Strategic Ambiguity and Arms Proliferation*

Sandeep Baliga   Tomas Sjöström
Northwestern University   Rutgers University

February 6, 2007

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. 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

Many countries have deliberately created ambiguity about their capabilities. The motivation of policy makers can be hard to decipher, especially in countries which lack democratic institutions and free press. However, analyses of countries such as Israel, North Korea and Iraq seem to suggest that a policy of ambiguity is meant to deter aggression without risking negative sanctions

---

1

and preemptive strikes. For example, Saddam Hussein possessed weapons of mass destruction in the early 1990's, but not in the late 1990's. He chose a policy of ambiguity in both situations. When revealing to his inner circle that Iraq had no WMDs, he "flatly rejected a suggestion that the regime remove all doubts to the contrary" fearing such a revelation would embolden his enemies to attack (Woods et al [31], p. 92 and Gordon and Trainor [10], p. 65).[1] Strategic ambiguity can help a country which *lacks* advanced weapons deter attacks, a policy which has been called "deterrence by doubt" (Gordon and Trainor [10], p. 65).[2] A country which *possesses* weapons of mass destruction may choose strategic ambiguity to avoid sanctions or preemptive strikes.

The conventional wisdom is that ambiguity is detrimental to world peace and, conversely, that arms inspections make us safer (e.g., Schrage [24]). This conventional wisdom is embodied in Article 3 of the *Treaty on the Non-proliferation of Nuclear Weapons,* known as the NPT [29]. The NPT requires that nations submit to inspection and verification of nuclear facilities by the IAEA. This 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, it is unclear if Pakistan and India have intercontinental ballistic missiles or are developing them (Norris and Kristensen [22], Norris, Kristensen and Handler [20]). Would the world be safer if these countries could be forced to allow arms inspections ?

Sobel [25] formalized the idea that ambiguity makes it difficult to distin-

---

[1]Israel has followed a similar policy: "While Israel prefers not to discuss the nuclear issue at all, the policy of intentional ambiguity is seen as a way of creating a deterrent, without making it explicit, a position that could invite sanctions or encourage an arms race in the Middle East." Greg Myre [19], *New York Times,* December 12, 2006. Cohen ([7], p. 244) documents Nasser's threat to attack Israel if he discovered a nuclear program. North Korea's Vice Minister of Foreign Affairs told visiting American scientists that their policy of ambiguity protected them against punitive actions: "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 [13]).

[2]Deterrence is clearly a major concern for many countries. "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" (Iranian newspaper editorial cited by Takeyh [26]). That the North Korean regime is motivated largely by fear of being attacked was conceded by South Korea's prime minister (Norris and Kristensen [21], p 67). Such fear can be caused by *uncertainty* about an opponent's true intentions. It does not have to be objectively justified in order to influence behavior.

guish the strong from the weak. In Sobel's model, "deterrence by doubt" can protect the weak from being attacked, while the strong prefer arms inspections. His argument relies on the assumption that a country's strength is exogenously given. Suppose instead weak countries can acquire advanced weapons at a cost, and these weapons can be used to defend against an attack. Ambiguity can only exist in equilibrium if the weak arm themselves (secretly) with positive probability. To give a weak country an incentive to arm, there must be a positive probability of war (or else secret weapons have no value). Since an equilibrium with ambiguity involves randomized strategies, ex post "mistakes" happen. A big power may attack a small power that turns out to be well armed, which is bad for both. The welfare analysis of ambiguity therefore becomes complex. Since strategic ambiguity is a substitute for actually acquiring WMDs, the equilibrium probability that the small power arms itself is smaller under ambiguity than under a regime of arms inspections. On the other hand, ambiguity leads to costly (equilibrium) mistakes, and a welfare analysis must trade-off these costs and benefits. We will show that 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. This tends to make a "mixed regime", where inspections happen with some probability, welfare optimal. Notice that, unlike in Sobel's model, the strong countries do not *necessarily* want to force arms inspections on the weak (say, by requiring them to sign the NPT). A weak country which is allowed to practise "deterrence by doubt" has fewer reasons to acquire WMDs, so *ambiguity reduces the risk of arms proliferation*, which benefits the strong country. Conversely, forcing a weak country to honor Article 3 of the NPT may increase its incentive to arm.

In our "arms proliferation game" there are two players, A and B, who are the leaders of countries 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 unarmed, 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 armed. Player A thinks there is a small probability that B is a "crazy" type who would share his weapons with terrorists. Player B's true type is *soft* (unverifiable) information. In Baliga and Sjöström [2], we studied how soft private information can trigger arms races and wars. In the current model, we consider whether revelation of *hard* information can promote peace and trust. Specifically, we assume weapons inspectors can verify whether or not B is armed (but not whether or not he is "crazy"). Player A's decision is

whether or not to attack B. The optimal decision depends on A's preferences (his type) and his beliefs. At most three types of equilibria can exist in this game. There is always an equilibrium with *full ambiguity* where B never allows arms inspections, and a *full disclosure* equilibrium where he always does. A more interesting possibility is a *communication equilibrium* involving informative cheap-talk.

With full disclosure, B's fear that A may attack him impels B to invest. With full ambiguity, B has less incentive to invest, which tends to make A better off. The only reason for A to want arms inspections in this case is if A is an *opportunistic type* who wants to attack B if and only if B is unarmed. (Perhaps A wants control of some resource or achieve "regime change" in country B.) Clearly, the opportunistic type benefits from information about B's true strength. Opportunistic types are the concern of classical deterrence theory. With full ambiguity, the probability that B invests is decreasing in the cost of investing and increasing in the value of advanced weapons in a conflict. If the cost of investing is low and/or the value of advanced weapons is large, then B is likely to invest even under full ambiguity. In this case, the opportunistic type prefers arms inspections. But if the cost is high and/or the value is small, then ambiguity makes A better off, regardless of type.

The fact that different types of player A can disagree about whether ambiguity is desirable suggests a role for cheap talk. Suppose player A can send either a "tough" or a "conciliatory" message. The conciliatory message encourages B to preserve ambiguity about his weapons capabilities, 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. In the communication equilibrium, A uses a "non-convex" strategy. If A is a peaceful "dove" or an aggressive "hawk", then he has an intrinsic preference for a particular action ("don't attack" for doves, "attack" for hawks). These extreme types do not need arms inspections in order to decide what to do, so they send the conciliatory message. It is the *intermediate* types who send the tough message. These intermediate types are opportunists who want to know if B is armed. (To be precise, in equilibrium some opportunists who are "almost" hawks or "almost" doves must also send the conciliatory message.) We show in the Appendix that any equilibrium where cheap-talk is effective in influencing B's investment decision can be taken to have two messages. This property follows from a key implication of incentive compatibility: a message which makes inspections more likely must

4

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 armed. 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.

Crawford and Sobel [8] and Green and Stokey [11] studied cheap talk games where only the receiver takes an action. In our model, both the sender (A) and the receiver (B) take actions, which gives rise to a non-convexity. When A's types are ordered in the natural way according to their propensity to attack, it is only the *intermediate* (opportunistic) types who have a *demand for information*. A similar non-convexity appears in Baliga and Sjöström [2]. In that paper, with probability close to one each player wants to avoid a conflict, but may be driven to attack out of *fear*. The paper showed how this fear could be eliminated by cheap-talk. Deterrence was not an issue. In the present paper, the key types are opportunistic types who attack out of *greed* if they know the opponent is unarmed. Preventing conflict is therefore a classic deterrence problem. Thus, the strategic logic here is quite different from the one in Baliga and Sjöström [2], although in both models incomplete information about the opponent's type plays a key role.

There is a related literature on financial intermediation and auditing, where costly inspections are used to verify incomes (Townsend [28], Diamond [9], Bond [5], Border and Sobel [6] and Mookherjee and Png [18]). 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 unarmed. 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 A (i.e., not to invest).

There exists an empirical literature which investigates the effect of concealed self-protective devices on crime (Ayres and Levitt [1], Lott and Mustard [16]). They argue that concealed self-protective devices generate a

5

positive externality for those who are unprotected, and hence will be undersupplied in equilibrium. We emphasize the *negative* externality B's self-protective weapons impose on A, and show that ambiguity may be welfare improving because it reduces the incentive for B to arm himself.

The paper proceeds as follows. In Section 2, we describe the model. In Section 3, we analyze equilibria without communication, where B either *always* or *never* allows inspections. In Section 4 we consider communication equilibria where A's message determines whether or not inspections occur. 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 capabilities by making an investment. His investment decision is binary: he either invests, or he doesn't invest. If B invests, he acquires weapons with probability $\sigma \in (0, 1)$. The cost of investing is $k > 0$. If B acquires advanced weapons then B is *armed*, otherwise B is *unarmed*. If B invests then he will be armed with probability σ, but if he doesn't invest then he is unarmed for sure. Player A cannot directly observe if B invests or is armed. However, "hard" information about B's weapons can be obtained 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 inspection publicly reveals whether B is armed or unarmed.

In the final stage of the game, A decides whether or not to attack B. If 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 is drawn from 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 armed with advanced weapons, then he can use them if A attacks. This yields an expected benefit $\gamma \in (0, \alpha)$ for B but an expected cost $c > 0$ for A. It is useful to define the *normalized cost of investing* to be

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

6

Player B has two possible types, "crazy" and "normal". His true type is his "soft" private information: while inspections can reveal if B is armed, they cannot reveal if he is crazy or normal. Player B's type is denoted $t \in \{z, n\}$ where z denotes crazy and n normal. Player A thinks the probability that B is crazy is τ. If B is armed but A does not attack, then A's and B's payoffs depend on B's type t: A suffers a cost d_t and B derives a benefit δ_t. The crazy type may share the advanced weapons with terrorists, or use them for some other purpose that could hurt A. Therefore, it is more costly for A if a crazy type obtains advanced weapons than if a normal type obtains them: $d_z > d_n$. We assume $d_n > 0$ because weapons proliferation could be costly to A even if B is not crazy (for example, terrorists may get hold of the technology even if B is normal). The advanced weapons are intrinsically more valuable to a crazy type than to a normal type: $\delta_z > \delta_n$. But they are more valuable in war than in peace: $\delta_z < \gamma$. We set $\delta_n = 0$ for simplicity (we would obtain the same results with $\delta_n > 0$). To summarize, we assume

$$0 < d_n < d_z \quad , \quad \delta_n = 0 < \delta_z < \gamma.$$

To simplify, we assume that if A attacks B then he eliminates the threat posed by B. (More generally, the threat could be reduced but not completely eliminated.)

The payoffs are summarized in the following matrix.

	B is armed	B is unarmed
A attacks	$a - c, \ -\alpha + \gamma$	$a, \ -\alpha$
No attack	$-d_t, \ \delta_t$	$0, \ 0$

This payoff matrix does not include B's cost of investment and the cost of 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 $-\alpha - k - \varepsilon$.

The solution concept is perfect Bayesian equilibrium. 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.

2.2   Time Line

The time line is as follows.

Time 0: Player A privately learns $a \in [a_0, a_1]$ and player B privately learns $t \in \{z, n\}$.

Time 1: Cheap-talk stage.

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

Time 3: If B invested, then he privately learns whether or not he has acquired weapons.

Time 4: The cost ε is realized. Player B decides whether or not to allow inspections. If inspections take place, then the inspectors publicly reveal whether or not B is armed.

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

2.3   Parameter Restrictions

Player A is a *dove* if $a < 0$, an *opportunist* if $0 < a < c - \left(\tau d_z + (1 - \tau) d_n \right)$, and a *hawk* if $a > c - \left(\tau d_z + (1 - \tau) d_n \right)$. The probability that A is a dove is $D \equiv F(0) > 0$. The probability that he is a hawk is $H \equiv 1 - F(c - \left(\tau d_z + (1 - \tau) d_n \right)) > 0$.

Assumption 1: $\tau d_z + (1 - \tau) d_n < c < d_z$.

The first inequality in Assumption 1 guarantees that opportunistic types can exist. Without opportunistic types, the whole problem of deterrence would be moot. Notice that the first inequality can be rewritten as

$$a > a - c - (- \left(\tau d_z + (1 - \tau) d_n \right)). \tag{1}$$

The left hand side of (1) is A's gain from attacking an unarmed B. The right hand side of (1) is A's *net* benefit of attacking an armed B if B is crazy with probability τ (i.e., the payoff from attacking *minus* the expected payoff from not attacking). The inequality implies that *given* A's prior τ, A is *less* inclined to attack when B is armed than when B is unarmed. This is why deterrence can be effective. Now the net benefit from attacking an armed B who is thought to be crazy for sure is $a - c - (-d_z)$. The second inequality of Assumption 1 says that $a < a - c - (-d_z)$. Thus, if A is *convinced* that B is crazy then A becomes *more* inclined to attack if he finds out that B is armed. If this 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

8

[25]. To support an equilibrium with ambiguity, A's disutility from advanced weapons in the hands of crazy types must be sufficiently big.[3]

Our second assumption guarantees that the cost of inspections is small enough so that it does not significantly influence the set of equilibria.

Assumption 2:

$$\bar{\varepsilon} \; < \; \min\{(F(c - (\tau d_z + (1 - \tau) d_n)) - F(c - d_z))(\alpha - \gamma - \delta_z),$$
$$(F(0) - F(c - d_z))\alpha, \; F(0) - F(c - b)\}$$

Our final assumption ensures that the cost of investing is small enough, so that A cannot achieve his "bliss point".

Assumption 3:

$$\frac{k}{\sigma} < (1 - F(c - d_z))(-\alpha + \gamma) + (1 - F(0))\alpha$$

If Assumption 3 were violated, then the cost of investing would be so high that there would be an equilibrium where a normal type of B never invests and always allows arms inspections. This case would not be very interesting. However, many results, such as the structure and existence of communication equilibrium, do not rely on Assumption 3.

We end this section with a useful preliminary result (the proof is in the Appendix):

Proposition 1 *In any perfect Bayesian equilibrium, the crazy type of player B invests with probability one.*

---

[3]Nuclear weapons in a "rouge nation" may not have the range to reach the United States, but in the hands of terrorists they could destabilize world security. Assumption 1 captures a scenario where the benefit from eliminating this threat can be positive or negative, depending on the assessment of the "type" of the "rouge leader".

3 Equilibria Without Communication

3.1 Equilibria with Full Disclosure

In an equilibrium with *full disclosure*, there is never any ambiguity about B's weapons on the equilibrium path. For example, an inspection may occur with probability one, or B may allow inspections if and only if he is armed (in which case a refusal to allow inspections reveals that B is unarmed). With full disclosure, 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's type. Without ambiguity, communication cannot prevent arms proliferation. (Since B has no incentive to reveal that he is crazy, we can without loss of generality assume that B sends no message.)

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

Proof. Suppose in equilibrium, B refrains from investing. With full disclosure, an unarmed B is attacked whenever $a \geq 0$, which happens with probability $1 - F(0)$. Consider a deviation where B invests and refuses inspections if successful, which triggers an attack with *at most* probability $1 - F(c - d_z)$ (because A will never attack if $a < c - d_z$). The gain from this deviation is *at least*

$$\sigma \left\{ (1 - F(c - d_z)) (-\alpha + \gamma) + \sigma F(c - d_z)\delta_t - (1 - F(0))(-\alpha) \right\} - k$$

This expression is strictly positive, by Assumption 3. Therefore, in any full disclosure equilibrium, player B invests with probability one.

It remains to show that an equilibrium with full disclosure exists. Let the equilibrium strategy specify that player B invests with probability one, and he allows inspections if and only if he is armed. If inspections reveal that B is armed, then A attacks if $a - c > \tau d_z + (1 - \tau) d_n$ (A thinks B is crazy with probability τ since both types are armed with probability σ). If B refuses inspections, then A infers that B is unarmed, so A attacks if $a \geq 0$. If B should allow inspections even though he is unarmed, he is still attacked if $a \geq 0$, so he has no reason to allow inspections in this case. Suppose B

deviates by refusing inspections when he is armed. This raises the probability of attack from $1 - F\left(c - \left(\tau d_z + (1 - \tau)\,d_n\right)\right)$ to $1 - F(0)$, which has a cost

$$\left(F\left(c - \left(\tau d_z + (1 - \tau)\,d_n\right)\right) - F(0)\right)\left(\alpha - \gamma - \delta_t\right)$$

The gain from the deviation is only ε. Assumption 2 guarantees that the cost exceeds the benefit, so B prefers to reveal if he is armed. Finally, given full disclosure, Assumption 3 guarantees that B prefers to invest. ∎

3.2   Equilibria with Full Ambiguity

Proposition 2 implies that B prefers to invest unless there is some ambiguity about his capabilities. In an equilibrium with *full ambiguity*, inspections never occur on the equilibrium path. Clearly, with full ambiguity communication cannot be effective. For ambiguity to deter A from attacking, the normal type of B must invest with sufficiently high probability. However, this requires that A attacks with sufficiently high probability, or else the normal type has no incentive to invest. The required equilibrium probabilities will depend on the normalized cost of investing, κ.

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

$$\kappa > 1 - F(\sigma\left(c - \tau d_z - (1 - \tau)d_n\right)), \tag{2}$$

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

Proof. Proposition 1 implies that player B's crazy type always invests. If B never allows inspections, then all of A's types simply want to minimize the probability that B's normal type invests. Therefore, the probability of investment must be independent of A's type. Let $\tilde{x}$ denote the probability that the normal type of B invests. Thus, B is armed with probability $\sigma\left(\tau + (1 - \tau)\tilde{x}\right)$.

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. Type $\tilde{a}$ expects $\tilde{a} - \sigma(\tau + (1 - \tau)\tilde{x})c$

by attacking, and $-\sigma\left(\tau d_z + (1-\tau)\tilde{x}d_n\right)$ by not attacking. The cut-off property implies that type $\tilde{a}$ must be indifferent between his two actions. This indifference condition holds if

$$\tilde{a} = \sigma\tau(c - d_z) + \sigma(1-\tau)\tilde{x}\,(c - d_n) \qquad (3)$$

If B deviates by allowing inspections, and he is found to be unarmed, then A attacks if and only if $a > 0$. But if B is found to be armed, then we may suppose A attacks if and only if $a > c - d_z$. This is supported by the off–the-equilibrium path belief that B is crazy (which is the belief most likely to support the equilibrium, since it punishes B's deviation most strictly). We will show that $\tilde{a} > 0$, so inspections always increase the probability of attack. This clearly implies that B prefers to refuse inspections.

If $0 < \tilde{x} < 1$ then B's normal type 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

$$-(1 - F(\tilde{a}))\alpha = -(1 - F(\tilde{a}))(\alpha - \sigma\gamma) - k \qquad (4)$$

which is the same as

$$\kappa - (1 - F(\tilde{a})) = 0 \qquad (5)$$

If $\kappa > 1 - F(\tilde{a})$ then the normal type's unique best response is not to invest, so $\tilde{x} = 0$. Similarly, if $\kappa < 1 - F(\tilde{a})$ then the unique best response implies $\tilde{x} = 1$. Define

$$\Gamma(x) \equiv \kappa - (1 - F(\sigma\tau(c - d_z) + \sigma(1-\tau)x\,(c - d_n)))$$

An equilibrium where $0 < \tilde{x} < 1$ requires that both (5)and (3) hold, which implies $\Gamma(\tilde{x}) = 0$. Now,

$$\frac{k}{\sigma} < (F(0)-F(c-d_z))\,(-\alpha)+(1-F(c-d_z))\gamma < (1-F(0))\gamma < (1 - F(\sigma\tau\,(c - d_z)))\,\gamma \qquad (6)$$

which implies $\kappa < 1 - F(\sigma\tau\,(c - d_z))$. (The first inequality in (6) follows from Assumption 3, the second follows from $c - d_z < 0$ and $\alpha > \gamma$, the third follows from $\sigma\tau(c - d_z) < 0$.) Therefore, $\Gamma(0) < 0$. 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 (2). Thus, there are two possible cases.

Case (i): (2) holds. In this case, 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 the normal type invests with probability 0 or 1). It is indeed an equilibrium, because $\tilde{a} > 0$. This follows from

$$\kappa - (1 - F(0)) < 0 = \kappa - (1 - F(\tilde{a})).$$

where the first inequality is due to (6) and the second to (5).

Case (ii): (2) is violated. In this case, $\Gamma(1) \leq 0$ so we must have $\tilde{x} = 1$. It is indeed an equilibrium, because (3) and Assumption 1 yield

$$\tilde{a} = \sigma(c - \tau d_z - (1 - \tau)d_n) > 0.$$

■

In the proof of Proposition 3, equilibrium is supported by the belief that B is crazy if inspections revealed that he is armed. These out-of-equilibrium beliefs qualify as reasonable according to standard arguments, such as the D1 criterion of Banks and Sobel [3]. Indeed, the equilibrium payoff of an armed normal type at the inspection stage is $(-\alpha + \gamma)(1 - F(\tilde{a}))$. If he allows inspections, and A attacks with probability p, then his expected payoff is $(-\alpha + \gamma)p - \varepsilon$. Thus, inspections would make the armed normal type weakly better off if and only if

$$(-\alpha + \gamma)p - \varepsilon \geq (-\alpha + \gamma)(1 - F(\tilde{a})). \tag{7}$$

Similarly, inspections would make the armed crazy type strictly better off if and only if

$$(-\alpha + \gamma)p + (1 - p)\delta_z - \varepsilon > (-\alpha + \gamma)(1 - F(\tilde{a})). \tag{8}$$

Since (7) implies (8), the D1 criterion suggests that the crazy type should be assigned probability one if out-of-equilibrium inspections reveal B is armed.

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

$$a^* \equiv \frac{\sigma d_n (c - \tau d_z - (1 - \tau) d_n)}{(1 - \sigma)c + \sigma d_n} \tag{9}$$

Assumption 1 implies

$$0 < a^* < \sigma(c - (\tau d_z + (1 - \tau) d_n)), \tag{10}$$

13

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:

$$\kappa > 1 - F(a^*), \tag{11}$$

where a^* is defined by (9). The inequalities (10) and (11) imply that (2) holds, so B refrains from investing with probability $1 - \tilde{x} > 0$ behind the veil of ambiguity. 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 $\tilde{a}$, defined by . 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 $\tilde{a} - \sigma (\tau + (1 - \tau) \tilde{x}) c$. Compare this to the outcome where B always invests and allows inspections. After the inspection, type $\tilde{a}$ attacks if and only if B is unarmed, which happens with probability $1 - \sigma$ (from (3), type $\tilde{a}$ is an opportunist). Thus, type $\tilde{a}$'s expected payoff would be

$$(1 - \sigma)\tilde{a} - \sigma (\tau d_z + (1 - \tau) d_n) .$$

Type $\tilde{a}$ prefers full ambiguity if and only if

$$(1 - \sigma)\tilde{a} - \sigma (\tau d_z + (1 - \tau) d_n) < \tilde{a} - \sigma (\tau + (1 - \tau) \tilde{x}) c \tag{12}$$

Using the definition of $\tilde{a}$, (12) is equivalent to $\tilde{x} < x^*$, where

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

The first inequality in Assumption 1 implies $x^* < 1$. Clearly, $\tilde{x} < x^*$ if $\tilde{x} = 0$. Suppose instead that $\tilde{x} > 0$. Since $\Gamma(0) < 0 = \Gamma(\tilde{x}) < \Gamma(1)$ and $\Gamma'(x) > 0$, we have $\tilde{x} < x^*$ if and only if $\Gamma(x^*) > 0$, which is equivalent to (11). 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:

$$\kappa < 1 - F(a^*). \tag{13}$$

If (2) holds, then B invests with probability $\tilde{x} < 1$. Therefore, hawks and doves strictly prefer full ambiguity to full disclosure. However, by a

14

similar reasoning as in case 1, we find that (13) implies that type $\tilde{a}$ strictly prefers full disclosure to full ambiguity (inequality (12) is reversed). If (2) 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 when $a \geq \tilde{a}$. With full disclosure, player A attacks if $a \geq c - \tau d_z - (1 - \tau) d_n$ when B is armed, and if $a \geq 0$ when B is unarmed. Therefore, when moving from full ambiguity to full disclosure, player B's expected utility (not including the cost of inspection) changes by an amount

$$\sigma (\alpha - \gamma - \delta_t) \left[F(c - \tau d_z - (1 - \tau) d_n) - F(\tilde{a}) \right] - (1 - \sigma) \alpha \left[F(\tilde{a}) - F(0) \right]. \tag{14}$$

The first term is positive. This term is due to the fact that with full ambiguity, a measure $F(c - \tau d_z - (1 - \tau) d_n) - F(\tilde{a})$ of "tough" opportunists attack B when he is armed. (Under full disclosure, B's weapons would be revealed and the tough opportunists would be deterred.) The second term is negative. It is due to the fact that with full ambiguity, a measure $F(\tilde{a}) - F(0)$ of "weak" opportunists do not attack B when he is unarmed. (Under full disclosure, the weak opportunists would attack the unarmed B.) Thus, disclosure deters "tough" opportunists when B is armed, but ambiguity deters "weak" opportunists when B is unarmed. Without making further assumptions on the distribution of A's types we cannot sign the expression in (14).

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 (11) holds. Player B prefers full ambiguity to full disclosure if and only if the expression in (14) 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 (2) 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

15

are desirable or not. In order to alleviate this conflict of interest, B's decision to allow inspections should be made to depend on A's type.

Remark The key to maintaining ambiguity in equilibrium is that B, whether he is armed or not, increases the probability of attack by allowing inspections. In our model, this property holds when B is unarmed because, by revealing this fact, B invites attack from opportunists who are otherwise "deterred by doubt". When B is armed, the property holds because B invites attack not only from opportunists but also from doves who fear he may be crazy. Other natural models would generate this property in alternative ways. For example, suppose A's type specifies not only his level of aggression a, but also his cost of fighting, high (c_H) or low (c_L). Assume that $c_L < \tau d_z + (1 - \tau) d_n < c_H$. When A has a low cost of fighting, he will attack an armed B even if he is a dove, as long as $c_L - \tau d_z + (1 - \tau) d_n \leq a \leq 0$. Such a "tough dove" prefers to preempt weapons proliferation by fighting an armed B, but he does not want to attack an unarmed B. As in our simpler model, if an unarmed B allows inspections, it can increase the probability of attack by removing "deterrence by doubt". But an armed B fears that revealing his weapons would invite attack from tough doves. Here attacking would be optimal for the tough dove, even if he thinks the probability that B is crazy conditional of revealing weapons is equal to the probability conditional on refusing inspections. Still, the tough dove would not attack without inspections, because they do not want to attack an unarmed opponent. Hence, the probability that an armed B is attacked can be raised if he discloses his weapons status, whatever it may be. In our model, we derive this property in a simpler way: some doves attack because they think an armed B who allows an inspection must be crazy, and this belief satisfies the standard refinement. Accordingly, there seems to be no benefit from making the model more complex than it is.

4   Communication Equilibrium

All the equilibria we have studied so far involve no interesting role for communication. We turn now to cheap-talk equilibria which are *effective*, i.e., the probability that B invests depends on A's type.

Inspections generate information about B's capabilities. 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 their optimal action depends on whether B is armed or not. Hence, they are willing to trade-off a higher probability of arms proliferation for a higher probability of inspections. This suggests that communication equilibria will involve just two informative messages. The intermediate types send a "tough" message which reveals they are opportunists. The tough message leads to inspections with probability one, but also induces B to invest with probability one. The extreme types send a "conciliatory" message which lead to inspections with probability strictly less than one, but also induces B to invest with probability strictly less than one. Thus, effective cheap-talk should involve just two messages, trading off a higher probability of investment against a higher probability of inspections. The next proposition confirms that this intuition is correct. If more than two messages are sent in some equilibrium, then the equilibrium messages can be divided into two classes, where all messages within a class lead to the same probability distribution over outcomes. Hence, any equilibrium can be replicated by a two-message equilibrium.

Proposition 5 *All equilibria with effective cheap-talk are equivalent to a two message equilibrium where, for some a' and a'' with $a'' > a' > 0$, player A sends a "tough" message if $a \in (a', a'')$ and a "conciliatory" message otherwise. After the tough message, player B invest with probability 1 and allows inspections iff he is armed. After the conciliatory message, the normal type of B arms with positive probability and refuses inspections with positive probability if he is armed.*

We sketch the proof of Proposition 5 here (the rigorous proof is in the Appendix). Let M be some arbitrary message space, and consider any equilibrium of the game. Let M^C be the set of "conciliatory" messages that minimize the probability that B invests. Let $M^T = M \backslash M^C$ be the set of "tough" messages. Any type of A who either *always* or *never* attacks in equilibrium must send a message in M^C, since all he cares about is reducing the probability that B invests. If cheap talk is effective in influencing B's decision to invest, then some types must send a message in M^T, and they must all be opportunists. They do not attack if inspections reveal B is armed, but attack otherwise. Therefore, any message $m^T \in M^T$ will cause B to invest and reveal any weapons he acquires. As all messages in M^T lead to the same

outcome, we can assume M^T is a singleton. Furthermore, B must invest with positive probability in response to any message $m^C \in M^C$, and must refuse inspections with positive probability when he has weapons (otherwise, all types of A would prefer to send a message in M^C). The proof is finished by showing that, in fact, all messages in M^C must cause B to invest with the same probability and refuse inspections with the same probability. Hence we can assume M^C is also a singleton.

All of A's types weakly prefer the communication equilibrium to full disclosure, and there is strict preference for some. Indeed, any type of A can induce the same equilibrium path as in the full disclosure equilibrium by sending the tough message. Any type that does *not* do so and sends the conciliatory message must prefer the communication equilibrium with effective cheap-talk. Hence, communication can broaden the set of circumstances where allowing some ambiguity improves the welfare of all of A's types. Player B faces a similar trade-off as in the full ambiguity equilibrium. Sometimes, A attacks after sending the conciliatory message when he would have been deterred by knowing B's true arms capabilities.

We will now construct a communication equilibrium. It 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 (13) 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. 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

$$\frac{H}{H+D} < \kappa. \tag{15}$$

Proposition 6 *Suppose*

$$\frac{H}{H+D} < \kappa < 1 - F(a^*) \tag{16}$$

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.

18

Proof. Consider the following strategies. There is a' and a'', where $0 < a' < a'' < c - \tau d_z - (1-\tau) d_n$, 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 armed. Player B's crazy type invests with probability one, regardless of message.

If A sends the tough message, then B's normal type invests with probability 1, and allow inspections if and only if the investment is successful. Since type $a \in (a', a'')$ who sends the tough message is opportunistic, he will attack if inspections are refused or if inspections reveal that B is unarmed. If they reveal that B is armed, then he will not attack.

If A sends the conciliatory message, then the normal type of B invests with probability $x \in (0, 1)$ and refuses inspections. If there is no inspection following the conciliatory message, then A attacks if $a \geq a''$ but not if $a \leq a'$. If there is a "surprise inspection" which reveals that B is armed, then A attacks if and only if $a > c - d_z$. This is justified by the out-of-equilibrium belief that B is a crazy type (which is consistent with the D1 criterion). If the surprise inspection reveals that B is unarmed, then A attacks if and only $a > 0$.

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

$$a'' = (\tau + (1-\tau)x)c - \tau d_z - (1-\tau)d_n < c - \tau d_z - (1-\tau)d_n \qquad (17)$$

If type a' sends the tough message, his expected payoff is $(1-\sigma)a' - \sigma\tau d_z - \sigma(1-\tau)d_n$. If type a' sends the conciliatory message, he will not attack (we verify later that this is optimal), and he gets expected payoff $-\sigma\tau d_z - \sigma(1-\tau)xd_n$. For type a' to be indifferent, we must have

$$a' = \frac{(1-x)\sigma(1-\tau)d_n}{1-\sigma} > 0 \qquad (18)$$

Define

$$x^* \equiv \frac{(1-\sigma)\tau(d_z - c) + (1-\tau)d_n}{(1-\sigma)(1-\tau)c + \sigma(1-\tau)d_n} < 1 \qquad (19)$$

19

where the inequality follows from the first inequality in Assumption 1. If $x = x^*$ is substituted into (17) and (18), we get $a' = a'' = a^*$, as defined in (9). Now (17) and (18) imply that a'' is increasing in x and a' is decreasing in x. Thus, $a'' > a'$ as long as $x > x^*$.

We now verify B's incentive to play according to his strategy. First, consider the decision to allow inspections. If he hears the tough message but is unarmed, then B realizes he will be attacked whether or not he allows inspections. He strictly prefers to refuse inspections to save the cost ε. If B is armed then his expected payoff from allowing inspections following the tough message is $\delta_t - \varepsilon$, while his expected payoff from refusing is $-(\alpha - \gamma)$. He prefers to allow inspections as $\delta_t - \varepsilon > -(\alpha - \gamma)$ by Assumption 2. Similarly, B strictly prefers to refuse inspections after the conciliatory message since inspections only increase the probability of an attack.

Next, consider the normal type's 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 the normal type'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's normal type must be indifferent between investing and not investing (since $0 < x < 1$), which is true if

$$(1 - F(a'') + F(a'))\kappa - (1 - F(a'')) = 0 \tag{20}$$

We can use (17) and (18) to substitute for a' and a'' in (20). Define

$$\Psi(x) \equiv \left(1 - F((\tau + (1 - \tau)x)c - \tau d_z - (1 - \tau)d_n) + F\left(\frac{(1 - x)\sigma(1 - \tau)d_n}{1 - \sigma}\right)\right)\kappa$$
$$-(1 - F((\tau + (1 - \tau)x)c - \tau d_z - (1 - \tau)d_n))$$

Notice that

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

and

$$\Psi(1) = (1 - F(c - \tau d_z - (1 - \tau) d_n) + F(0))\kappa - (1 - F(c - \tau d_z - (1 - \tau) d_n))$$

The indifference condition (20) is verified, together with (17) and (18), if x is chosen such that $\Psi(x) = 0$. Now, (16) is equivalent to $\Psi(x^*) < 0 < \Psi(1)$. By continuity, there is $x \in (x^*, 1)$ such that $\Psi(x) = 0$. Since $x > x^*$, $a'' > a'$.

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. Also, since B's normal type always weakly prefers to invest, it is optimal for the crazy type to always invest.

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

$$a' - \sigma \left(\tau + (1 - \tau) x\right) c = a' - \sigma(a'' + \sigma\tau d_z + \sigma (1 - \tau) d_n) < (1 - \sigma)a' - \sigma(\tau d_z + (1 - \tau) d_n)$$

where the equality uses (17), 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 conciliatory message and then not attack. If type a'' chooses such a strategy, then he gets

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

where the equality uses (18), and the inequality is due to $a'' > a'$. The right hand side expression is what type a'' gets in equilibrium. ∎

The communication equilibrium has the same "non-convex" structure as the cheap-talk equilibrium in Baliga and Sjöström [2]. 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). Therefore, we can separate the extreme types from the intermediate types. 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 of A, and lowers the risk of arms proliferation. However, in Baliga and Sjöström [2], all decisions were made *simultaneously* so the issue of strategic ambiguity did not arise. Since the current model allows the

possibility of arms inspections taking place before A decides to attack, we can study the costs and benefits of strategic ambiguity. As we have shown, strategic ambiguity is required for effective communication.

As noted, all of A's types weakly prefer the communication equilibrium to full disclosure (and there is strict preference for some). Consider B's pay-off. With full disclosure, if B acquires weapons then he is attacked with probability $1 - F(c - \tau d_z - (1 - \tau) d_n)$, otherwise he is attacked with probability $1 - F(0)$. In the communication equilibrium, if A sends the tough message then B is attacked if and only if he has no weapons. If A sends the conciliatory message, then B is attacked with probability

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

If we move from the communications equilibrium to full disclosure, B's expected payoff changes by

$$\sigma(\alpha - \gamma + \delta_t) \left[F(c - \tau d_z - (1 - \tau) d_n) - F(a'') \right] - \qquad (21)$$
$$(1 - \sigma) \left[F(a') - F(0) \right] \alpha - \left[1 - (F(a'') - F(a')) \right] E\varepsilon.$$

The interpretation is similar to (14). The first term is positive and is due to the fact that there is a measure $F(c - \tau d_z - (1 - \tau) d_n) - F(a'')$ of "tough" opportunists who send the conciliatory message but then attack B, even though B may be armed. (Under full disclosure, the tough opportunists are deterred by B's weapons.) 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, even though B may be un-armed. (Under full disclosure, the weak opportunists attack whenever B is unarmed.) Again, disclosure deters "tough" opportunists when B is armed, but ambiguity deters "weak" opportunists when B is unarmed. Without making further assumptions on the distribution F, we cannot sign the expression in (21). But, as $\delta_z > \delta_n = 0$, if the crazy type of B prefers the communication equilibrium to full disclosure then so does the normal type.

Proposition 7 *All of A's types prefer the communication equilibrium studied in Proposition 6 to full disclosure. Player B faces a trade-off. Both types of B prefer the communication equilibrium to full disclosure if and only if the expression (21) is negative for the crazy type.*

There may be other communication equilibria which exist under a broader set of circumstances than the one in Proposition 6. But they all share the same qualitative properties: they must be equivalent to two message equilibria, A must prefer the communication equilibrium to the full disclosure equilibrium and B faces a trade-off.

5   Concluding Comments

In policy debates, it is often argued that U.S. policy should be to eliminate ambiguity (e.g., Schrage [24]). Sobel [25] 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 help weak nations, but be harmful to strong nations. In this paper, we have shown that if a weak nations's incentive to acquire weapons 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 capabilities. If such mistakes are very likely, then strategic ambiguity hurts the small power. We stress instead another positive aspect of ambiguity: it reduces the small power's incentive to acquire weapons. Accordingly, strategic ambiguity tends to benefit the big power.

If A is an opportunistic type, then he wants information about B's true capabilities in order to avoid costly mistakes. In a communication equilibrium, opportunistic types send "tough" messages. This can be interpreted as demands that B signs the NPT. Player B optimally responds by revealing his true capabilities. 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.

We have analyzed a game with two players. In reality, several big powers with divergent interests may interact with the same small power. The communication stage then becomes a game with multiple senders and one receiver. Such games have been analyzed by Battaglini [4] and Krishna and Morgan [15]. A dovish big power (that would never in any circumstance attack) would strive to reduce the likelihood of arms proliferation, while a more opportunistic big power may instead try to force inspections. Our model provides some insights into these many-player interactions, but a complete analysis is left for future work.

References

[1] Ayres, I. and S. Levitt (1998): "Measuring Positive Externalities from Unobservable Victim Precaution: An Empirical Analysis of Lojack," *Quarterly Journal of Economics* **113**: 43-77.

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

[3] Banks, J. and J. Sobel (1987), "Equilibrium Selection in Signaling Games," *Econometrica* **55**: 647-662.

[4] Battaglini, M. (2002): "Multiple Referrals and Multidimensional Cheap Talk," *Econometrica*, **70**: 1379-1401.

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

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

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

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

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

[10] Gordon, M. and B. Trainor (2006): *Cobra II: The Inside Story of the Invasion and Occupation of Iraq*, (New York: pantheon Books).

[11] Green, J. and N. Stokey (1980, 2007): "A Two-Person Game of Information Transmission," *Journal of Economic Theory*, forthcoming.

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

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

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

[15] Krishna, V. and J. Morgan (2001): "A Model of Expertise," *Quarterly Journal of Economics*, **116**: 747-775.

[16] Lott, J. and D. Mustard (1997): "Right-to-Carry Concealed Weapons and the Importance of Deterrence," *Journal of Legal Studies* **26**:1-68

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

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

[19] Myre, Greg: "In a Slip, Israel's Leader Seems to Confirm Its Nuclear Arsenal," *New York Times,* December 12, 2006.

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

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

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

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

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

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

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

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

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

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

[30] Woods K. M., J, Lacey and W. Murray (2006): "Saddam's Delusions: The View from the Inside," *Foreign Affairs*, **85**: 2-26.

[31] Woods K. M. with M. R. Peese, M. E. Stout, W. Murray and J.G. Lacey (2006): "Iraqi Perspectives Project," mimeo, Joint Center for Operational Analysis.

6   Appendix

This appendix contains the proofs of Propositions 1 and 5. We restate the Propositions for completeness.

Proposition 1 *In any perfect Bayesian equilibrium, the crazy type of player B invests with probability one.*

Proof. Consider any perfect Bayesian equilibrium. Let $x(m, t)$ denote the probability that type $t \in \{z, n\}$ invests after player A has sent message m. If $x(m, n) > 0$ for some $m \in M$ then $x(m, z) = 1$. This follows from $\delta_z > \delta_n$, which makes the crazy type strictly more willing to invest. Conversely, $x(m, z) < 1$ implies $x(m, n) = 0$.

In order to obtain a contradiction, suppose $x(m, z) < 1$ for some m, and let $M^* \subseteq M$ be the set of messages that minimize $x(m, z)$. If $m^* \in M$ then $x(m^*, z) < 1$ and $x(m^*, n) = 0$.

Claim 1: If $m^* \in M^*$ then $0 < x(m^*, z) < 1$ for .

Proof of claim: By hypothesis, $x(m^*, z) < 1$. Suppose $x(m^*, z) = 0$. In this case, B will be unarmed for sure following message m^*. Clearly, all of A's types will send a message in M^*, since a zero probability that B invests is the best possible outcome for A. Since B will be known to be unarmed, player A attacks if and only if $a \geq 0$, which happens with probability $1 - F(0)$. Suppose B changes to a strategy where he invests, and refuses inspections if the investment succeeds. Since A never attacks when $a - c < -d_z$, the probability of an attack can be at most $1 - F(c - d_z)$, and B's expected improvement will be at least

$$\sigma \left\{ (1 - F(c - d_z)) \left(-\alpha + \gamma \right) + F(c - d_z)\delta_t - (1 - F(0))(-\alpha) \right\} - k$$

This is strictly positive by Assumption 3. This contradiction proves the claim.

Claim 2: Player A must send a message in M^* if $a > 0$ or $a < c - d_z$.

Proof of claim: If $a - c > -d_n$ then A always attacks, so at the message stage he simply wants to minimize the probability that B is armed. He does this by sending a message in M^*. Similarly, if $a - c < -d_z$ then A never attacks, and again he wants to minimize the probability that B is armed. Finally, consider the case $0 < a \leq c - d_n$. If A sends $m^* \in M^*$ and then (regardless of what happens at the inspections stage) attacks for sure, then his expected payoff is $a - \sigma \tau x(m^*, z)c$. Suppose instead he sends $m' \notin M^*$. Following this message, B's crazy type will be armed with probability $\sigma x(m', z)$ and his normal type will be armed with probability $\sigma x(m', n)$. If B is unarmed, A prefers to attack (since $a > 0$). If B is armed, then A prefers to attack if and only if B is crazy (since $-d_z < a - c < -d_n$). Therefore, A's

maximum possible payoff from sending message m' is

$$
\begin{aligned}
& (1 - \tau\sigma x(m', z) - (1 - \tau)\sigma x(m', n))a + \tau\sigma x(m', z)(a - c) + (1 - \tau)\sigma x(m', n)(-d_n) \\
= \; & (1 - \sigma(1 - \tau) x(m', n)) a - \sigma\tau x(m', z)c - \sigma(1 - \tau) x(m', n)d_n \\
< \; & a - \sigma\tau x(m^*, z)c
\end{aligned}
$$

since $a > 0$ and $x(m^*, z) < x(m', z)$. This proves the claim.

Notice that claim 2 implies $M^* \neq \varnothing$.

Claim 3: If any $m^* \in M^*$ was sent and B is armed, then B will not allow inspections.

Proof of claim: To obtain a contradiction, suppose that following $m^* \in M^*$, there is a positive probability that inspections are allowed and reveal that B is armed. Since only crazy types invest $(x(m^*, z) > 0 = x(m^*, n))$, B will be known to be crazy once he reveals that he is armed. Thus, all types of A with $a \geq c - d_z$ will attack. Since type $a < c - d_z$ will never attack in any situation, if B is armed he is strictly better off not revealing it, since inspections are costly and will not reduce the probability of attack. This contradiction proves the claim.

Claim 4: Following any $m^* \in M^*$, there must be positive probability that B is unarmed and refuses inspections.

Proof of claim: Suppose that if B observes $m^* \in M^*$ and is unarmed, then he allows inspections with probability one. It follows that if B refuses inspections, he will be known to be armed. Therefore, following message m^* there is never ambiguity about whether or not B is armed. Since $m^* \in M^*$ both minimizes the probability that B arms and also fully reveals if B is armed or not, all types of A will either send m^*, or else some equivalent message in M^* (i.e., a message which also fully reveals if B is armed). Since $x(m^*, n) = 0$ for $m^* \in M^*$, the normal type will never invest in this equilibrium, will always allow inspections, and will be attacked whenever $a > 0$. So his expected payoff is

$$
(1 - F(0))(-\alpha) - E\varepsilon \tag{22}
$$

where $E\varepsilon$ is the expected value of ε. Now, in no circumstance would A ever attack if $a < c - d_z$. Therefore, if B invests and then allows inspections if and only if the investment fails, his expected payoff is at least

$$
\sigma\left\{(1 - F(c - d_z))(-\alpha + \gamma) + F(c - d_z)\delta_t\right\} + (1 - \sigma)\left\{(1 - F(0))(-\alpha) - E\varepsilon\right\} - k \tag{23}
$$

28

Assumption 3 implies that (23) is strictly greater than (22) for both types of B. Thus, the normal type can improve on his equilibrium payoff by investing. This contradiction proves the claim.

Now we can complete the proof of the Proposition. Let ζ denote the probability that A attacks, conditional on a message in M^* having been sent and inspections refused. For any $m^* \in M^*$, $x(m^*, z) > 0 = x(m^*, n)$, so following message m^* player B will be either unarmed or armed and crazy. In either case, A prefers to attack if $a \geq 0$. Thus, by claim 2, if $a \geq 0$ then A sends a message in M^* and then attacks for sure. Type $a < c - d_z$ also sends a message M^*, but never attacks. Therefore, $1 - F(0) \leq \zeta \leq 1 - F(c - d_z)$. Following any $m^* \in M^*$, when B is unarmed he at least weakly prefers to refuse inspections. Thus, conditional on some message in M^* having been sent and player B not investing, B's expected payoff is $-\zeta\alpha$. On the other hand, if after any $m^* \in M^*$ B invests and then refuses inspections, then his expected payoff conditional on a message in M^* is

$$
\begin{aligned}
& \sigma \left\{ \zeta \left(-\alpha + \gamma \right) + (1 - \zeta)\delta_t \right\} + (1 - \sigma)\zeta(-\alpha) - k \\
\geq \; & \sigma(1 - F(c - d_z)) \left(-\alpha + \gamma \right) + (1 - \sigma)\zeta(-\alpha) - k \\
> \; & -\sigma(1 - F(0))\alpha - (1 - \sigma)\zeta\alpha \geq -\zeta\alpha
\end{aligned}
$$

where the first inequality is due to $\zeta \leq 1 - F(c - d_z)$ and $\delta_t \geq 0$, the second to Assumption 3, and the third to $1 - F(0) \leq \zeta$. Therefore, both types of B will strictly prefer to invest in response conditional on receiving a message in M^*, a contradiction of $x(m^*, n) = 0$. ∎

We now turn to our characterization of the set of effective cheap-talk equilibria. Player A sends a message $m \in M$, where M is an arbitrary message space. Of course, by relabeling messages, we obtain many equilibria that are equivalent, in the sense of producing the same outcomes. We are interested in the final outcomes, not the particular labeling of messages.

Proposition 5 *All equilibria with effective cheap-talk are equivalent to a two message equilibrium where, for some a' and a'' with $a'' > a' > 0$, player A sends a "tough" message if $a \in (a', a'')$ and a "conciliatory" message otherwise. After the tough message, player B invest with probability 1 and allows inspections iff he is armed. After the conciliatory message, the normal type of B arms with positive probability and refuses inspections with positive probability if he is armed.*

Proof. Without loss of generality, we may assume player A's types do not randomize over messages. Let $A(m)$ be the set of types of player A who send

29

message m in equilibrium. (That is, $a \in A(m)$ if type a sends m.) Without loss of generality, we may assume $A(m) \neq \varnothing$ for all $m \in M$ (since a message which is never sent can be dropped). We know that the crazy type invests with probability 1. Let $x(m)$ be the probability that the normal type of player B invests when player A sends $m \in M$. By Bayes' rule, the probability that B is crazy conditional on being armed and message m having been sent is

$$\tau(m, \text{armed}) \equiv \frac{\tau}{\tau + (1 - \tau)x(m)}.$$

Notice that $\tau(m,\text{armed}) \geq \tau$. 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 $x(m^C) < x(m^T) \leq 1$. If communication is effective in influencing B's investment decision, some types of A must send a message in M^C, and some types must send a message in M^T.

When B is unarmed, whether he is normal or crazy is payoff irrelevant, so there is no reason to distinguish the unarmed normal type from the unarmed crazy type. Abusing terminology, let B's *ex post type* be denoted $t \in \{z, n, u\}$, where n denotes that B is *armed and normal*, z that B is *armed and crazy*, and u denotes that B is *unarmed*. Let $I(m, t, \varepsilon)$ be the probability that player B allows inspections following message m, when his ex post type is $t \in \{z, n, u\}$ and the cost of inspection is $\varepsilon \in [0, \bar{\varepsilon}]$. This formulation is without loss of generality, because all unarmed players must make the same decision at the inspection stage (for the same ε).

Conditional only on m and t, the probability of inspections is

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

The probability that B is crazy conditional on being armed and allowing inspections following message $m \in M$ is

$$\tau(m, \text{allow,armed}) \equiv \frac{\tau I(m, z)}{\tau I(m, z) + (1 - \tau) x(m) I(m, n)}.$$

The probability that B is crazy conditional on being armed and refusing inspections after message $m \in M$ is

$$\tau(m, \text{refuse,armed}) \equiv \frac{\tau (1 - I(m, z))}{\tau (1 - I(m, z)) + (1 - \tau) x(m) (1 - I(m, n))}.$$

30

Also, let $att(m,$allow,armed$)$ be the probability that A attacks after message m if B allows inspections which reveal he is armed. Similarly, $att(m,$allow,unarmed$)$ is the probability that A attacks if B allows inspections which reveal he is unarmed, and $att(m,$refuse$)$ is the probability that A attacks if B refuses inspections. The proof has 14 steps.

Step 1: Suppose after message m, there is positive probability that inspections reveal B is armed. Then,

(a) $att(m,$allow,armed$) < att(m,$refuse$)$;

(b) if type $a \in A(m)$ sends message m and then attacks when inspections reveal B is armed, then he must also attack if inspections are refused; and

(c) $\tau(m,$allow,armed$) \geq \tau(m,$armed$) \geq \tau(m,$refuse,armed$)$ (the inequalities are strict if B refuses inspections with positive probability when armed).

Proof: (a) Both armed types of B want to minimize the probability of attack as $\delta_t \geq 0 > -\alpha + \gamma$. Also, inspections are costly for B. Hence, when armed, B will allow inspections after message $m \in M$ only if they lower the probability of attack.

(b) If $a \in A(m)$ and type a attacks if inspections are refused, then if $a' > a$ and $a' \in A(m)$, type a' also attacks. Hence, for any message $m \in M$, there is a cut-off type $\tilde{a}$ such that if inspections are refused, then type $a \in A(m)$ attacks if and only if $a \geq \tilde{a}$. Similarly, there is a type $\hat{a}$ such that if inspections reveal that B is armed, then type $a \in A(m)$ attacks if and only if $a \geq \hat{a}$. Part (a) implies $\hat{a} > \tilde{a}$, which implies (b).

(c) The armed type $t \in \{n, z\}$ with cost shock ε is willing to allow an inspection if and only if

$$att(m, \text{allow,armed}) \left(-\alpha + \gamma\right) + \left(1 - att(m, \text{allow,armed})\right) \delta_t - \varepsilon$$
$$\geq \quad att(m, \text{refuse}) \left(-\alpha + \gamma\right) + \left(1 - att(m, \text{refuse})\right) \delta_t.$$

If this condition holds for $t = n$, then it holds strictly for $t = z$, because $\delta_z > \delta_n$ and $1 - att(m,$allow,armed$) > 1 - att(m,$refuse$)$ by part (a). As crazy types are more willing to allow inspections than normal types, (c) is proved. This completes the proof of step 1.

The remaining steps will establish that, without loss of generality, we can assume $M^C = \{m^C\}$ and $M^T = \{m^T\}$. Moreover, we will show that

$$I(m^T, z) \quad = \quad I(m^T, n) = 1,$$
$$I(m^T, u) \quad = \quad I(m^C, u) = 0$$

31

and either
$$I(m^C, z) = I(m^C, n) = 0$$

or
$$1 = I(m^C, z) > I(m^C, n) > 0.$$

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.

Step 2: Doves and hawks only send messages in M^C.

Proof: Suppose $a \in A(m)$, where a is a dove or a hawk. We claim $m \in M^C$. Notice that if, along the equilibrium path following message m, type a either *never* attacks or *always* attacks, then clearly $m \in M^C$, for otherwise type a could increase his expected payoffs by reducing the probability that B invests. There are three cases to consider.

Case 1: Suppose inspections never occur on the equilibrium path after message m is sent.

For a dove with $a \leq 0$, the expected payoff from not attacking is

$$-\sigma\left(\tau d_z + (1 - \tau) x(m) d_n\right) \geq -\sigma\left(\tau d_z + (1 - \tau) d_n\right) > a - \sigma c$$

by Assumption 1, so his best-response is to never attack. Hence, $m \in M^C$.

For a hawk with $a \geq c - \left(\tau d_z + (1 - \tau) d_n\right)$, the expected payoff from attacking is

$$
\begin{aligned}
& a - \sigma\left(\tau + (1 - \tau) x(m)\right) c \\
= {} & (1 - \sigma\left(\tau + (1 - \tau) x(m)\right)) a + \sigma\left(\tau + (1 - \tau) x(m)\right)(a - c) \geq \\
& (1 - \sigma\left(\tau + (1 - \tau) x(m)\right)) a - \sigma\left(\tau + (1 - \tau) x(m)\right)(\tau(m, \text{armed}) d_z + (1 - \tau(m, \text{armed})) d_n) \\
= {} & (1 - \sigma\left(\tau + (1 - \tau) x(m)\right)) a - \sigma\left(\tau d_z + (1 - \tau) x(m) d_n\right) \\
> {} & -\sigma\left(\tau d_z + (1 - \tau) x(m) d_n\right)
\end{aligned}
$$

as $a > 0$ and $\tau(m,\text{armed}) \geq \tau$. Hence, a hawk's best-response is to always attack. Hence, $m \in M^C$.

Case 2: Suppose inspections always occur on the equilibrium path after message m is sent.

Suppose type $a \in A(m)$ is a dove. If inspections reveal that B is unarmed, the dove does not attack. If in addition he doesn't attack when inspections reveal that B is armed, then he never attacks, so $m \in M^C$. Suppose instead

32

that type a attacks if inspections reveal B is armed. Then all types $a' \in A(m)$ with $a' \geq a$ also attack if inspections reveal B is armed. By part (a) of step 1, an armed type must expect that an attack occurs with higher probability if he refuses inspections. Hence, there must be some types $a' \in A(m)$ with $a' < a$ who do not attack when inspections reveal B is armed. Hence, there is $a' \in A(m)$ who never attacks, so $m \in M^C$.

Suppose type $a \in A(m)$ is a hawk. Following message m, type a attacks inspections reveal B is unarmed. If inspections reveal B is armed, the payoff from attacking is

$$a - c \geq - \left(\tau d_z + (1 - \tau) d_n \right) \geq - \left(\tau(m, \text{armed}) d_z + (1 - \tau(m, \text{armed})) d_n \right)$$

and hence a always attacks. Therefore, $m \in M^C$.

Case 3: Suppose inspections sometimes (but not always) occur on the equilibrium path after message m is sent.

Sub-case 3.1: If B always allows inspections when armed, then B's equilibrium strategy fully reveals his capabilities with probability one. Hence, the argument in Case 2 shows $m \in M^C$.

Sub-case 3.2: Suppose armed types sometimes accept and sometimes reject inspections on the equilibrium path in response to message $m \in M$.

Suppose type $a \in A(m)$ is a dove. Then type a does not attack if inspections reveal B is unarmed. Suppose type a attacks if inspections are refused. This implies that

$$-\lambda \left(\tau(m, \text{refuse, armed}) d_z + (1 - \tau(m, \text{refuse,armed})) d_n \right) \leq a - \lambda c < -\lambda c \tag{24}$$

where λ is the probability that B is armed conditional on refusing inspections. (The first inequality of (24) says that type a prefers to attack, the second is due to $a < 0$.) Notice that (24) implies that

$$-c > - \left(\tau(m, \text{refuse,armed}) d_z + (1 - \tau(m, \text{refuse,armed})) d_n \right). \tag{25}$$

We than have

$$
\begin{aligned}
a - c \quad &\geq \quad - \left(\tau(m, \text{refuse, armed}) d_z + (1 - \tau(m, \text{refuse, armed})) d_n \right) \tag{26} \\
&> \quad - \left(\tau(m, \text{allow, armed}) d_z + (1 - \tau(m, \text{allow, armed})) d_n \right)
\end{aligned}
$$

The first inequality in (26) is due to the fact that, since the first inequality in (24) holds for $\lambda \leq 1$, (25) implies it also holds when λ is replaced by 1. The

second inequality in (26) is due to $\tau(m,\text{allow,armed}) > \tau(m,\text{refuse,armed})$, from part (c) of step 1. Now, (26) implies that type a strictly prefers to attack when B allows inspections which reveal that he is armed. Hence, any dove in $A(m)$ who attacks when inspections are refused also attacks when inspections show B is armed. More generally, if type $a \in A(m)$ attacks when inspections show B is armed, then any type $a' > a$ in $A(m)$ also attacks when inspections show B is armed. But this implies $att(m,\text{allow,armed}) \geq att(m,\text{refuse})$, contradicting part (a) of step 1. This contradiction proves that type a does not attack if inspections are refused. If type a attacks when inspections reveal that B is armed then, by step 1 (b), he must also attack if inspections are refused, which contradicts the previous sentence. So in fact a dove $a \in A(m)$ never attacks, hence $m \in M^C$.

Suppose type $a \in A(m)$ is a hawk. Then,

$$a - c \geq -\left(\tau d_z + (1-\tau)\, d_n\right) > -\left(\tau(m,\text{allow,armed})d_z + (1 - \tau(m,\text{allow,armed}))\, d_n\right) \tag{27}$$

(The first inequality is due to the definition of hawk, the second is due to $\tau(m,\text{allow,armed}) > \tau(m,\text{armed}) \geq \tau$ from step 1 (c).) Now (27) implies that type a attacks if inspections reveal B is armed. Part (b) of step 1 implies type a also attacks if inspections are refused. Since hawks certainly attack if inspections reveal B is unarmed, type a always attacks. Hence $m \in M^C$.

Sub-case 3.3: Finally, suppose B never allows inspections when armed, after m was sent. Then he must sometimes allow inspections when unarmed, otherwise we are in case 1. Hence, the probability of attack must be lower after inspections reveal B is unarmed than after he refuses inspections:

$$att(m,\text{allow,unarmed}) < att(m,\text{refuse}). \tag{28}$$

By an argument like in the proof of step 1 part (b), (28) implies that if there is any type in $A(m)$ who attacks when inspections reveal B is unarmed, he must also attack when inspections are refused, hence he always attacks. In this case, $m \in M^C$. Suppose instead that no type in $A(m)$ attacks when inspections reveal B is unarmed, i.e., only doves send message m. But then in response to message m, a normal type of B should refrain from investing, and then reveal that he is unarmed. Therefore, $x(m) = 0$, which certainly implies $m \in M^C$.

Step 3: $I(m^T, u) = 0$ for all $m^T \in M^T$.

Proof: By step 2, message $m^T \in M^T$ reveals that A must be an opportunist. Therefore, if B is unarmed, then he will refuse inspections following

34

message $m^T \in M^T$. (If he were to allow inspections, he would pay $\varepsilon > 0$ only to be attacked for sure.)

Step 4: $I(m^T, n) = I(m^T, z) = x(m^T) = 1$ for all $m^T \in M^T$. If message $m^T \in M^T$ is sent and there is no inspection, then an attack occurs with probability one.

Proof: Inspections must occur with positive probability following m^T, or else all types of A would prefer to send $m^C \in M^C$. Since $I(m^T, u) = 0$ by step 3, whenever inspections occur they must reveal that B is armed, so step 1 applies. Specifically, if some type $a \in A(m^T)$ attacks after the inspections, then he must also attack if inspections are refused, by step 1 (b). That means he always attacks (since $I(m^T, u) = 0$), so he is better of sending $m^C \in M^C$, a contradiction. Hence, no type in $A(m^T)$ attacks after inspections have revealed B is armed. If some type $a \in A(m^T)$ doesn't attack when inspections are refused, then this type never attacks (since $I(m^T, u) = 0$), so he is better of sending $m^C \in M^C$, a contradiction. Hence, when inspections are refused, the probability of attack is one. On the other hand, as we have seen, if inspections reveal B is armed then the probability of attack is zero. This clearly means B strictly prefers to allow inspections when armed, hence $I(m^T, n) = I(m^T, z) = 1$ for any $m^T \in M^T$. It also implies B strictly prefers to invest when he hears message $m^T \in M^T$, so $x(m^T) = 1$. Indeed, if after hearing message $m^T \in M^T$ player B invests and allows inspections if and only if he is armed, his payoff is $\sigma(\delta_t - E\varepsilon) - (1 - \sigma)\alpha - k$. By not investing he gets $-\alpha$. Assumption 3 implies that he prefers to invest.

Step 5: We can assume, without loss of generality, that $M^T = \{m^T\}$ is a singleton.

Proof: Steps 2 and 4 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 no arms or inspections are refused. Also, $x(m^T) = 1$ for all $m^T \in M^T$. Therefore, all messages in M^T lead to the same outcome, so we may as well assume there is only one such message.

Step 6: Following any $m^C \in M^C$, $I(m^C, n) < 1$.

Proof: Since the only reason to send the message $m^T \in M^T$ is to improve the chance of an inspection, steps 3 and 4 imply we cannot have $I(m^C, n) = I(m^C, z) = 1$. By an argument as in step 1 (c), crazy types are more willing to allow inspections than normal types, so we must have $I(m^C, z) > I(m^C, n)$.

Step 7: Doves never attack on the equilibrium path and hawks always attack.

Proof: Consider the doves. By step 2, we know they send only messages in M^C. In the proof of step 2, case 1, we established that if inspections never occur on the equilibrium path, doves do not attack. So assume inspections occur with positive probability on the equilibrium path following m^C. First, suppose $I(m^C, u) > 0 = I(m^C, z) = I(m^C, n)$. Now, if inspections reveal B is unarmed, then all types with $a > 0$ attack. For the unarmed to be willing to allow inspections, the inspections must strictly reduce the probability of attack. This implies that some doves ($a < 0$) who send m^C must attack if inspections are refused. These doves get no more than $-\sigma(\tau + (1-\tau)x(m^C))c$ from sending m^C, since they end up attacking whenever B is armed. If instead they send m^T and never attack, they get $-\sigma(\tau d_z + (1-\tau)d_n)$. Thus, for them to prefer to send m^C, we need

$$-\sigma(\tau + (1 - \tau)x(m^C))c \geq -\sigma(\tau d_z + (1 - \tau)d_n) \qquad (29)$$

If type $a > 0$ sends m^T, as shown above, in equilibrium he will attack whenever B is unarmed, and he gets

$$(1 - \sigma)a - \sigma(\tau d_z + (1 - \tau)d_n) < a - \sigma(\tau + (1 - \tau)x(m^C))c$$

The inequality uses $1 - \sigma < 1$ and (29). But, the right hand side is what he gets if he sends m^C and always attacks. Thus, no type $a > 0$ will send m^T, so m^T is never sent in equilibrium, a contradiction. Therefore, $I(m^C, u) > 0 = I(m^C, z) = I(m^C, n)$ is impossible. Thus, if inspections sometimes occur, they must sometimes occur when B is armed, so we can use step 1. By step 6, they must also be refused with positive probability when B is armed. Suppose type $a = 0$ sends m^C and inspections are refused. Let $\tilde{\sigma}$ denote the probability B is armed, conditional on inspections being refused after message $m^C \in M^C$. Then, after the refusal, type $a = 0$ prefers to attack if and only if

$$-\tilde{\sigma}c \geq -\tilde{\sigma}\left(\tau(m, \text{refuse})d_z + (1 - \tau(m, \text{refuse}))d_n\right) \qquad (30)$$

We will show this inequality leads to a contradiction. Step 1 (c) implies $\tau(m,\text{refuse}) \leq \tau(m,\text{allow,armed})$. Therefore, (30) implies

$$-c \geq -\left(\tau(m, \text{allow,armed})d_z + (1 - \tau(m, \text{allow,armed}))d_n\right)$$

36

so (30) implies type $a = 0$ also prefers to attack if inspections reveal B is armed. Hence, type $a = 0$ prefers to always attack, which is to say

$$-\sigma \left(\tau + (1-\tau)x(m^C)\right) c \geq -\sigma \left(\tau d_z + (1-\tau)d_n\right).$$

But then,as $1 - \sigma < 1$, for opportunists with $a > 0$ we have

$$a - \sigma \left(\tau + (1-\tau)x(m^C)\right) c > (1-\sigma) a - \sigma \left(\tau d_z + (1-\tau)d_n\right). \qquad (31)$$

The right-hand-side of (31) is the expected payoff to an opportunist from sending message $m^T \in M^T$. Hence, (31) implies no opportunist sends a message in M^T and in fact all types send messages in M^C. But this contradicts our assumption that the equilibrium displays effective communication. This contradiction shows (30) cannot hold, so type $a = 0$ strictly prefers *not* to attack if inspections are refused, and hence so do all types with $a < 0$. If there is positive probability that inspections reveal B is armed, then if type $a < 0$ attacks in this case, he must also attack if inspections are refused, by step 1 (b), which contradicts the previous sentence. Thus, if type $a < 0$ sends m^T, he will not attack if inspections are refused, or if they reveal that B is armed. Since he will certainly not attack if inspections show B is unarmed, type $a < 0$ will never attack. A similar argument shows that hawks always attack on the equilibrium path.

Step 8: $I(m^C, u) = 0$ 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 (by step 7). Clearly, if B hears message $m^C \in M^C$ and is unarmed, 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 unarmed. Therefore, $I(m^C, u) = 0$.

Step 9: $0 < x(m^C) < 1$ for $m^C \in M^C$, so B's normal type is indifferent between investing and not investing when he hears m^C.

Proof: Suppose $x(m^C) = 0$ for $m^C \in M^C$. Then, after message m^C, only crazy types are armed. Clearly, any type $a > 0$ prefers to send m^C rather than m^T, in view of $m^T = 1$. But then, by step 2, no type sends m^T, which contradicts communication being effective. Thus, $x(m^C) = 0$ is impossible. On the other hand, $x(m^C) < x(m^T) = 1$.

Step 10: There are cut-off points a' and a'', where $0 < a' \leq a'' < c - \tau d_z - (1-\tau) d_n$, such that A sends m^T if $a \in (a', a'')$. He sends some $m^C \in M^C$ if $a < a'$ or $a > a''$.

Proof: The proof of Step 7 established that type $a = 0$ strictly prefers not to attack if inspections are refused on the equilibrium path. 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 $c - \tau d_z - (1 - \tau) d_n$ 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, $0 < a' \leq a'' < c - \tau d_z - (1 - \tau) d_n$.

Step 11: For all $m^C \in M^C$, $I(m^C, n) = I^* < 1$.

Proof: Step 6 established $I(m^C, n) < 1$. In this case, if both the crazy type and the normal type never allow inspections after all messages in M^c, we are done. Now, if type z is willing to allow inspections, they must reduce the probability of attack. Hawks always attack by step 7. Therefore, if inspections reveal that B is armed, some "tough" opportunists ($a'' < a < c - \tau d_z - (1 - \tau) d_n$) must not attack. If $I(m^C, n) = 0$, then when inspections reveal B is armed A can infer that B is crazy, so all opportunists would attack, contradicting the previous sentence. Therefore, $0 < I(m^C, n) < I(m^C, z) = 1$. By definition, $x(m^C)$ is the same for all $m^C \in M^C$. This implies we must have $I(m^C, n)$ constant for all m^C that are sent in equilibrium. Otherwise, all "tough" opportunists ($a'' < a < c - \tau d_z - (1 - \tau) d_n$) must send such messages as they attack on the equilibrium path iff inspections are refused. But then, the incentive of the normal type of B to arm is higher after such messages and $x(m^c)$ cannot be constant for all messages $m^c \in M^c$, a contradiction.

Step 12: We can assume, without loss of generality, that $M^C = \{m^C\}$ is a singleton.

Proof: Step all messages in M^C yields the same outcome, so we may assume there is only one such message.

Step 13 proves that any equilibrium with $I(m^C, n) = I(m^C, z) = 0$ for all $m^C \in M^C$ is outcome-equivalent to the communication equilibrium described in Section 4. ∎