Strategic Ambiguity and Arms Proliferation

Sandeep Baliga
Northwestern University

Tomas Sjöström
Rutgers University

March 13, 2006

Abstract

A big power is facing a small power that may have developed WMDs. The small power can create *strategic ambiguity* by not allowing arms inspections. We study the impact of strategic ambiguity on arms proliferation and the probability of conflict. Creating strategic ambiguity is a substitute for actually acquiring new weapons: ambiguity reduces the incentive for the small power to invest in a weapons program, which reduces the risk of arms proliferation. Therefore, strategic ambiguity tends to benefit the big power. On the other hand, strategic ambiguity may hurt the small power because it does not always protect it from an attack. Cheap-talk messages can be used to trigger inspections when they are most valuable to the big power. To preserve incentive compatibility, the "tough" messages which make inspections more likely must imply a greater risk of arms proliferation.

1 Introduction

The North Korean regime claims it has developed nuclear weapons in order to deter a U.S. attack (Pinkston and Diamond [15]). It is known that the North Koreans have access to weapons-grade plutonium, but the existence of operational nuclear weapons has not been verified (Hecker [8], Norris and Kristensen [13]). There is *strategic ambiguity* about North Korea's nuclear capability. The North Koreans may have chosen a policy of strategic ambiguity because they feel threatened but lack actual deterrence capability. Similar considerations may explain Saddam Hussein's behavior after 1991.

On the other hand, if the North Koreans really do possess operational nuclear weapons, they may fear that if they reveal it (for example by conducting a nuclear test) then other countries will act against them, perhaps out of fear that the weapons will end up in the hands of extremists. In this case they may also prefer a policy of ambiguity. North Korea's Vice Minister of Foreign Affairs told a visiting delegation of American scientists that revealing too much would be dangerous: "If you go back to the United States and say that the North already has nuclear weapons, this may cause the U.S. to act against us" (Hecker [8]). In fact, the U.S. national security advisor Stephen Hadley has threatened "punitive action" if North Korea conducts a nuclear test (New York Times, July 25, 2005). The North Koreans may not consider this an empty threat. After all, the stated motive for attacking Iraq in 2003 was evidence of the existence of weapons of mass destruction (WMDs).

But there is a wide-spread belief that the Americans attacked Iraq only because they knew that Iraq did *not* have nuclear weapons. Takeyh [19] argues that Iran's leaders think they are next in line to be attacked. The U.S. engagement in Iraq may rule out an attack at the present time, but the Iranians hope to establish a deterrence capability before their time is up. Iranian newspaper editorials cited by Takeyh [19] lend support to this view: "Based on Bush's record after 9/11, one can only conclude that the U.S. has not invaded our two immediate neighbors to the east and the west just to fight Al Qaeda. Consequently, astute political observers warn that Iran is next on the U.S. list of direct targets" (*Iran News*). "In the contemporary world, it is obvious that having access to advanced weapons shall cause deterrence and therefore security, and will neutralize the evil wishes of great powers to attack other nations" (*Jumhuri-ye Islami*). Similarly, it is possible that North Korea's nuclear weapons program is motivated by a desire to protect their current regime against a perceived threat (as opposed to, say, an ambition to unite North and South Korea). Perceptions drive behavior, whether or not these perceptions are accurate. Simply dismissing a leader's fears as unfounded cannot change the fact that perceptions matter.

The standard argument against ambiguity, and in favor of weapons inspections, is embodied in Article 3 of the *Treaty on the Non-proliferation of Nuclear Weapons,* known as the NPT [22]. Incomplete information generates fear which causes arms races and wars. Therefore, strategic ambiguity is thought to make the world a more dangerous place. Simply talking to the opponent may not reduce fear, because talk is cheap. However, if the opponent reveals that he is unarmed then fear is reduced. Therefore, the NPT requires

that nations submit to inspection and verification of nuclear facilities by the IAEA. The NPT is meant to promote peace and trust. However, Israel, India and Pakistan have not signed the NPT, North Korea has withdrawn from it, and Iran is close to violating it. The exact quality and quantity of WMDs in these countries is unknown. For example, Pakistan and India possess short-range nuclear weapons but it is unclear whether they have intercontinental ballistic missiles or are developing them (Norris and Kristensen [14], Norris, Kristensen and Handler [12]). Why wouldn't any leader who desires peace sign the NPT, disarm, and allow arms inspectors to verify it? The problem is obvious: a leader who doesn't trust his opponents is unwilling to disarm and then reveal that he is defenseless. The NPT cannot create trust where no trust exists to begin with. Therefore, a withdrawal from the NPT is not necessarily a signal that the leader has aggressive intentions, neither is it necessarily a signal that a nuclear arsenal exists. Strategic ambiguity can be a substitute for WMDs. The ambiguity is in itself a deterrence, which reduces the value of acquiring WMDs. Therefore, ambiguity may lead to less arms proliferation. The standard argument in favor of weapons inspections does not take this into account.

Sobel [18] gives an argument in favor of ambiguity. A strong (i.e., well armed) country is less likely to be attacked, so if disclosure is possible then all strong countries will reveal their weapons. In equilibrium, any country which does not reveal its weapons will be known to be weak. Thus, weak countries cannot rely on strategic ambiguity for protection and they will be attacked. (See Grossman [7] and Milgrom [10] for similar "unraveling" arguments.) If disclosure is impossible, then weak and strong countries cannot be distinguished, which reduces the risk of war. This argument suggests that small powers (whose capabilities are uncertain) may be better off *ex ante* if their strength is never revealed, but big powers (who are known to be strong) prefer to eliminate all ambiguity so they can prey on the weak. As Sobel points out, in practice it may be difficult to prevent well armed countries from disclosing their weapons. For example, during the cold war each nuclear power disclosed its capabilities in order to deter an attack, and it is unclear how they could have been prevented from doing so. However, in the asymmetric post cold war world, it is not necessarily in the interest of a small power to reveal that it has advanced weapons. Doing so might trigger a preemptive attack from a big power, perhaps due to fears that the advanced weapons might end up in the hands of extremists. If "hard" information about capabilities can be a negative signal about "soft" information,

then the unraveling result does not hold, and it may be possible to sustain strategic ambiguity even though disclosure is possible.

But Sobel's [18] argument in favor of ambiguity relies on the assumption that a country's strength is exogenously given. If this assumption is dropped then the argument becomes quite different. Suppose small powers can acquire advanced weapons at a cost. In this case, ambiguity requires that small powers arm themselves with a positive probability which is determined in equilibrium, together with a positive probability that the big power attacks. In an equilibrium with ambiguity, some attacks will be "mistakes", i.e., they could have been avoided, had the small power's true strength been publicly known. But the equilibrium probability that the small power arms itself is smaller than it would be with full disclosure. Thus, ambiguity about the small power's arsenal reduces the risk of arms proliferation, which is good for the big power. Moreover, if cheap-talk is allowed, then those leaders who are more likely to make "mistakes" can send a message which increases the probability of an inspection, and thereby reduce the likelihood of a mistake.

We consider an "arms proliferation game". There are two players, A and B, who are the leaders of two countries, also denoted A and B. Country A is a big power that is known to possess advanced weapons. Country B is a small power which initially is weak, i.e., it lacks advanced weapons. But B can try to acquire advanced weapons by making an investment. If this succeeds, then B will become strong. Player A fears that if B acquires advanced weapons then these weapons may end up in the hands of extremists. Player B knows if this fear is justified, but this is *soft* (unverifiable) information. In Baliga and Sjöström [1], we studied how soft private information can lead to arms races and wars. However, we did not allow the players to reveal any *hard* information. The standard argument discussed above suggests that arms inspections promote peace and trust. To study this issue, we suppose B's true strength is hard information which can be verified by weapons inspectors. Player A's decision is whether or not to attack player B. The optimal decision depends on A's preferences (his type) and his beliefs. At most three equilibria can exist in this game. There is an equilibrium with *full ambiguity* where B never allows arms inspections, and a *full disclosure* equilibrium where he always does. Finally, a *communication equilibrium* involving informative cheap-talk may exist.

With full disclosure, B's fear of A impels B to try to increase his strength by investing. With full ambiguity, B has less incentive to invest, which tends to make A better off. However, if A is an *opportunistic type*, then he would

like to attack if B is weak, in order to obtain control of some resource or simply achieve "regime change". Therefore, just like the players in Sobel's [18] model, the opportunistic type values information about B's strength. With full ambiguity, the probability that B invests is decreasing in the cost of investing and increasing in the value of advanced weapons in case of a conflict. If the cost of investing is low and/or the value of advanced weapons is large, then ambiguity does not significantly reduce the risk of weapons proliferation, so the opportunistic type prefers arms inspections in this case. But if the cost is high and/or the value is small, then ambiguity makes A better off regardless of type.

If different types of player A disagree about whether ambiguity is desirable, then it is possible to construct a communication equilibrium. Player A sends either a "tough" or a "conciliatory" message. The conciliatory message allows B to preserve ambiguity about his true strength, which reduces B's incentive to invest and thereby reduces the risk of arms proliferation. The tough message, which can be interpreted as insisting that B signs the NPT, removes the ambiguity and thereby increases the risk of arms proliferation. Player A uses a "non-convex" strategy. If A is a peaceful "dove" or an aggressive "hawk", then he has a strong intrinsic preference for a particular action ("don't attack" for doves, "attack" for hawks). These extreme types do not need to know B's strength in order to decide what to do. Since they want to reduce the risk of arms proliferation, they send the conciliatory message. It is the *intermediate* types who send the tough message. These intermediate types are opportunists who want to know B's true strength. (However, some opportunists who are "almost" hawks or "almost" doves send the conciliatory message.) We show in the Appendix that this is the only equilibrium where cheap-talk is effective in influencing B's investment decision. The uniqueness (modulo relabelling the messages) comes from the fact that incentive compatibility requires that a message which makes inspections more likely *must* increase the risk of arms proliferation.

In the communication equilibrium, player A can trigger inspections by sending the tough message. By revealed preference, all of A's types prefer the communication equilibrium to the equilibrium with full disclosure. But B may prefer full disclosure if ambiguity is not an effective deterrent. In both the communication equilibrium and the full ambiguity equilibrium, some opportunistic types will attack even though B is strong. The frequency of such "mistakes" determines whether or not ambiguity is good for the small power. There are parameter values where strategic ambiguity about the small

power's arsenal is good for the big power but bad for the small power. The key point is that ambiguity is a substitute for actually acquiring advanced weapons, so more ambiguity means less arms proliferation.

The literature on cheap talk games where the sender has private information was pioneered by Crawford and Sobel [5]. In our model, both the sender (A) and the receiver (B) take actions. Although A's types are ordered in a natural way according to their propensity to attack, only *intermediate* (opportunistic) types have a *demand for information*. This non-convexity cannot exist in sender-receiver games of the kind studied by Crawford and Sobel [5] where the sender takes no action. There is a related literature on financial intermediation and auditing, where costly inspections are used to verify incomes (Townsend [21], Diamond [6], Bond [2], Border and Sobel [3] and Mookherjee and Png [11]). In this literature, the value of information is traded off against the resource cost of inspections. In our model, inspections do not consume significant real resources. We instead focus on a commitment problem: player A cannot commit not to attack if player B reveals that he is weak. We show that the optimal policy may be to forego inspections in order to allow B the security to take an action which is good for both A and B (i.e., not to invest).

The paper proceeds as follows. In Section 2, we describe the model. In Section 3, we analyze equilibria without communication, where player B's strength is either *always* revealed or *never* revealed. In Section 4 we consider communication equilibria where A's message determines whether or not B reveals his true strength. Section 5 concludes.

2 The Arms Proliferation Game

2.1 Strategies and Payoffs

There are two players, A and B. Initially, player B has no advanced weapons, but he can try to improve his arsenal by making an investment. The decision is binary: he either invests, or he doesn't invest. If B doesn't invest, then he will not acquire advanced weapons. If B invests, then there are three possible outcomes. With probability $\tau\sigma$, B acquires advanced weapons, and these will eventually be shared with extremists (perhaps because B's scientists are corrupt, or because extremists gain power in country B). With probability $(1-\tau)\sigma$, B acquires advanced weapons, but extremists will not. With prob-

ability $1-\sigma$, B does not acquire advanced weapons. Assume $0<\tau<1$ and $0<\sigma<1$. The cost of investing is $k>0$. If B acquires advanced weapons then B is *strong*, otherwise B is *weak*. Notice that if B invests then he will be strong with probability σ, but if he doesn't invest then he is weak for sure. Player A cannot observe B's true strength or his investment.

Player B true strength is hard information which can be verified by (perfectly reliable) inspectors. If there is an inspection, then B incurs a small cost ε which is drawn from a distribution with support $[0,\bar{\varepsilon}]$ and density h. The inspectors will publicly announce B's strength ("strong" or "weak"). However, the inspectors cannot verify whether or not advanced weapons will be acquired by extremists (this is B's soft information).

In the final stage of the game, A decides whether or not to attack B. If player A attacks, then A gets a benefit a and B suffers a cost α. We refer to a as player A's *type*. It is A's private information. Player B thinks a has a continuous distribution with support $[a_0,a_1]$, where $a_0<0<a_1$. The density is denoted f and the c.d.f. is denoted F. If B is strong, i.e. has advanced weapons, then he can use them against A if A attacks. This yields an expected benefit $\gamma\in(0,\alpha)$ for B but an expected cost $c>0$ for A. If there is no attack and extremists acquire advanced weapons, then A suffers a cost $b>0$. To simplify, we assume B does not care if extremists get weapons (a small cost or benefit could be introduced without changing the main results). Again to simplify, we assume that if A attacks B then he eliminates the threat posed by extremists. More generally, the threat could be reduced but not completely eliminated. It is useful to define the *normalized cost of investing* to be

$$\kappa\equiv\frac{k}{\sigma\gamma}.$$

The payoffs are summarized in the following matrix.

	B is strong	B is weak
A attacks	$a-c,\ -\alpha+\gamma$	$a,\ -\alpha$
No attack	$-\tau b,\ 0$	$0,\ 0$

This payoff matrix assumes that A believes that, if B is strong and there is no attack, then extremists will acquire advanced weapons with probability τ. Although this is always true on the equilibrium path, off the equilibrium path A's beliefs may differ (see Section 2.3). Also, the payoff matrix does not include the cost of B's investment or the cost of the inspection. For example,

if B invests but does not acquire advanced weapons, there is an inspection, and A attacks, then B's final payoff is $-\alpha - k - \varepsilon$.

2.2 Time Line

The time line is as follows.

Time 0: Player A privately learns a.

Time 1: Cheap-talk.

Time 2: Player B decides whether or not to invest.

Time 3: If B invested, then he privately observes the outcome.

Time 4: The cost ε is realized and player B decides whether or not to allow inspections. If inspections take place, then the inspectors publicly announce B's true strength.

Time 5: Player A decides whether or not to attack.

2.3 Equilibrium

The solution concept is sequential equilibrium (Kreps and Wilson [9]). Along the equilibrium path, each player's beliefs are computed from the equilibrium strategies using Bayesian updating. Given these beliefs, each player's behavior must be sequentially rational.

Along the equilibrium path, if an inspection reveals that B is strong, then A thinks advanced weapons will fall into the hands of terrorists with probability τ unless he attacks, and so A attacks if and only if $a - c > -\tau b$. If instead the inspection reveals that B is weak, then A attacks if and only if $a > 0$. If along the equilibrium path there is no inspection, then A's action is determined by his beliefs about B's strength. If B's equilibrium strategy implies that he has invested with probability η, then B is strong with probability $\eta\sigma$, so A attacks if and only if $a - \eta\sigma c > -\eta\sigma\tau b$.

A "surprise inspection" is an inspection that happens even though A thinks B will refuse inspections with probability one. Following a surprise inspection (or a "surprise refusal"), A's action will depend on off-the-equilibrium path beliefs. There is some freedom to specify these beliefs, since they cannot be derived from Bayes' rule. In particular, the surprise inspection (or refusal) may be considered by A to be a signal about information that B obtained at time 3. However, the beliefs must respect the fact that, when B invests, he does not know if his weapons will end up in the hands of extremists.

2.4 Parameter Restrictions

The objective of our research is to understand how incomplete information about the opponent's motives and capabilities can trigger arms races and conflicts. Accordingly, we will assume that player A's optimal decision depends on his preferences (his type) and his beliefs. Player A is a *dove* if $a < 0$, an *opportunist* if $0 < a < c - \tau b$, and a *hawk* if $a > c - \tau b$. The probability that A is a dove is $D \equiv F(0) > 0$. The probability that he is a hawk is $H \equiv 1 - F(c - \tau b) > 0$.

Assumption 1: $\tau b < c < b$.

Notice that A's net benefit from attacking a weak B is a. A strong player B generates two kinds of fear in player A: fear that advanced weapons will be used against him if he attacks, and fear that extremists get advanced weapons if he does not attack. Whether or not the benefit to preemption outweighs the risks depends on which fear dominates, i.e., it depends on A's beliefs about how B's advanced weapons will be used. The *ex ante* net benefit from attacking a strong B is $a - c - (-\tau b)$. The first inequality in Assumption 1 says that $a > a - c - (-\tau b)$, so a strong B is *less* likely to be attacked, given that A thinks extremists will get weapons with probability τ. Without this inequality, there would be no opportunistic types, advanced weapons would never deter attacks in equilibrium, and the problem would not be interesting. Now notice that if A is *certain* that advanced weapons will end up in the hands of extremists, then the net benefit from attacking a strong B is $a - c - (-b)$. The second inequality says that $a < a - c - (-b)$, so a strong B is *more* likely to be attacked if A has these pessimistic beliefs. If the second inequality were violated, then regardless of A's beliefs, disclosing advanced weapons would always make A less likely to attack. In this case the unique equilibrium would involve full disclosure, just as in Sobel [18]. To support an equilibrium with ambiguity, A's fear that extremists get advanced weapons must be sufficiently strong. For example, nuclear weapons in the hands of North Korea or Iran may not have the range to reach the continental United States, but they do have the potential to destabilize world security in the hands of terrorists.

We are interested in the case where B's decision to allow inspections is driven mainly by his fear of A, not by the cost of an inspection. Thus,

we assume the cost of inspections is very small, which also simplifies the exposition considerably. It suffices to assume:

Assumption 2:

$$\frac{\bar{\varepsilon}}{\alpha - \gamma} < \min\{F(c - \tau b) - F(0), F(0) - F(c - b)\}$$

Our final assumption ensures that investing is cheap enough that A cannot achieve his "bliss point":

Assumption 3:

$$\kappa < 1 - F(0).$$

If Assumption 3 is violated, then the cost of investing is so high that there is an equilibrium where B never invests and always allows arms inspections. (This statement is implied by the proof of Proposition 2.) By making Assumption 3, we make sure player A cannot attain this "bliss" outcome. However, many results, such as the structure and existence of communication equilibrium, do not rely on Assumption 3.

3 Equilibria Without Communication

3.1 Equilibria with Full Disclosure

In an equilibrium with *full disclosure*, on the equilibrium path there is always an inspection. In the cheap-talk stage, all of A's types will send a message which minimizes the probability that B invests. Thus, cheap-talk cannot be effective, i.e., the probability that B invests cannot depend on a. In other words, with full disclosure, communication cannot prevent arms proliferation.

Lemma 1 *Full disclosure implies that player B invests with strictly positive probability.*

Proof. In order to derive a contradiction, suppose there is a full disclosure equilibrium where player B invests with probability zero. Along the equilibrium path, inspections will reveal that B is weak, and A attacks if and only if $a > 0$. Therefore, B's equilibrium payoff is

$$(1 - F(0))(-\alpha) - \varepsilon \tag{1}$$

If inspections should reveal that B is strong, then there is some $\tilde{a}$ such that A attacks if and only if $a \geq \tilde{a}$. If B refuses inspections, there is some $\hat{a}$ such that A attacks if and only if $a \geq \hat{a}$. Therefore, if B is strong, sequential rationality implies that he allows inspections if

$$(1 - F(\tilde{a}))(-\alpha + \gamma) - \varepsilon > (1 - F(\hat{a}))(-\alpha + \gamma) \tag{2}$$

and he refuses inspections if the inequality is reversed. There are two (non-exclusive) cases.

Case I: Suppose the inequality (2) holds for some ε. Then, if the inspection reveals that B is strong, player A's assessment must be that B deviated at time 2 by investing. However, there was no deviation at time 4, since (2) holds. When player B deviated at time 2, he did not know what the outcome of the investment would be. Therefore, player A must think that the probability that extremists will acquire advanced weapons if A doesn't attack is τ. Thus, A attacks if and only if $a - c > -\tau b$. That is, $\tilde{a} = c - \tau b$. Therefore, if inspections reveal that B is strong, the probability of an attack is $1 - F(c - \tau b)$.

Case II: Suppose the inequality in (2) is reversed for some ε. Now suppose B refuses an inspection. If in this case player A's assessment is that B is weak, then A attacks if and only if $a > 0$ (so $\hat{a} = 0$). Alternatively, A's assessment might be that B deviated from his equilibrium strategy at time 2 and invested, and B became strong and followed his equilibrium strategy by refusing inspections at time 4 (since inequality (2) is reversed). In this case, A should attack if and only if $a - c > -\tau b$ (so $\hat{a} = c - \tau b$). More generally, player A's belief following the "surprise refusal" might put positive probability on both these events.[1] In this case $0 \leq \hat{a} \leq c - \tau b$. Therefore, if B refuses inspections, the probability of an attack is at most $1 - F(0)$.

Now, it must necessarily be the case that either case I or case II applies (or both). In either case, if B is strong, he can guarantee that the probability of an attack is no greater than $1 - F(0)$. Therefore, the expected payoff from investing is no smaller than

$$\sigma(1 - F(0))(-\alpha + \gamma) + (1 - \sigma)(1 - F(0))(-\alpha) - k - \varepsilon \tag{3}$$

Since we are assuming that B does not invest, (1) must be at least as big as (3). But it can be checked that this is equivalent to the violation of

[1] By the definition of sequential equilibrium, we can disregard the possibility that B has deviated from the equilibrium strategy on more than one occasion.

Assumption 3. This contradiction implies that B must invest with positive probability. ■

We now can prove the following result.

Proposition 2 *There is an equilibrium with full disclosure. Full disclosure implies that player B invests with probability one. Cheap-talk cannot reduce the probability that B invests.*

Proof. Lemma 1 implies that, if there is full disclosure, then along the equilibrium path the inspections will sometimes reveal that B is strong. In this case, player A will attack iff and only if $a - c > -\tau b$. Therefore, the payoff from investing is given by (3), while the payoff from not investing is given by (1). But, Assumption 3 implies that (3) is strictly greater than (1). Therefore, in any full disclosure equilibrium, player B invests with probability one.

It remains to show that such an equilibrium exists. Since we know that player B prefers to invest, the only thing to check is that he also prefers to allow inspections. Suppose B refuses inspections. Let A's assessment be that B is weak. In this case, A attacks if and only if $a > 0$. Therefore, by refusing inspections, B raises the risk of an attack from $1 - F(c - \tau b)$ to $1 - F(0)$. This is more costly for the weak than for the strong. For the strong B, the cost of refusing inspections is $(F(c - \tau b) - F(0))(\alpha - \gamma)$ while the benefit is ε. Assumption 2 guarantees that the cost exceeds the benefit, so B prefers to allow inspections. ■

3.2 Equilibria with Full Ambiguity

Proposition 2 implies that B prefers to invest unless there is some ambiguity about his strength. In this section we consider equilibria with *full ambiguity*. In such an equilibrium, inspections never occur on the equilibrium path. Clearly, with full ambiguity communication cannot be effective. To maintain ambiguity, B must invest with positive probability. However, this requires that A attacks sufficiently often. In equilibrium, the probability of investment and the probability of attack are determined simultaneously, and they will depend on the normalized cost of investing, κ.

Proposition 3 *There is an equilibrium with full ambiguity. Full ambiguity implies that B invests with probability $\tilde{x}$, where $0 < \tilde{x} < 1$ if*

$$\kappa > 1 - F(\sigma(c - \tau b)), \tag{4}$$

and $\tilde{x} = 1$ otherwise. Cheap-talk cannot reduce the probability that B invests.

Proof. If B never allows inspections, then all of A's types want to minimize the probability that B invests. Therefore, the probability of investment must be independent of A's type. Let $\tilde{x}$ denote the probability that B invests. Then B is strong with probability $\sigma\tilde{x}$.

The equilibrium must satisfy a cut-off property: there is $\tilde{a}$ such that if there is no inspection then A attacks if and only if $a > \tilde{a}$. In equilibrium, doves will not attack but hawks will. Type $\tilde{a} \in (a_0, a_1)$ must be indifferent between attacking and not attacking. Since type $\tilde{a}$ expects $\tilde{a} - \sigma\tilde{x}c$ by attacking, and $-\sigma\tilde{x}\tau b$ by not attacking, the indifference condition yields

$$\tilde{a} = \sigma\tilde{x}(c - \tau b) \tag{5}$$

If B deviates by allowing inspections, and he is found to be weak, then A attacks if and only if $a > 0$. But if the weapons inspectors discover that B is strong, then we may suppose A attacks if and only if $a > c - b$. This is supported by the off–the-equilibrium path belief that advanced weapons are about to end up in the hands of extremists (which is the belief most likely to support the equilibrium, since it punishes B's deviation most strictly). Assumption 1 implies that B does not want to allow inspections.

If $0 < \tilde{x} < 1$ then B must be indifferent between investing and not investing. Since B is attacked with probability $1 - F(\tilde{a})$, he is indifferent between investing and not investing if

$$-(1 - F(\tilde{a}))\alpha = -(1 - F(\tilde{a}))(\alpha - \sigma\gamma) - k \tag{6}$$

which is the same as

$$\kappa - (1 - F(\tilde{a})) = 0 \tag{7}$$

If $\kappa > 1 - F(\tilde{a})$ then B's unique best response is $\tilde{x} = 0$, and if $\kappa < 1 - F(\tilde{a})$ then B's unique best response is $\tilde{x} = 1$. Define

$$\Gamma(x) \equiv \kappa - (1 - F(\sigma x(c - \tau b)))$$

An equilibrium where $0 < \tilde{x} < 1$ requires that both (7) and (5) hold, which implies $\Gamma(\tilde{x}) = 0$. By Assumption 3, $\kappa < 1 - F(0)$, so $\Gamma(0) < 0$. Since $\Gamma'(x) > 0$, there is $\tilde{x} \in (0, 1)$ such that $\Gamma(\tilde{x}) = 0$ if and only if $\Gamma(1) > 0$, which is equivalent to (4). Thus, if (4) holds then there is $\tilde{x} \in (0, 1)$ such that $\Gamma(\tilde{x}) = 0$, and this is the only candidate for a full ambiguity equilibrium.

(Since $\Gamma(0) < 0 < \Gamma(1)$, it is not possible that B invests with probability 0 or 1). On the other hand, if (4) is violated then $\Gamma(1) \leq 0$ so B must set $\tilde{x} = 1$.

■

In equilibrium, ambiguity has its price, because some opportunistic types attack even though B is strong. The welfare implications of ambiguity depend on the probability of such "mistakes". It is useful to define

$$a^* \equiv \frac{\tau b \sigma (c - \tau b)}{(1 - \sigma) c + \sigma \tau b} \tag{8}$$

Notice that Assumption 1 implies $0 < a^* < \sigma (c - \tau b)$, so type a^* is an opportunist. To look at the implications of ambiguity, we can distinguish two cases.

Case 1: Suppose the normalized cost of developing advanced weapons is high:

$$\kappa > 1 - F(a^*), \tag{9}$$

where a^* is defined by (8). As $a^* < \sigma(c - \tau b)$, (4) is satisfied, so B invests with probability $\tilde{x} < 1$ behind the veil of ambiguity. Player B can rely on ambiguity for protection, and with positive probability he refrains from investing. Clearly, hawks and doves strictly prefer full ambiguity to full disclosure (under full disclosure B invests with probability one). Among the opportunists, it is not hard to see that the one most likely to want inspections is precisely type $\tilde{a} = \sigma \tilde{x}(c - \tau b)$. The smaller is $\tilde{x}$, the more likely it is that type $\tilde{a}$ prefers full ambiguity. Type $\tilde{a}$'s expected utility under full ambiguity is $\tilde{a} - \sigma \tilde{x} c$. Suppose instead that there are inspections, but B invests with probability one. After the inspection, type $\tilde{a}$ attacks if and only if B is weak, which happens with probability $1 - \sigma$. Thus, type $\tilde{a}$'s expected payoff would be $(1 - \sigma)\tilde{a} - \sigma \tau b$. He prefers full ambiguity if and only if

$$(1 - \sigma)\tilde{a} - \sigma \tau b < \tilde{a} - \sigma \tilde{x} c \tag{10}$$

Using the definition of $\tilde{a}$, (10) is equivalent to $\tilde{x} < x^*$, where

$$x^* \equiv \frac{\tau b}{(1 - \sigma)c + \sigma \tau b}$$

Clearly, $\tilde{x} < x^*$ if $\tilde{x} = 0$. Suppose instead that $\tilde{x} > 0$. Since $\Gamma(0) < 0 < \Gamma(1)$ and $\Gamma'(x) > 0$, we have $\tilde{x} < x^*$ if and only if $\Gamma(x^*) > 0$, which is equivalent

to (9). Thus, in case 1 ambiguity reduces the risk of arms proliferation sufficiently to make all of A's types better off.

Case 2: Suppose the normalized cost of developing advanced weapons is low:

$$\kappa < 1 - F(a^*). \tag{11}$$

If (4) holds, then B invests with probability $\tilde{x} < 1$. Therefore, hawks and doves strictly prefer full ambiguity to full disclosure. However, by a similar reasoning as in case 1, we find that (11) implies that type $\tilde{a}$ strictly prefers full disclosure to full ambiguity (inequality (10) is reversed). If (4) is violated, then B invests with probability one under full ambiguity, so some opportunistic types strictly prefer full disclosure, because it allows them to make better decisions. In case 2, ambiguity does not significantly reduce the risk of arms proliferation. Therefore, there are always opportunistic types of A who prefer disclosure.

So far, we have considered only A's welfare. Now consider the situation from the point of view of B. With full ambiguity, player A attacks if and only $a \geq \tilde{a}$, so player B's expected utility is

$$-\sigma(\alpha - \gamma)\,(1 - F(\tilde{a})) - (1 - \sigma)\,\alpha\,(1 - F(\tilde{a})) - k.$$

The difference in B's payoffs between full disclosure and full ambiguity is

$$\sigma(\alpha - \gamma)\,[F(c - \tau b) - F(\tilde{a})] - (1 - \sigma)\,\alpha\,[F(\tilde{a}) - F(0)] - \varepsilon. \tag{12}$$

The first term is positive and is due to the fact that in the equilibrium with full ambiguity, there is a measure $F(c - \tau b) - F(\tilde{a})$ of "tough" opportunists who attack. However, there is a chance that B is strong. Under full disclosure this would be revealed, and the tough opportunists would be deterred. The second term is negative and is due to the fact that there is a measure $F(\tilde{a}) - F(0)$ of "weak" opportunists who do not attack under full ambiguity. However, there is a chance that B is weak. Under full disclosure this would be revealed, and the weak opportunists would attack. >From an ex ante point of view, B faces a trade-off: disclosure deters "tough" opportunists when B is strong, but ambiguity deters "weak" opportunists when B is weak. Without making further assumptions on the distribution of A's types we cannot sign the expression in (12).

We summarize these findings in the following proposition:

Proposition 4 *All of A's types prefer full ambiguity to full disclosure if and only if (9) holds. Player B prefers full ambiguity to full disclosure if and only if the expression in (12) is negative.*

Clearly, hawks and doves always at least weakly prefer full ambiguity to full disclosure, since they have nothing to gain from inspections (their actions will not depend on the arms inspector's report). However, there is a case, namely if (4) holds in Case 2, where hawks and doves *strictly* prefer ambiguity while some opportunistic types *strictly* prefer disclosure. In this case there is a conflict of interest among A's types about whether inspections are desirable or not. This conflict of interest might be alleviated if A could use messages to manipulate B's decision to allow inspections. We now investigate such equilibria.

4 Communication Equilibrium

Inspections generate information about B's strength. This information may be more or less useful, depending on A's type. The extreme types (hawks and doves) dislike arms proliferation, but they do not benefit from inspections *per se*, because they will not act on the information that is generated. The intermediate types are opportunistic and have a strong desire for inspections. This allows us to construct an equilibrium with two informative messages. The intermediate types send a "tough" message that leads to inspections, but also induces B to invest. The extreme types send a "conciliatory" message which does not lead to inspections, but also reduces the risk that B will invest. In the Appendix, we show that all equilibria with effective communication must have this form. In particular, the restriction to two messages is without loss of generality.

The communication equilibrium exists if and only if two conditions are satisfied. First, we must be in case 2 of Section 3.2. Otherwise, all of A's types would prefer ambiguity, and no-one would send the "tough" message. Thus, the first condition is that (11) must hold. Second, the prior probability that A is a hawk must be small. Otherwise, B would invest for sure after the conciliatory message, which is sent by both hawks and doves. In any case, it is intuitively clear that if A is likely to be a hawk then ambiguity will not prevent B from investing. Specifically, the second condition turns out to be

$$\frac{H}{H+D} < \kappa. \tag{13}$$

Proposition 5 *Suppose*

$$\frac{H}{H+D} < \kappa < 1 - F(a^*) \tag{14}$$

There is a communication equilibrium where, for some a' and a'', player A sends a "tough" message if $a \in (a', a'')$ and a "conciliatory" message otherwise. Player B is more likely to invest and more likely to allow inspections following the tough message.

Proof. Consider the following strategies. There is a' and a'', where $0 < a' < a'' < c - \tau b$, such that A sends the tough message if $a \in (a', a'')$ and the conciliatory message otherwise. Player B allows inspections if and only if he hears the tough message and is strong.

If A sends the tough message, then B invests, and B allows inspections if and only if he is strong. Since type $a \in (a', a'')$ is opportunistic, he will attack if inspections are refused or if inspections reveal that B is weak. If they reveal that B is strong, then he will not attack.

If A sends the conciliatory message, then B invests with probability $x \in (0, 1)$ and refuses inspections. If there is no inspection then A attacks if $a \geq a''$ but not if $a \leq a'$. If there is a "surprise inspection" which reveals that B is strong, then A attacks if and only if $a > c - b$ (this can be justified by the out-of-equilibrium belief that advanced weapons will end up in the hands of extremists). If the surprise inspection reveals that B is weak, then A attacks if and only $a > 0$.

Types a' and a'' are indifferent between the two messages. Suppose type a'' sends the tough message. With probability σ, player B is strong and A gets $-\tau b > a'' - c$. With probability $1 - \sigma$, player B is weak and A gets $a'' > 0$. Thus, the expected payoff is $(1 - \sigma)a'' - \sigma\tau b$. If type a'' sends the conciliatory message then B is strong with probability $x\sigma$. Type a'' will attack, and get expected payoff $a'' - x\sigma c$ (we verify later that attacking is optimal). For type a'' to be indifferent between the two messages, we must have

$$a'' = xc - \tau b < c - \tau b \tag{15}$$

If type a' sends the tough message, his expected payoff is $(1-\sigma)a' - \sigma\tau b$. If type a' sends the conciliatory message, he will not attack (we verify later

that this is optimal), and he gets expected payoff $-x\sigma\tau b$. For type a' to be indifferent, we must have

$$a' = \frac{(1-x)\sigma\tau b}{1-\sigma} > 0 \tag{16}$$

Define

$$x^* \equiv \frac{\tau b}{(1-\sigma)c + \sigma\tau b} < 1 \tag{17}$$

If $x = x^*$ then $a' = a'' = a^*$, as defined in (8). Now (15) and (16) imply that a'' is increasing in x and a' is decreasing in x. Thus, $a'' > a'$ if $x > x^*$. If $x \leq x^*$ then $a'' \leq a'$ which is impossible: player A's strategy only makes sense if $a'' > a'$. Thus, we must have $x > x^*$. Consider B's incentive to play according to his strategy. First, consider the decision to allow inspections. If he hears the tough message but is weak, then B thinks he will be attacked whether or not he allows inspections. He strictly prefers to refuse inspections to save the cost, ε. If he is strong then B's expected payoff from allowing inspections following the tough message is $-\varepsilon$, while his expected payoff from refusing is $-(\alpha - \gamma)$. He prefers to allow inspections as $\alpha - \gamma > \varepsilon$. Similarly, B strictly prefers to refuse inspections after the conciliatory message since inspections only increase the probability of an attack.

Next, consider the decision to invest. If B hears the tough message, then his expected payoff from investing is $\sigma(-\varepsilon)+(1-\sigma)(-\alpha)$. His expected payoff from not investing is $-\alpha$. Since $\alpha > \varepsilon$, he prefers to invest.

Now consider B's investment decision following the conciliatory message. If B hears the conciliatory message, then he thinks A will attack if $a \geq a''$ but not if $a \leq a'$. Accordingly, if B invests his expected payoff is

$$-\frac{1-F(a'')}{F(a')+1-F(a'')}(\alpha - \sigma\gamma) - k$$

If he does not invest, his expected payoff is

$$-\frac{1-F(a'')}{F(a')+1-F(a'')}\alpha$$

Player B must be indifferent between investing and not investing (since $0 < x < 1$), which is true if

$$(1-F(a'')+F(a'))\,\kappa - (1-F(a'')) = 0$$

We can use (15) and (16) to substitute for a' and a''. Define

$$\Psi(x) \equiv \left(1 - F(xc - \tau b) + F\left(\frac{(1-x)\sigma\tau b}{1-\sigma}\right)\right)\kappa - (1 - F(xc - \tau b)) \quad (18)$$

Notice that

$$\Psi(x^*) \equiv \kappa - (1 - F(a^*))$$

and

$$\Psi(1) \equiv (1 - F(c - \tau b) + F(0))\kappa - (1 - F(c - \tau b))$$

By definition, the equilibrium requires that the indifference condition $\Psi(x) = 0$ holds. Now, (14) is equivalent to $\Psi(x^*) < 0 < \Psi(1)$. By continuity, there is $x \in (x^*, 1)$ such that $\Psi(x) = 0$.

Notice that A's extreme types ($a < a'$ and $a > a''$) are less interested in inspections than the intermediate types. Since types a' and a'' are indifferent between the two messages, it is indeed optimal for the intermediate types to send the tough message, and for the extreme types to send the conciliatory message.

It remains to verify two assertions made above. First, it should not be optimal for type a' to send a conciliatory message and then attack. If type a' chooses such a strategy, then he gets

$$a' - x\sigma c = a' - \sigma(a'' + \tau b) < (1 - \sigma)a' - \sigma\tau b$$

where the equality uses (15), and the inequality is due to $a'' > a'$. The right hand side expression is what type a' gets in equilibrium.

Second, it should not be optimal for type a'' to send a tough message and then not attack. If type a'' chooses such a strategy, then he gets

$$-\sigma x\tau b = -\sigma\tau b + (1 - \sigma)a' < -\sigma\tau b + (1 - \sigma)a''$$

where the equality uses (16), and the inequality is due to $a'' > a'$. The right hand side expression is what type a'' gets in equilibrium. ■

In a previous paper (Baliga and Sjöström [1]), we considered a model where two symmetric players decide whether or not to arm themselves. The decisions were made simultaneously, without knowing the action taken by the opponent. We showed that if the players are motivated mainly by fear, then cheap-talk can dramatically reduce the probability of an arms race. The arms proliferation game of the current paper is different in one crucial regard: B

can reveal his strength before A makes a decision. With full disclosure, all of A's types will send whatever message minimizes the probability that B develops advanced weapons, so cheap-talk will be ineffective. Strategic ambiguity is therefore a necessary part of an equilibrium with effective communication. The communication equilibrium has the same "non-convex" structure as in Baliga and Sjöström [1]. Different types trade off "coordination" and "cooperation" at different rates. The intermediate types put a high value on coordination: they need information in order to make an optimal decision. The extreme types mainly want the opponent to cooperate (by not investing). If the prior probability that A is a hawk is low enough, specifically if $H/(H+D) < \kappa$, then the conciliatory message reduces B's fear and lowers the risk of arms proliferation.

By revealed preference, all of A's types weakly prefer the communication equilibrium to full disclosure (and there is strict preference for some). Consider B's payoff. With full disclosure, B's payoff is

$$-\sigma\left(1-F(c-\tau b)\right)(\alpha-\gamma)-(1-\sigma)\left(1-F(0)\right)\alpha-k.$$

If A sends the tough message in the communication equilibrium, then B's payoff is

$$-(1-\sigma)\,\alpha-k$$

If A sends the conciliatory message, then B's payoff is

$$-\frac{1-F(a'')}{F(a')+1-F(a'')}\left(\alpha-\sigma\gamma\right)-k=-\frac{1-F(a'')}{F(a')+1-F(a'')}\alpha.$$

Therefore, B's ex ante expected payoff in the communication equilibrium is

$$\begin{aligned} & -\left(1-F(a'')\right)(\alpha-\sigma\gamma)-\left(F(a'')-F(a')\right)(1-\sigma)\,\alpha-k \\ = \; & -\left(1-F(a'')\right)\sigma(\alpha-\gamma)-(1-\sigma)\alpha\left(1-F(a')\right)-k \end{aligned}$$

The difference in B's payoffs between full disclosure and communication equilibrium is

$$\sigma(\alpha-\gamma)[F(c-\tau b)-F(a'')]-(1-\sigma)\,\alpha[F(a')-F(0)]-[F(a'')-F(a')]\,\varepsilon. \quad (19)$$

The interpretation is similar to (12). The first term is positive and is due to the fact that there is a measure $F(c-\tau b)-F(a'')$ of "tough" opportunists who send the conciliatory message but then attack. However, there is a

chance that B is strong. Under full disclosure this would be revealed, and the tough opportunists would be deterred. The second term is negative and is due to the fact that there is a measure $F(a') - F(0)$ of "weak" opportunists who send a conciliatory message and then do not attack. However, there is a chance that B is weak. Under full disclosure this would be revealed, and the weak opportunists would attack. Again, B faces a trade-off: disclosure deters "tough" opportunists when B is strong, but ambiguity deters "weak" opportunists when B is weak. Without making further assumptions on the distribution F, we cannot sign the expression in (19).

Proposition 6 *All of A's types prefer the communication equilibrium to full disclosure. Player B prefers the communication equilibrium to full disclosure if and only if the expression (19) is negative.*

5 Conclusion

In policy debates, it is often argued that U.S. policy should be to eliminate ambiguity (e.g., Schrage [17]). Sobel [18] pointed out that ambiguity makes it more difficult to distinguish the weak from the strong, which protects the weak from being attacked. This suggests that ambiguity may be good for small powers (whose capabilities are uncertain) but not for big powers (who are known to be strong). However, we argue that once the small power's incentive to arm itself is taken into account, the opposite may be true. For ambiguity to be part of an equilibrium, the small power (B) must have an incentive to invest with positive probability, which means attacks must be sufficiently likely. Some of these attacks will be "mistakes": the leader of the big power (A) attacks even though he would have been deterred, had he known the small power's true strength. (In the equilibrium with full ambiguity and the communication equilibrium, this happens when $a \in (\tilde{a}, c - \tau b)$ and $a \in (a'', c - \tau b)$, respectively.) If such mistakes are very likely, then strategic ambiguity hurts the small power. We stress instead another positive aspect of ambiguity: the small power's incentive to arm itself is reduced by ambiguity. Therefore, strategic ambiguity tends to benefit the big power, at least if the leader is a type who will not make any "mistakes".

If A is an opportunistic type, then he needs information about B's true strength in order to avoid making mistakes. In a communication equilibrium, opportunistic types send "tough" messages, which can be interpreted as a

demand to sign the NPT. Player B responds by revealing his true strength. Dovish types instead send "conciliatory" messages. If B hears a conciliatory message, then he maintains a policy of ambiguity, but he is less likely to actually acquire advanced weapons. Unfortunately, hawks have an incentive to masquerade as doves and send a conciliatory message as well. Therefore, the nature of the equilibrium set depends on the relative likelihood of hawks and doves, $H/(H+D)$. If $H/(H+D)$ is too large then a conciliatory message will not reassure B, who suspects a "false dove", and communication will be ineffective.

A second determinant of the equilibrium set is the normalized cost of investing, κ. Recall that κ is higher the bigger is the cost k of investing; the smaller is the probability σ that B will acquire advanced weapons; and the smaller is the value of advanced weapons, γ. With ambiguity, the probability that B invests is decreasing in κ. If κ is small then B very likely will attempt to get advanced weapons, whether there is ambiguity or not. Thus, the smaller is κ, the more likely it is that the opportunistic type prefers inspections. But if κ is high then ambiguity makes A better off regardless of type. To see this effect in a special case, suppose c is very high, so a confrontation with a strong small power is very costly for the big power. As c increases, a^* approaches $\tau b\sigma/\left(1-\sigma\right)$, and (4) and (13) will both be satisfied. The crucial condition (11) becomes

$$\kappa < 1 - F(\frac{\tau b\sigma}{1-\sigma}) \tag{20}$$

If (20) holds, then some opportunistic types prefer inspections. In this case a communication equilibrium necessarily exists. Thus, there are two possibilities. If κ is small enough that (20) holds, then a communication equilibrium exists. This is the best equilibrium for A, regardless of type. However, if κ is big enough that (20) is violated, then A prefers ambiguity, regardless of type. In this case the full ambiguity equilibrium is the best for all of his types.

References

[1] Baliga, S. and T. Sjöström (2004), "Arms Races and Negotiations," *Review of Economic Studies* **71**:351-369.

[2] Bond, P. (2004): "Bank and Nonbank Financial Intermediation," *Journal of Finance* **59**: 2489-2530.

[3] Border, K. and J. Sobel (1987): "Samurai Accountant: A Theory of Auditing and Plunder," *Review of Economic Studies* **54**: 525-540.

[4] Cohen, A. (1998) *Israel and the Bomb* (New York City: Columbia University Press)

[5] Crawford, V. and J. Sobel (1982), "Strategic information transmission," *Econometrica* **50**: 1431-1451.

[6] Diamond, D. (1984): "Financial Intermediation and Delegated Monitoring," *Review of Economic Studies*, **51**: 393-414.

[7] Grossman, S. (1981) "The Informational Role of Warranties and Private Disclosure about Product Quality", *Journal of Law and Economics* **24**: 461-483

[8] Hecker, S. (2004) "Visit to the Yonugbyon Nuclear Scientific Research Center in North Korea", written statement to the Senate Committee on Foreign Relations

[9] Kreps, D. and R. Wilson (1982) "Sequential Equilibrium," *Econometrica* **50**: 863-894

[10] Milgrom, P. (1981) "Good News and Bad News: Representation Theorems and Applications," *Bell Journal of Economics* **12**: 380-391

[11] Mookherjee, D. and I. Png (1989): "Optimal Auditing, Insurance and Redistribution," *Quarterly Journal of Economics* **104**: 399-415.

[12] Norris, R.S., H. M. Kristensen and J. Handler (2002), "Pakistan's Nuclear Forces, 2001," *Bulletin of the Atomic Scientists* **58:** 70-71

[13] Norris, R.S. and H. M. Kristensen (2005), "North Korea's Nuclear Program, 2005," *Bulletin of the Atomic Scientists* **61:** 64-67

[14] Norris, R.S. and H. M. Kristensen (2005), "India's Nuclear Forces, 2005," *Bulletin of the Atomic Scientists* **61:** 73-75

[15] Pinkston and Diamond (2005), Special Report on the Shutdown of North Korea's 5MW(e) Nuclear Reactor, Mimeo, Monterrey Institute of International Studies

[16] Schelling, T.C. (1960) *The Strategy of Conflict* (Cambridge, MA: Harvard University Press)

[17] Schrage, M. (2003), "No Weapons, No Matter. We called Saddam's Bluff," Op-Ed in Washington Post, May 11, 2003.

[18] Sobel, J. (1992): "How and (when) to Communicate with Enemies," in *Equilibrium and Dynamics*, edited by M. Majumdar, Macmillan, pages 307-321.

[19] Takeyh, R. (2005), "WMD, Terrorism and Proliferation," Testimony Before Subcommittee on Prevention of Biological and Nuclear Attack, Committee on Homeland Security, September 8, 2005

[20] Thucydides (1972) *The History of the Peloponessan War* (London: Penguin Classics)

[21] Townsend, R. M. (1979): "Optimal contracts and costly state verification, *Journal of Economic Theory* **61**: 265-298.

[22] *Treaty on the Non-proliferation of Nuclear Weapons*, http://www.iaea.org/Publications/Documents/Treaties/npt.html

6 Appendix

In this appendix, we characterize the set of all cheap-talk equilibria. Player A sends a message $m \in M$, where M is an arbitrary message space. Of course, the labelling of messages is arbitrary. We will show that if the cost of inspections is small, then any equilibrium with effective communication must be *outcome-equivalent* to the communication equilibrium discussed in Section 4. That is, investment, inspections and attack decisions are identical (but messages may be re-labelled).

Proposition 7 *For $\bar{\varepsilon} > 0$ small enough, all equilibria with effective cheap-talk are outcome-equivalent to the communication equilibrium described in Section 4.*

Proof. Without loss of generality, we may assume in equilibrium each message in M is sent with positive probability by some type or set of types. Let

$x(m)$ be the probability that player B invests when player A sends $m \in M$. The set of messages that minimize $x(m)$ is denoted $M^c \subseteq M$. Let $M^t \equiv M \backslash M^c$. By definition, if $m^c \in M^c$ and $m^t \in M^t$, then $x(m^c) < x(m^t) \leq 1$. Let B's true strength be denoted $\omega \in \{s, w\}$, where s denotes that B is strong and w denotes that B is weak. Let $I(m, \omega, \varepsilon)$ be the probability that player B allows inspections following message m, when his strength is $\omega \in \{s, w\}$ and cost of inspection $\varepsilon \in [0, \bar{\varepsilon}]$. Conditional only on m and ω, the probability of inspections is

$$I(m, \omega) \equiv \int_0^{\bar{\varepsilon}} I(m^c, \omega, \varepsilon) h(\varepsilon) d\varepsilon$$

Notice that whether or not arms fall into the hands of extremists is payoff irrelevant for B. At time 4, either B strictly prefers to allow inspections for all ε, or he strictly prefers to refuse inspections for all ε, or there is some $\varepsilon^* \in [0, \bar{\varepsilon}]$ such that he prefers to allow inspections if and only if $\varepsilon \leq \varepsilon^*$. Therefore, the probability that B allows inspections does not depend on the risk that his advanced weapons will fall into the hands of extremists. This implies that, as long as A thinks B plays according to his equilibrium strategy, player A must believe that, conditional on B being strong, advanced weapons will fall into the hands of terrorists with probability τ. At time 5, player A plays a best response given his type and his beliefs. We leave out-of-equilibrium beliefs unspecified.

The proof will establish that, without loss of generality, we can let $M^c = \{m^c\}$ and $M^t = \{m^t\}$. Moreover, $I(m^t, s) = 1$, and $I(m^t, w) = I(m^c, w) = I(m^c, s) = 0$. This corresponds with the communication equilibrium from Section 4, if m^c is interpreted as the conciliatory message and m^t as the tough message. It is then easy to show that investment and attack decisions must also be the same. The proof has 12 steps.

Step 1: Doves and hawks will only send messages in M^c.

Proof: On the equilibrium path, hawks always attack and doves never attack. Therefore, a hawk's payoff is $a - \sigma x(m)c$ which is decreasing in $x(m)$. A dove's payoff is $-\sigma x(m)\tau b$ which is also decreasing in $x(m)$. Hence, doves and hawks will only send messages in M^c,

Step 2: $I(m^t, w) = 0$ for all $m^t \in M^t$.

Proof: By step 1, message $m^t \in M^t$ reveals that A must be an opportunist. Therefore, if B is weak, then he will refuse inspections following message $m^t \in M^t$. (If he were to allow inspections, he would pay $\varepsilon > 0$ only to be attacked for sure.)

Step 3: $I(m^t, s) = 1$ for all $m^t \in M^t$. If message $m^t \in M^t$ is sent and there is no inspection, then an attack must occur with probability one.

Proof: If $I(m^t, s) = I(m^t, w) = 0$ then no type would send m^t. Indeed, message $m^c \in M^c$ would lead to a strictly smaller probability that B invests, and perhaps also trigger an inspection (which can do no harm to A). Therefore, to induce some type of A to send m^t, we must have $I(m^t, s) > 0$ (since $I(m^t, w) = 0$ by step 2). Thus, inspections occur with positive probability following m^t. Suppose that an opportunistic type of A sends message $m^t \in M^t$. Step 2 implies that if, along the equilibrium path, inspections occur following $m^t \in M^t$, then they will surely reveal that B is strong, and the opportunistic type will not attack. If the opportunistic A's strategy also specifies that he does not attack when inspections are refused, then his expected payoff is $-\sigma\tau bx(m^t)$, since he never attacks. However, A would be strictly better off sending $m^c \in M^c$ and not attacking, since this yields $-\sigma\tau bx(m^c)$ where $x(m^c) < x(m^t)$. This argument shows that, if message $m^t \in M^t$ is sent and there is no inspection, then an attack must occur with probability one. However, an inspection that reveals that B is strong would deter the opportunistic types from attacking. Therefore, a strong player B strictly prefers to allow inspections.

Step 4: $I(m^c, s) < 1$ for all $m^c \in M^c$.

Proof: Since the only reason to send the message $m^t \in M^t$ is to improve the chance of an inspection, steps 2 and 3 imply that we must have $I(m^c, s) < 1$ for all $m^c \in M^c$.

Step 5: $I(m^c, w) = 0$ for all $m^c \in M^c$.

Proof: If message $m^c \in M^c$ is sent and inspections are refused, hawks attack but not doves. Clearly, if B hears message $m^c \in M^c$ and is weak, he has no reason to allow inspections. Indeed, inspections cannot convince the hawks not to attack, and if some opportunist also sends m^c, he will attack if inspections reveal that B is weak. Therefore, $I(m^c, w) = 0$.

Step 6: We can assume, without loss of generality, that there is only one message in M^t, so $M^t = \{m^t\}$.

Proof: Steps 1 and 3 imply that any type of player A who sends a message $m^t \in M^t$ must be an opportunist who attacks if and only if inspections reveal weakness or inspections are refused. Since $I(m^t, w) = 0$ and $I(m^t, s) = 1$ for all $m^t \in M^t$, $x(m^t)$ must also be the same for all $m^t \in M^t$ or else all types of A prefer the message in M^t that minimizes $x(m^t)$. Therefore, all messages in M^t lead to the same outcome, so we may as well assume there is only one such message.

Step 7: $0 < x(m^c) < 1$ for $m^c \in M^c$, so B must be indifferent between investing and not investing when he hears m^c.

Proof: If $x(m^c) = 0$ for $m^c \in M^c$, then no type of A would send $m^t \in M^t$, because an outcome where B invests with probability zero strictly dominates all other outcomes for A. However, we are assuming cheap-talk is effective, so some type of A must send a message in M^t (otherwise the probability that B invests would be independent of A's type). Thus, $x(m^c) = 0$ is impossible. On the other hand, we already know that $x(m^c) < x(m^t) \leq 1$ for $m^t \in M^t$.

Step 8: Type $a = 0$ strictly prefers to announce a message in M^c and not attack if player B refuses inspections.

Proof: If B is weak then type $a = 0$ gets a payoff of zero for sure. If along the equilibrium path inspections reveal that B is strong, then he will not attack and get $-\tau b$ (because $-\tau b > -c$). Accordingly, if type $a = 0$ sends $m^t \in M^t$, his payoff is $-\sigma x(m^t)\tau b$. If he sends $m^c \in M^c$ and does not attack when B refuses inspections, then his payoff is $-\sigma x(m^c)\tau b > -\sigma x(m^t)\tau b$. If he sends $m^c \in M^c$ and attacks if player B refuses inspections, his payoff is

$$\sigma x(m^c)\left(I(m^c, s)(-\tau b) + (1 - I(m, s))(-c)\right) < -\sigma x(m^c)\tau b$$

as $c > \tau b$ and $I(m^c, s) < 1$. Hence, type $a = 0$ strictly prefers to announce a message in M^c and not attack if player B refuses inspections.

Step 9: There are cut-off points a' and a'', where $0 < a' \leq a'' < c - \tau b$, such that A sends some $m^c \in M^c$ if $a \in (a', a'')$. He sends m^t if $a < a'$ or $a > a''$.

Proof: Step 8 established what type $a = 0$ strictly prefers to do. Some "weak" opportunists with a close to zero must have the same strict preference as type $a = 0$. Let $a' > 0$ be the supremum of all such types. All types such that $a < a'$ must send $m^c \in M^c$ and then not attack if there are no inspections. Similarly, it can be shown that "tough" opportunists with types just less than $c - \tau b$ must be sending messages in M^c and attacking if player B refuses inspections. Let a'' be the infimum of all such types. All types such that $a > a''$ must send m^c and then attack if there are no inspections. Necessarily, $0 < a' \leq a'' < c - \tau b$.

Step 10: For small enough $\bar{\varepsilon} > 0$, $I(m^c, s) = 0$ for all $m^c \in M^c$.

Proof: Suppose there is m^c such that $I(m^c, s) > 0$. Since $I(m^c, w) = 0$, all "tough" opportunists will send a message m^c which maximizes $I(m^c, s)$. Since $I(m^c, s) < 1$ by step 4, there must be ε such that B is indifferent between allowing and not allowing inspections when he is strong. However,

by allowing inspections he prevents an attack by "tough" opportunists. Since the cost of inspections is less that $\bar{\varepsilon}$, for arbitrarily small $\bar{\varepsilon}$, it must be the case that the conditional probability that A is a tough opportunist, given any m^c, is arbitrarily small. This means a'' must be very close to $c - \tau b$. That is, $a'' \approx c - \tau b$ for $\bar{\varepsilon}$ very small.

For type $a'' \approx c - \tau b$, the value of inspections is very low, since he is almost indifferent between attacking and not attacking a strong B, and strictly prefers attacking a weak B. Therefore, for type a'' to be indifferent between m^c and m^t, it must be the case that $x(m^c) \approx x(m^t)$. Now if type a' sends m^t, she will attack if and only if B is weak. But if she sends m^c, she will never attack. Since the probability that B is strong in the two cases is almost the same ($\sigma x(m^c) \approx \sigma x(m^t)$), to keep type a' indifferent between m^c and m^t we must have $a' \approx 0$.

Since $a'' \approx c - \tau b$ and $a' \approx 0$ for $\bar{\varepsilon}$ very small, there are almost no "tough" or "weak" opportunists. Therefore, among the types of A that send messages in M^c, a fraction approximately equal to $H/(H + D)$ are hawks, a fraction approximately equal to $D/(H + D)$ are doves, and only a vanishingly small fraction are opportunists. Hawks attack but doves don't. Integrating over messages sent in M^c, when $\bar{\varepsilon}$ is small, B's payoff from investing when he hears a message in M^c is approximately

$$-\frac{H}{H+D}(\alpha - \sigma\gamma) - k. \tag{21}$$

His payoff from not investing is approximately

$$-\frac{H}{H+D}\alpha. \tag{22}$$

This implies that he is strictly better off by investing if

$$\kappa < \frac{H}{H+D}$$

and strictly better off by not investing if the inequality is reversed. In either case we contradict step 7, which established that B must be indifferent between investing and not investing after each message $m^c \in M^c$. This contradiction shows that for small enough $\bar{\varepsilon}$, $I(m^c, s) > 0$ is impossible.

Step 11: We can assume, without loss of generality, that there is only one message in M^c, so $M^c = \{m^c\}$.

Proof: Since $I(m^c, w) = I(m^c, s) = 0$, all messages in M^c yields the same outcome, so we may assume there is only one such message.

Step 12: $x(m^t) = 1$. The cut-off types a' and a'' (defined in step 9) and the investment probability $x = x(m^c)$, are determined by equations (15), (16) and $\Psi(x) = 0$ (where Ψ is defined by (18)).

Proof: If message $m^t \in M^t$ is sent and B is weak, then he is attacked with probability one. Therefore, he prefers to invest when he hears message $m^t \in M^t$. It must be the case if $a' < a''$, a' and a'' are indifferent between reporting m^c and m^t. These indifference conditions yield (15) and (16). By step 7, B must be indifferent between investing and not investing when he hears m^c. The equation $\Psi(x) = 0$ is this indifference condition.

This proves that any equilibrium with effective cheap-talk is outcome-equivalent to the communication equilibrium described in Section 4. ■