to arm. In stage one, fairly tough types say Hawk. Normal and very tough types say Dove. If both players say Hawk, then neither builds weapons in stage two. (The hawkish message reveals that both players are fairly tough, but not tough enough to be dominant strategy types.) If both players say Dove, then a player who is very tough will build new weapons in stage two, while normal types will not. Finally, if one player says Hawk and the other says Dove, then both players build in stage two. The partitioning is such that if μ is small then the fraction of normal types will be close to one. Therefore, the probability will be close to one that both players say Dove and then refrain from building new weapons.

Disregard for the moment the issue of incentives in stage one. Suppose an exchange of dovish messages convinces each player that the opponent is *either* normal *or* very tough. Why is it now part of an equilibrium for normal types to choose N? The multiplier condition is still the key, but now it must be applied to the cumulative distribution over types *conditional on two dovish messages.* This conditional cumulative distribution function has a horizontal part on the interval of fairly tough types. Graphically, if the horizontal part crosses the ray through the origin with slope $1/d$, then the multiplier condition is violated and the contagion is blocked. The separation of fairly tough types from the other types generates a convex part of the conditional cumulative distribution function, which is what is needed to prevent the contagion from taking hold. Intuitively, because some types who have a lower cost of arming than the normal types are eliminated, the conditional distribution becomes more conducive to cooperation among the normal types. Of course, occasionally a normal type will meet a very tough type, in which case the normal type's realized pay-off will be $-d$. Still, the normal types are willing to trust an opponent who claims to be a dove as long as very tough types are *sufficiently rare*. This is guaranteed by the fact that the multiplier condition is violated. Most of the time, an exchange of dovish messages will be followed by peaceful coexistence.

In the previous paragraph, we discussed stage two without regard for the incentives in stage one, as if the messages had been exogenously generated public signals. In fact, the messages are endogenously generated by the players. We need to show that a player prefers to say Hawk if and only if he is fairly tough. Intuitively, it works as follows. First, the normal types do not want to deviate from their "sincere dovish strategy" in stage one, because saying Hawk will get them involved in arms races more often. Second, the very tough types do not want to deviate from their "insincere dovish strategy" in stage one either. They want to minimize the probability that the opponent arms, and this is done by masquerading as doves. This allows them to arm unilaterally against the normal types, who cannot tell a very tough opponent from a normal one. Finally, consider the fairly tough types. They are rewarded for saying Hawk in stage one by a guarantee that they will always coordinate with the opponent: they coordinate on (N, N) with other fairly tough types, and on (B, B) with everyone else. Suppose a fairly tough type deviates by saying Dove. In stage two he can either behave like a normal type and choose N, or like a very tough type and choose B. The first option is not attractive because his propensity to arm is high enough that he does not like the gambles normal types take. The second option is not attractive either, because it implies arming against everyone. By saying Hawk, he at least gets to coordinate on (N, N) with other fairly tough types.

We now state our main theorem. It will be proved under the regularity assumption that F has an infinite Taylor series representation.

Theorem 2. *Suppose the multiplier condition is satisfied. For any $\delta > 0$ there is $\bar{\mu} > 0$ such that if $0 < \mu < \bar{\mu}$ then there is a perfect Bayesian equilibrium of the cheap-talk extension of the arms race game where N is played with at least probability $1 - \delta$.*

In the remainder of this section we prove this theorem. We need the following lemma, which is proved in the Appendix.

Lemma 1. *Suppose the multiplier condition is satisfied. For sufficiently small $\mu > 0$, there exists a triple (c^L, c^*, c^H) such that*

$$\mu < c^L < c^* < c^H < \bar{c} \tag{4}$$

$$[F(c^H) - F(c^L)]c^L = (1 - F(c^H))\mu \tag{5}$$

$$[1 - 2(F(c^H) - F(c^L))]c^H = F(c^L)d \tag{6}$$

$$(1 - F(c^H))(\mu - c^*) + F(c^L)(-c^*) = F(c^L)(-d). \tag{7}$$

If $\mu \to 0$ then $c^H \to 0$.

Let (c^L, c^*, c^H) be as defined by Lemma 1 and consider the following strategies in the cheap-talk extension of the arms race game. Player i is *normal* if $c_i > c^H$, *fairly tough* if $c^L \leq c_i \leq c^H$, and *very tough* if $c_i < c^L$. In stage 1, the cheap-talk stage, player i says Hawk if he is fairly tough. Otherwise, he says Dove. (Players are allowed to say something else than "Hawk" or "Dove", but this will not happen in equilibrium.) In stage 2, the arms race stage, player i behaves as follows. If $c_i \leq \mu$ then he chooses B no matter what announcements were made in stage one. If $c_i > \mu$ then player i plays as follows: (i) if both players said Hawk then player i chooses N; (ii) if one player said Dove and the other said Hawk then player i chooses B; (iii) if both players said Dove then player i chooses N if and only if $c_i \geq c^*$; (iv) if any message except Dove or Hawk was sent by any player then player i chooses B.

Intuitively, equation (5) defines a type c^L who is indifferent between saying Dove and following an *insincere dovish strategy*, and saying Hawk and following a *hawkish strategy*; equation (6) defines a type c^H who is indifferent between saying Dove and following a *sincere dovish strategy,* and saying Hawk and following a *hawkish strategy*; and equation (7) defines a type c^* who is indifferent between playing B and N when both players have announced Dove.

We describe the *equilibrium* announcements made in stage 1, and the actions played *on the equilibrium path* in stage 2, in the table below. For example, if player 1 is very tough ($c_1 < c^L$) and player 2 is fairly tough ($c^L \leq c_2 \leq c^H$), then player 1 says Dove and player 2 says Hawk. Both players proceed to choose B, *i.e.* there is an arms race.

	$c_2 < c^L$ (Dove)	$c^L \leq c_2 \leq c^H$ (Hawk)	$c_2 > c^H$ (Dove)
$c_1 < c^L$ (Dove)	BB	BB	BN
$c^L \leq c_1 \leq c^H$ (Hawk)	BB	NN	BB
$c_1 > c^H$ (Dove)	NB	BB	NN

We claim that for small enough $\mu > 0$, these strategies form a perfect Bayesian equilibrium of the cheap-talk extension of the arms race game. Notice for future reference that (5) and the fact that $\mu < c^L$ implies that there are more normal types than fairly tough types:

$$1 - F(c^H) > F(c^H) - F(c^L). \tag{8}$$

In fact the R.H.S. of (8) will be close to zero and the L.H.S. close to one for μ small (for then both c^L and c^H will be close to zero).

Lemma 2. *The strategies specified above are sequentially rational in the action stage for all types, following all messages.*

Proof. If $c_i \leq \mu$, then it is clearly in player i's interest to choose B no matter what happened in stage 1. Suppose $c_i > \mu$. If both players announced Hawk in stage 1, then the opponent is expected to choose N in stage 2, and N is a best response against N. If one player said Dove and the other Hawk, or someone said something else than "Hawk" or "Dove", then the opponent is expected to choose B in stage 2, and B is a best response against B. Finally, suppose both players said Dove in stage 1. In this case, player i thinks his opponent is either a normal type who will choose N, or a very tough type who will choose B (recall that fairly tough types are the only ones who say Hawk in equilibrium). Now there are $1 - F(c^H)$ normal types and $F(c^L)$ very tough types. Then (7) implies that player i is indifferent between B and N if he is of type $c_i = c^*$. Clearly it is a best response for player i to choose B if $c_i < c^*$ and N if $c_i \geq c^*$. ‖

We now turn to the cheap-talk stage. Notice that for any type the expected pay-off from following the strategies specified above is greater than what he gets from playing BB for sure. Hence, no type has an incentive to send any message other than Hawk or Dove.

Lemma 3. *Player i prefers to say Dove if $c_i \leq \mu$.*

Proof. The dominant strategy type will go on to choose B for sure, so his objective is simply to maximize the probability of his opponent choosing N. If player i says Hawk, his opponent will choose N if and only if the opponent is a fairly tough type (who says Hawk according to his equilibrium strategy), an event which occurs with probability $F(c^H) - F(c^L)$. If player i says Dove, his opponent will choose N if and only if the opponent is a normal type (who says Dove according to his equilibrium strategy), an event which occurs with probability $1 - F(c^H)$. By (8), the dominant strategy type prefers to say Dove. ‖

Lemma 4. *Player i prefers to say Dove if $\mu < c_i < c^L$ and Hawk if $c^L \leq c_i < c^*$.*

Proof. Suppose $\mu < c_i < c^*$. First, suppose player i says Hawk. Then if the other player also says Hawk both will choose N. This event occurs with probability $F(c^H) - F(c^L)$. Otherwise, both choose B. The expected pay-off to player i is

$$[1 - (F(c^H) - F(c^L))](-c_i). \tag{9}$$

Suppose instead player i says Dove. Since $c_i < c^*$ player i will then choose B at the action stage whatever his opponent (player j) has said. Player j will choose N if and only if he is a normal type (who says Dove according to his equilibrium strategy), an event which occurs with probability $1 - F(c^H)$. Therefore, player i's expected pay-off from saying Dove is

$$(1 - F(c^H))(\mu - c_i) + F(c^H)(-c_i). \tag{10}$$

But, equation (5) implies that (9) equals (10) if $c_i = c^L$ so player i of type c^L is indifferent between Dove and Hawk. If $c_i > c^L$, (10) is smaller than (9) so player i prefers to say Hawk. If $\mu < c_i < c^L$, (10) is bigger than (9) so player i prefers to say Dove in this case. ‖

Lemma 5. *Player i prefers to say Hawk if $c^* \leq c_i \leq c^H$ and Dove if $c_i > c^H$.*

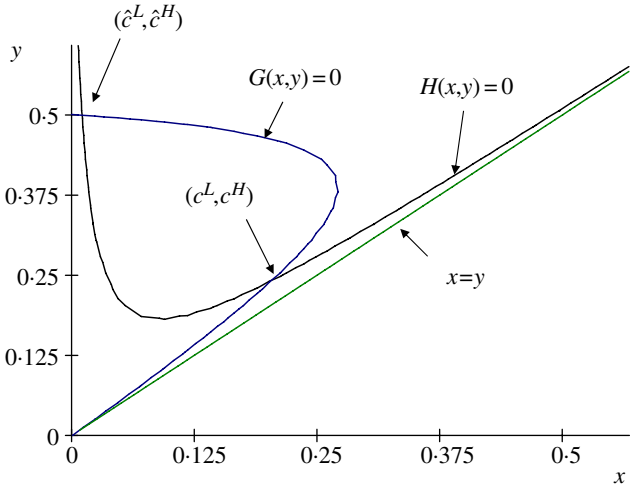

FIGURE 1
Constructing the equilibrium

Proof. Suppose $c_i \geq c^*$. First, suppose player i says Hawk. If the opponent is a fairly tough type (who says Hawk according to his equilibrium strategy) then both will choose N in the action stage. This event occurs with probability $F(c^H) - F(c^L)$. Otherwise, both will choose B. Therefore, player i's expected pay-off from saying Hawk is

$$[1 - (F(c^H) - F(c^L))](-c_i). \tag{11}$$

Suppose instead player i says Dove. If the opponent is a fairly tough type (who says Hawk according to his equilibrium strategy) then both will choose B at the action stage. Otherwise, the opponent will say Dove, player i will choose N at the action stage, and the opponent will choose B if he is very tough and N if he is normal. Therefore, player i's expected pay-off from saying Dove is

$$[F(c^H) - F(c^L)](-c_i) + F(c^L)(-d). \tag{12}$$

But, equation (6) implies that (11) equals (12) if $c_i = c^H$ so player i of type c^H is indifferent between saying Dove and Hawk. By (8), if $c_i > c^H$ then (11) is smaller than (12) so player i prefers to say Dove. If $c^* \leq c_i < c^H$, (11) is bigger than (12) so player i prefers to say Hawk. ‖

These results show that the strategy profile specified is a perfect Bayesian equilibrium of the cheap-talk extension of the arms race game. Since $c^L \to 0$ and $c^H \to 0$ as $\mu \to 0$, it is evident from the construction of the strategies that the fraction of types that play B in equilibrium goes to zero as μ goes to zero. This completes the proof of our main theorem.

We end with a few remarks. First, there are other informative equilibria. For example, we have been focusing on the solution labelled (c^L, c^H) in Figure 1, but there is another intersection of the two curves, labelled $(\hat{c}^L, \hat{c}^H)$, which also satisfies (5) and (6). One can check that $\mu < \hat{c}^L < \hat{c}^* < \hat{c}^H < \bar{c}$, where $\hat{c}^*$ solves $(1 - F(\hat{c}^H))(\mu - \hat{c}^*) + F(\hat{c}^L)(-\hat{c}^*) = F(\hat{c}^L)(-d)$. In fact, the arguments made above go through and so $(\hat{c}^L, \hat{c}^H)$ represents an alternative perfect Bayesian equilibrium. However, as $\mu \to 0$, $(\hat{c}^L, \hat{c}^H)$ converges to $(0, c^{\text{med}})$, where c^{med} is the median cost, $F(c^{\text{med}}) = \frac{1}{2}$. This implies that (N, N) is the outcome of the game approximately half of the time, much worse than the equilibrium represented by (c^L, c^H).

In equilibrium normal types get taken advantage of by very tough types, but this cannot be avoided. *Any* cheap-talk equilibrium which does not involve an arms race with probability one must have the property that dominant strategy types sometimes arm unilaterally against peaceful opponents. The only way to prevent such unilateral arms build-up would be to get the dominant strategy type to reveal his true nature, thereby alerting the opponent that he should arm. But this would not be incentive compatible since the dominant strategy type does not want his opponent to arm. Our equilibrium allows the dominant strategy types to remain undetected until the time when actions are chosen.

Recent research has studied the role of informative public signals in global games models of the Carlsson and van Damme (1993) type (see Metz (2000), Hellwig (2001), Morris and Shin (2002*a*)). A sufficiently precise public signal may restore multiple equilibria in that model by making the common shock sufficiently commonly known (see Hellwig, 2001). This argument does not apply to our model, since we have independent types and no unobservable, common shock. However, as discussed above, one can view our cheap-talk stage as endogenously generating a public signal via voluntary disclosures. Whether or not multiple continuation equilibria exist following this disclosure depends on whether or not the cumulative distribution of types conditional on the public information satisfies the multiplier condition.

5. CONCLUSION

This paper makes two contributions. First, we provide a formalization of Schelling's (1960) surprise attack dilemma, using incomplete information about preferences rather than an imperfect warning system. Our model is very close to Schelling's *informal* argument: "if I go downstairs to investigate a noise at night, with a gun in my hand, and find myself face to face with a burglar who has a gun in his hand, there is a danger of an outcome that neither of us desires. Even if he prefers to leave quietly, and I wish him to, there is a danger that he may *think* I want to shoot, and shoot first. Worse, there is danger that he may think that *I* think *he* wants to shoot. Or he may think that *I* think *he* thinks *I* want to shoot. And so on" (Schelling, 1960, p. 207). We show that, even if types are independent and the true pay-off matrix is very likely to be that of a stag hunt game, a *multiplier condition* implies the existence of a unique Bayesian–Nash equilibrium. In this equilibrium, there is an arms race with probability one.

Our second contribution is to show how cheap-talk can resolve Schelling's dilemma. The cheap-talk equilibrium has several interesting properties. Fairly tough types, with an intermediate propensity to arm, say Hawk in order to minimize the probability of a coordination failure. Very tough types and normal types say Dove in order to minimize the probability that the opponent arms. When a very tough type meets a normal type, there is an exchange of dovish messages followed by a unilateral arms build-up by the very tough type. When two normal types meet, they coexist peacefully. For the normal type, this gamble is worth taking as long as the very tough types are rare. The introduction of cheap-talk raises *all* types' expected pay-offs, since without cheap-talk the unique equilibrium involves an arms race with probability one (when the multiplier condition holds).

Much further work remains to be done. In this article, a player is in effect a nation which behaves as a unitary actor with well-defined preferences. But in reality, the population may contain elements that have widely different preferences, and they may try to influence the course of events. Extremists on both sides may not be happy with negotiations that reduce tensions between the two nations. Rioting may be one way for them to signal their strength. If a riot in one country makes the leader of the other country fearful, then he will be more likely to arm, which may lead to escalation, which may make the extremists better off. What will negotiations look like in such circumstances?

APPENDIX

The following result was discussed in the text.

Theorem A1. *Suppose the multiplier condition is satisfied. Suppose there is $\hat{c} \in (0, \bar{c})$ such that each player $i \in \{1, 2\}$ says Hawk if $c_i < \hat{c}$ and Dove if $c_i > \hat{c}$. Then, there will be an arms race with probability one.*

Proof. Suppose player i says Hawk if $c_i < \hat{c}$ and Dove if $c_i > \hat{c}$. The probability that a player says Dove is $1 - F(\hat{c}) > 0$, and the probability that a player says Hawk is $F(\hat{c}) > 0$. Type $\hat{c}$ is indifferent between saying Hawk and Dove, he can send either message (it does not matter). In order to obtain a contradiction, suppose some types sometimes play N.

If both players say Hawk, then it is easy to check that both must choose B, due to the multiplier condition. Let p denote the probability that a player who says Hawk will choose N when he hears the opponent say Dove. Dominant strategy types with cost less than $\hat{c}$ say Hawk, and they will surely choose B. Therefore, $p < 1$. Let q^H (resp. q^D) denote the probability that a player who says Dove will choose N when he hears the opponent say Hawk (resp. Dove). For dominant strategy types to be willing to say Hawk, the probability that the opponent chooses N must be greater when the dominant strategy type says Hawk than when he says Dove, which is equivalent to

$$q^H(1 - F(\hat{c})) \geq pF(\hat{c}) + q^D(1 - F(\hat{c})). \tag{A.1}$$

Since we assume N is sometimes played, (A.1) implies $q^H > 0$. That is, some types who say Dove must choose N if the opponent says Hawk. To make them behave that way, some types who say Hawk must choose N if the opponent says Dove. That is, $p > 0$. There is then some type $\tilde{c}$, where $\mu \leq \tilde{c} < \hat{c}$, who says Hawk and is indifferent between choosing N and B when he hears Dove. Also, since $p > 0$, the inequality (A.1) implies $q^D < 1$. That is, if both players say Dove, then there is a non-zero probability of an arms race. But since type $\hat{c}$ has a lower cost than any other type that says Dove, if type $\hat{c}$ says Dove and the opponent says Dove, then type $\hat{c}$ certainly prefers to choose B.

If type $\hat{c}$ says Hawk and the opponent says Dove, type $\hat{c}$ will strictly prefer N to B, because $\hat{c} > \tilde{c}$. Therefore, type $\hat{c}$'s expected pay-off from saying Hawk is strictly greater than $-\hat{c} + \mu q^H(1 - F(\hat{c}))$, which is what he would get by always choosing B. But,

$$-\hat{c} + \mu q^H(1 - F(\hat{c})) \geq -\hat{c} + \mu[pF(\hat{c}) + q^D(1 - F(\hat{c}))] \tag{A.2}$$

by (A.1). Thus, by saying Hawk, type $\hat{c}$ gets strictly more than the R.H.S. of (A.2).

Suppose instead that type $\hat{c}$ says Dove. We know that he will prefer to play B when the opponent says Dove. If in addition he plays B when the opponent says Hawk, then by saying Dove he gets the R.H.S. of (A.2). This is strictly less than what he expects from saying Hawk, which is a contradiction of the definition of $\hat{c}$. Thus, type $\hat{c}$ must strictly prefer to play N when he says Dove and the opponent says Hawk. Since his cost of building is lower than any other type who ever says Dove, all types who say Dove must play N when the opponent says Hawk. That is, $q^H = 1$. But then all types who say Hawk, except the dominant strategy types, will choose N when the opponent says Dove. Therefore, $\tilde{c} = \mu$. Type $\hat{c}$'s expected pay-off from saying Dove is

$$F(\hat{c})[p \times 0 + (1 - p) \times (-d)] + (1 - F(\hat{c}))[q^D(\mu - \hat{c}) + (1 - q^D)(-\hat{c})].$$

His expected pay-off from saying Hawk is

$$F(\hat{c})(-\hat{c}) + (1 - F(\hat{c})) \times 0.$$

The last two expressions must be equal, by definition of $\hat{c}$. This implies

$$F(\hat{c})(1 - p)d - (1 - F(\hat{c}))q^D\mu = (2F(\hat{c}) - 1)\hat{c}. \tag{A.3}$$

Now, (A.1) and $q^H = 1$ imply that

$$(2F(\hat{c}) - 1)\hat{c} \leq ((1 - p)F(\hat{c}) - q^D(1 - F(\hat{c})))\hat{c}. \tag{A.4}$$

Substituting from (A.3) into the L.H.S. of (A.4), and rearranging, yields

$$F(\hat{c})(1 - p)(d - \hat{c}) \leq q^D(1 - F(\hat{c}))(\mu - \hat{c}).$$

However, the L.H.S. is strictly positive and the R.H.S. is non-positive, given $\mu < \hat{c} < d$, a contradiction. ‖

Before proving Lemma 1, we need a preliminary technical result. This is where we use the assumption that F has a Taylor series representation. Formally, assume there are coefficients $a_0, a_1, a_2 \ldots$ such that

$$F(c) = \sum_{j=0}^{\infty} a_j c^j \tag{A.5}$$

for all $c \in [0, \bar{c}]$.

Lemma A1. *If the multiplier condition is satisfied, then there exists $\gamma > 0$ such that $F'(c)d > 1$ for all $c \in (0, \gamma)$.*

Proof. Since F has a power series representation, the function $\lambda(c) \equiv F(c)d - c$ also has a power series representation

$$\lambda(c) = \sum_{j=0}^{\infty} k_j c^j. \tag{A.6}$$

The multiplier condition says that $\lambda(c) \geq 0$ for all $c \geq 0$. Moreover, $k_0 = 0$ since $\lambda(0) = 0$. Also, as $\lambda(\bar{c}) = d - \bar{c} > 0$ by assumption, there is j such that $k_j \neq 0$. Let $n \geq 1$ be the *smallest* integer such that $k_n \neq 0$. For small enough $c > 0$ the expression in (A.6) will be dominated by the term $k_n c^n$. Hence, we must have $k_n > 0$ for $\lambda(c) \geq 0$ to be true for c close to zero. The derivative of $\lambda(c)$ is

$$\lambda'(c) = \sum_{j=1}^{\infty} j k_j c^{j-1}$$

which for small enough $c > 0$ is dominated by the term $n k_n c^{n-1} > 0$. Hence, $\lambda'(c) = F'(c)d - 1 > 0$ for $c > 0$ close enough to zero. ‖

We now prove Lemma 1.

Define two functions H and G as follows:

$$H(x, y) \equiv [F(y) - F(x)]x - (1 - F(y))\mu \tag{A.7}$$

$$G(x, y) \equiv [1 - 2(F(y) - F(x))]y - F(x)d. \tag{A.8}$$

Then, equations (5) and (6) are equivalent to the statement that $(x, y) = (c^L, c^H)$ solves the equation system

$$\begin{aligned} H(x, y) &= 0 \\ G(x, y) &= 0. \end{aligned} \tag{A.9}$$

To analyse this system consider the shape of the two curves defined by (A.9), restricting our attention to x and y in $[0, \bar{c}]$. We have

$$H(0, y) = -(1 - F(y))\mu.$$

Therefore, there is a unique $y \in [0, \bar{c}]$, namely, $y = \bar{c}$, such that $H(0, y) = 0$. Similarly,

$$H(\bar{c}, y) = -(1 - F(y))(\mu + \bar{c})$$

so that there is a unique $y \in [0, \bar{c}]$, namely, $y = \bar{c}$, such that $H(\bar{c}, y) = 0$. For $0 < x < \bar{c}$ we notice that

$$H(x, 0) = -F(x)x - \mu < 0$$

$$H(x, \bar{c}) \equiv [1 - F(x)]x > 0$$

and

$$\frac{\partial H(x, y)}{\partial y} = F'(y)(x + \mu) > 0.$$

Therefore, there is a unique $y \in (0, \bar{c})$ that satisfies $H(x, y) = 0$. We may write $y = \phi(x)$, where $H(x, \phi(x)) \equiv 0$ for all $x \in [0, \bar{c}]$. Notice that for all $x > 0$, $\mu \to 0$ implies $\phi(x) \to x$.

Next, we turn to the G function. We have

$$G(x, 0) \equiv -F(x)d$$

so that there is a unique $x \in [0, \bar{c}]$, namely, $x = 0$, such that $G(x, 0) = 0$. Let c^{med} denote the median type ($F(c^{\text{med}}) = 1/2$). We have

$$G(x, c^{\text{med}}) \equiv \left[1 - 2\left(\tfrac{1}{2} - F(x)\right)\right]c^{\text{med}} - F(x)d = F(x)(2c^{\text{med}} - d).$$

Notice that $2c^{\text{med}} = c^{\text{med}}/F(c^{\text{med}}) \leq d$ by the multiplier assumption. If $c^{\text{med}}/F(c^{\text{med}}) = d$ then $G(x, c^{\text{med}}) = 0$ for all $x \in [0, \bar{c}]$. However, if $c^{\text{med}}/F(c^{\text{med}}) < d$, then there is a unique $x \in [0, \bar{c}]$, namely $x = 0$, such that $G(x, c^{\text{med}}) = 0$.

Now suppose $0 < y < c^{\text{med}}$. Then,

$$G(0, y) \equiv [1 - 2F(y)]y > 0$$

and

$$G(y, y) \equiv y - F(y)d \leq 0$$

by assumption. Moreover, for $y < c^{\text{med}}$,

$$\frac{\partial G(x, y)}{\partial x} = F'(x)(2y - d) = F'(x)\left(\frac{y}{F(c^{\text{med}})} - d\right) < F'(x)\left(\frac{y}{F(y)} - d\right) \leq 0 \tag{A.10}$$

using the multiplier condition. Hence, for each $y \in (0, c^{\text{med}})$, there is a unique $x \in (0, y]$ such that $G(x, y) = 0$. We may write $x = \theta(y)$, where $G(\theta(y), y) \equiv 0$ for all $y \in (0, c^{\text{med}})$.

Claim. *If $0 < x \leq y$ and x and y are sufficiently close to zero, then*

$$0 < \frac{d\theta(y)}{dy} < 1.$$

Proof. Totally differentiating $G(\theta(y), y) \equiv 0$, we obtain

$$\frac{d\theta(y)}{dy}\frac{\partial G(x, y)}{\partial x} + \frac{\partial G(x, y)}{\partial y} = 0.$$

We calculate

$$\frac{\partial G(x, y)}{\partial x} \equiv F'(x)(2y - d)$$

and

$$\frac{\partial G(x, y)}{\partial y} \equiv 1 - 2(F(y) - F(x)) - 2F'(y)y.$$

Therefore,

$$\frac{d\theta(y)}{dy} = -\frac{\partial G(x, y)/\partial y}{\partial G(x, y)/\partial x} = \frac{1 - 2(F(y) - F(x)) - 2F'(y)y}{F'(x)(d - 2y)}. \tag{A.11}$$

For small enough x and y, (A.11) is strictly positive since both the numerator and denominator are strictly positive. To show that (A.11) is strictly smaller than 1, it suffices to show that

$$\frac{1 - 2(F(y) - F(x)) - 2F'(y)y}{F'(x)(d - 2y)} < \frac{1}{F'(x)d} \tag{A.12}$$

since, for small enough x, $F'(x)d > 1$ by Lemma A1. But (A.12) is equivalent to

$$(F'(y)d - 1)y + (F(y) - F(x))d > 0. \tag{A.13}$$

The first term in (A.13) is strictly positive for small enough y, by Lemma A1, while the second term is non-negative since $y \geq x$. Thus, (A.13) is satisfied. ‖

Figure 1 is a typical depiction of (A.9). Notice that the G function does not involve μ, hence the function θ does not involve μ either. However, for all $x > 0$, $\mu \to 0$ implies $\phi(x) \to x$. Since $0 < d\theta(y)/dy < 1$, for small enough $\mu > 0$ the two curves $y = \phi(x)$ and $x = \theta(y)$ must have an intersection arbitrarily close to the point $(0, 0)$ in the positive quadrant, and with $y > x > 0$ as depicted in the figure. This point is denoted $(x, y) = (c^L, c^H)$. Notice that $c^H > c^L$.

It remains only to show that $\mu < c^L < c^* < c^H$. Notice that for (c^L, c^H) close to zero we are guaranteed that

$$1 - F(c^H) > F(c^H) - F(c^L). \tag{A.14}$$

The equation $H(c^L, c^H) = 0$ implies

$$[F(c^H) - F(c^L)]c^L = (1 - F(c^H))\mu. \tag{A.15}$$

Since $c^L > 0$ and $\mu > 0$, (A.14) and (A.15) imply $\mu < c^L$.

Finally, let c^* solve (7). That is, let c^* satisfy

$$(1 - F(c^H) + F(c^L))c^* = (1 - F(c^H))\mu + F(c^L)d. \tag{A.16}$$

Clearly c^* exists, because $1 - F(c^H) + F(c^L) > 0$. We claim that $c^L < c^* < c^H$. The equation $G(c^L, c^H) = 0$ implies

$$[1 - F(c^H) - (F(c^H) - F(c^L))]c^H = F(c^L)(d - c^L). \tag{A.17}$$

Since $c^L < c^H$, (A.17) implies

$$(1 - F(c^H) - (F(c^H) - F(c^L)))c^L < F(c^L)(d - c^L) \tag{A.18}$$

and also

$$(1 - F(c^H))c^H - (F(c^H) - F(c^L))c^L > F(c^L)(d - c^H). \tag{A.19}$$

Now substitute from (A.15) into (A.18) and (A.19) to get

$$(1 - F(c^H) + F(c^L))c^L < (1 - F(c^H))\mu + F(c^L)d$$

and

$$(1 - F(c^H) + F(c^L))c^H > (1 - F(c^H))\mu + F(c^L)d.$$

But these two inequalities and equation (A.16) imply $c^L < c^* < c^H$.

Acknowledgements. We thank three anonymous referees and the editor, Mark Armstrong, for many valuable comments. We also thank Eric Maskin, Josef Perktold, Ariel Rubinstein and Bill Zame, as well as the participants in many seminars, for their comments. Tomas Sjöström acknowledges financial support from National Science Foundation grant SES-0111527. Any errors are our responsibility.

REFERENCES

AUMANN, R. (1990), "Nash Equilibria are Not Self-Enforcing", in J. J. Gabszewicz, J.-F. Richard and L. A. Wolsey (eds.) *Economic Decision-Making: Games, Econometrics and Optimization* (Amsterdam: Elsevier).
AUSTEN-SMITH, D. (1990), "Information Transmission in Debate", *American Journal of Political Science*, **34**, 124–152.
BALIGA, S. and MORRIS, S. (2002), "Coordination, Spillovers and Cheap-Talk", *Journal of Economic Theory*, **105**, 450–468.
BANKS, J. and CALVERT, R. (1992), "A Battle-of-the-Sexes Game with Incomplete Information", *Games and Economic Behavior*, **4**, 347–372.
CARLSSON, H. and VAN DAMME, E. (1993), "Global Games and Equilibrium Selection", *Econometrica*, **61**, 989–1018.
CHURCHILL, W. (1931) *The World Crisis (abbr. and revised)* (New York: C. Scribner).
CRAIG, G. (1978) *Germany 1866–1945* (New York: Oxford University Press).
CRAWFORD, V. and SOBEL, J. (1982), "Strategic Information Transmission", *Econometrica*, **50**, 1431–1451.
GREEN, J. and STOKEY, N. (1980), "A Two Person Game of Information Transmission" (H.I.E.R. Discussion Paper No. 751, Harvard University).
HARSANYI, J. (1973), "Games With Randomly Disturbed Payoffs: A New Rationale for Mixed Strategy Equilibrium Points", *International Journal of Game Theory*, **2**, 1–23.
HELLWIG, C. (2001), "Public Information, Private Information and Multiplicity of Equilibria in Coordination Games" (Mimeo, L.S.E.).
JERVIS, R. (1976) *Perception and Misperception in International Politics* (Princeton, NJ: Princeton University Press).
JERVIS, R. (1978), "Cooperation under the Security Dilemma", *World Politics*, **30**, 167–214.
KISSINGER, H. (1994) *Diplomacy* (New York: Touchstone).
KYDD, A. (1997), "Game Theory and the Spiral Model", *World Politics*, **49**, 371–400.
LEFFLER, M. (1992) *A Preponderance of Power: National Security, the Truman Administration and the Cold War* (Stanford: Stanford University Press).
MATTHEWS, S. and POSTLEWAITE, A. (1989), "Pre-play Communication in Two-person Sealed-bid Double Auctions", *Journal of Economic Theory*, **48**, 238–263.
METZ, C. (2000), "Public and Private Information in Self-fulfilling Currency Crises" (Mimeo, University of Kassel).
MORRIS, S. and SHIN, H. (2002*a*), "The Social Value of Public Information", *American Economic Review*, **92**, 1521–1534.
MORRIS, S. and SHIN, H. (2002*b*), "Heterogeneity and Uniqueness in Interaction Games" (Mimeo, Yale University).
MORRIS, S. and SHIN, H. (2003), "Global Games: Theory and Applications", in M. Dewatripont, L. Hansen and S. Turnovsky (eds.) *Advances in Economics and Econometrics (Proceedings of the Eighth World Congress of the Econometric Society)* (Cambridge, UK: Cambridge University Press).
O'NEILL, B. (1999) *Honor, Symbols and War* (Ann Arbor: University of Michigan Press).
REAGAN, R. (1983) *Public Papers of the Presidents of the United States, Ronald Reagan, Book 1* (Washington D.C.: U.S. Government Printing Office).
RUBINSTEIN, A. (1989), "The Electronic Mail Game: Strategic Behavior under Almost Common Knowledge", *American Economic Review*, **79**, 385–391.
SCHELLING, T. C. (1960) *The Strategy of Conflict* (Cambridge, MA: Harvard University Press).
SONTAG, R. (1933) *European Diplomatic History, 1871–1932* (New York: Appleton-Century-Crofts).
THUCYDIDES (1972) *The History of the Peloponnesian War* (London: Penguin Classics).
WAINSTEIN, L. (1971), "The Dreadnought Gap", in R. Art and K. Waltz (eds.) *The Use of Force* (Boston: Little Brown).