Strategic Ambiguity and Arms Proliferation

Sandeep Baliga

Northwestern University

Tomas Sjöström

Rutgers University

A big power is facing a small power that may have developed weapons of mass destruction. The small power can create strategic ambiguity by refusing arms inspections. We study the impact of strategic ambiguity on arms proliferation and welfare. Strategic ambiguity is a substitute for actually acquiring weapons: ambiguity reduces the incentive for the small power to invest in weapons, which reduces the threat of arms proliferation. But strategic ambiguity hides information, and this can lead to costly mistakes. Cheap-talk messages can be used to trigger inspections when such mistakes are particularly costly. Tough messages that trigger inspections always imply a greater risk of arms proliferation.

I. Introduction

Countries sometimes try to create ambiguity about their military capabilities. For example, Saddam Hussein possessed weapons of mass destruction (WMDs) in the early 1990s but not in the late 1990s. He appears to have deliberately chosen a policy of ambiguity in both situations. The motivation of policy makers can be hard to decipher, es-

We thank Canice Prendergast, three referees, and seminar audiences at Berkeley, Columbia, Göteborg, Indiana, Kellogg Cooperation and Conflict conference, Rice, the Stockholm School of Economics, University of British Columbia, and the Econometric Society 2007 winter meetings. We also thank Kane Sweeney for excellent research assistance. Sandeep Baliga gratefully acknowledges research support from the Zell Center for Risk Research. All remaining errors are our own responsibility.

[*Journal of Political Economy*, 2008, vol. 116, no. 6]

pecially in countries that lack democratic institutions and a free press. However, a policy of ambiguity may be an attempt to deter aggression without risking negative sanctions or preemptive strikes.

A country that lacks WMDs may use strategic ambiguity to create "deterrence by doubt" (see Gordon and Trainor 2006, 65). For example, when Saddam Hussein revealed to his inner circle that Iraq had no WMDs, he "flatly rejected a suggestion that the regime remove all doubts to the contrary" because he thought such a revelation would embolden his enemies to attack (Woods, Lacey, and Williamson 2006, 6). However, a country that possesses WMDs may rely on strategic ambiguity to avoid sanctions or preemptive strikes. For example, Israel's policy of strategic ambiguity on nuclear weapons may be "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" (Myre 2006, 5).[1]

The conventional wisdom is that ambiguity is detrimental to world peace and, conversely, that arms inspections promote peace and trust (e.g., Schrage 2003). This view is embodied in article 3 of the Treaty on the Non-proliferation of Nuclear Weapons, known as the NPT (http://www.iaea.org/Publications/Documents/Treaties/npt.html). The NPT requires that nations submit to inspection and verification of nuclear facilities by the International Atomic Energy Agency. 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 are unknown. For example, it is unclear if Pakistan and India have intercontinental ballistic missiles (Norris et al. 2002; Norris and Kristensen 2005). Would the world be safer if arms inspections could be forced on these countries?

A game-theoretic analysis of strategic ambiguity is found in Sobel (1992). In his model, military capability is exogenous. Weak countries benefit from ambiguity because "deterrence by doubt" protects them from being attacked. Strong countries dislike ambiguity because it prevents them from taking advantage of the weak. In contrast, our emphasis is on arms proliferation. We assume that military capability is endogenous: weak countries can acquire advanced weapons at a cost.

In our model, there are parameter values such that if there is no ambiguity, weak countries prefer not to acquire advanced weapons. For such parameter values, strong countries surely prefer to eliminate all ambiguity (say, by enforcing the NPT). However, there are other parameter values such that if there is no ambiguity, then weak countries will try to acquire advanced weapons, but strategic ambiguity (deter-

[1] Similarly, North Korea's vice minister of foreign affairs once told visiting American scientists that their policy of ambiguity protected them against sanctions: "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 2004, 21).

rence by doubt) can be a substitute for a costly weapons program. In this parameter region, which we focus on, allowing the weak to practice deterrence by doubt can reduce the threat of arms proliferation, and this can make all countries (weak and strong) better off. This result stands in opposition to the conventional arguments in favor of arms inspections.

Advanced weapons may bring various benefits. Some potential benefits, such as international or domestic prestige, require the weapons to be publicly known. Strategic ambiguity, however, requires that countries sometimes refrain from revealing weapons they actually possess (otherwise there can be no deterrence by doubt). In our model, secret weapons are valuable because they can be used in self-defense if the country is attacked. If they are revealed, they may deter an aggressor. But we abstract from other benefits such as international prestige that require the weapons to be publicly known. Such benefits could be added, but they cannot be too important, or else strategic ambiguity would not be incentive compatible.

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 that initially is unarmed; that is, it lacks advanced weapons. But B can try to acquire advanced weapons by making an investment. If this succeeds, then B will become armed. With a small probability, B is a "crazy" type who might share his weapons with terrorists. Player B's type is *soft* (unverifiable) information. In Baliga and Sjöström (2004), we studied how soft private information can trigger arms races and wars. Now we will consider whether revelation of *hard* information can promote peace and trust. We assume that B's military capability is hard information that can be verified by weapons inspectors. The inspectors are assumed to be perfectly reliable and never make any mistakes, and the cost of inspections is very small. This allows us to study the strategic incentives in an idealized world with no "frictions" generated by imperfect or costly inspections.

Player A must decide whether or not to attack player B. The optimal decision depends on A's preferences (his type) and his beliefs. Player A can be a peaceful *dove*, an aggressive *hawk*, or an *opportunistic type*. As in classical deterrence theory, opportunistic types are deterred from attacking if they think that B is a normal type who is armed. This does not mean that if B's weapons program was a success, it must be optimal for him to reveal it. Player A might interpret this as a signal that B is crazy and attack because the cost of allowing the crazy type to possess advanced weapons is very high. Thus, the fear of sending a negative signal about soft information can prevent B from revealing hard information. There is no "unraveling" as in models with hard but no soft information (e.g., Grossman 1981). In our model there is always a *full ambiguity* equilibrium in which B's military capabilities are never revealed, as well as a *full disclosure* equilibrium in which they are always revealed.

The opportunistic type of A will attack if he discovers that B is unarmed. Accordingly, with full disclosure, B has a strong incentive to make the investment and try to acquire new weapons. With full ambiguity, B's incentive to invest is smaller since he is protected by deterrence by doubt. If the parameters are such that B is unlikely to invest with full ambiguity, then all of A's types benefit from the ambiguity. However, if the probability that B invests is almost the same with full ambiguity as with full disclosure, then A's opportunistic type prefers the latter. Even if B invests, there is a chance that the investment fails, and the opportunistic type wants this to be revealed so he can attack the unarmed. But hawks and doves have a strong intrinsic preference for a particular action ("do not attack" for doves, "attack" for hawks), so they do not need information about B's capabilities in order to decide what to do. They always prefer ambiguity because it minimizes the probability that B invests.

The fact that different types of player A can disagree about whether ambiguity is desirable suggests a role for communication. In a *mixed inspections equilibrium*, player A sends either a "tough" or a "conciliatory" cheap-talk message. The conciliatory message encourages B to preserve ambiguity and to refrain from investing, which reduces the risk of arms proliferation. The tough message can be interpreted as a demand for weapons inspections. Player B will be more likely to invest knowing that his capabilities will be revealed. Accordingly, if A's types are ordered according to their propensity to attack, then A will use a "nonconvex" strategy. *Intermediate* (opportunistic) types have a *demand for information*—they want to find out if B is armed—so they send the tough message. Extreme types on both sides (hawks and doves) send the conciliatory message in order to minimize the risk of arms proliferation. For the equilibrium to exist, B must be willing to refrain from investing when he receives the conciliatory message. This requires that the prior probability that A is a hawk is not too big. Thus, the mixed inspections equilibrium exists only for some parameter values.

We show in the Appendix that if inspections do not consume significant real resources, then any equilibrium must be equivalent (in terms of payoffs, investment, and attack probabilities) to either the full disclosure, full ambiguity, or mixed inspections equilibrium. Thus, we focus on these three without loss of generality. In particular, there is no reason to include more than two messages in A's message space. Incentive compatibility for A requires that one message (the tough one) makes inspections more likely but increases the risk of arms proliferation, whereas the other message (the conciliatory one) has the opposite effect. Player A's choice among these two messages captures the trade-off between ambiguity and arms proliferation.

In the mixed inspections equilibrium, A can trigger inspections by sending a tough message. By revealed preference, all of A's types prefer the mixed inspections equilibrium to the equilibrium with full disclosure. But B may prefer full disclosure if ambiguity is not an effective deterrent. In both the mixed inspections and the full ambiguity equilibria, 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 such that strategic ambiguity about the small power's arsenal is good for the big power (because it reduces the risk of arms proliferation) but bad for the small power (because of the mistakes).

Crawford and Sobel (1982) and Green and Stokey (2007) initiated the study of cheap-talk games. In Matthews (1989), the receiver picks a proposal in a one-dimensional space that the sender may veto. In our model, the receiver (B) takes two actions (he invests and decides whether to allow inspections). Nevertheless, like Matthews, we find that at most two messages are sent in equilibrium, a tough and an accommodating one. In Baliga and Sjöström (2004), we considered the role of cheap talk in a model in which each player could decide to be aggressive and each felt threatened by the opponent. The current paper investigates the role of hard information, and the strategic situation is quite different. If B reveals that he is unarmed, he convinces A that he is not a threat. However, this might trigger an attack from an opportunistic type, so B faces a classic deterrence problem. In our earlier paper, opportunistic types did not exist, and there was no hard information, no deterrence problem, and no role for strategic ambiguity.

Townsend (1979) initiated a literature on *costly state verification*, where the value of information is traded off against the resource cost of inspections (for a recent contribution, see Bond [2004]). We do not study this trade-off. Indeed, we assume that inspections do not consume significant real resources. In our model, inspections have a different, more subtle, cost: player A cannot commit not to attack if inspections reveal that B is unarmed, and this may cause B to arm himself. By forgoing inspections, A in effect commits not to attack B when B is unarmed, which lowers the risk of arms proliferation.

Finally, there exists an empirical literature that investigates the effect of concealed self-protective devices on crime (Lott and Mustard 1997; Ayres and Levitt 1998). Concealed self-protective devices generate a 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 we show that ambiguity may be welfare improving because it reduces the incentive for B to arm himself.

In Section II, we describe the model. In Section III, we study equilibria without communication. Section IV considers the role of cheap talk. Section V presents conclusions. A characterization of the equilibrium set is in the Appendix.

II. The Arms Proliferation Game

A. *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: either he invests or he does not invest. The cost of investing is $k > 0$. Player B's investment is successful with probability $\sigma \in (0, 1)$. If it is successful, then B acquires advanced weapons and becomes *armed*. Otherwise B remains *unarmed*. Thus, if B invests, then he will become armed with probability σ; but if he does not invest, then he remains unarmed for sure. Player A cannot directly observe if B invests or is armed. Hard information about B's weapons can be obtained by (perfectly reliable) inspectors. If there is an inspection, then B incurs a fixed cost $\varepsilon > 0$. The inspection publicly reveals whether B is armed or unarmed. The cost of inspections could be monetary or psychological (e.g., "loss of face").

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 that 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 cumulative distribution function is denoted F. In addition, if B is armed when A attacks, the weapons will yield a benefit $\gamma \in (0, \alpha)$ to B and a cost $c > 0$ to A. It is useful to define the *normalized cost of investing* to be

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

Player B has two possible types, "crazy" and "normal."[2] 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. The prior 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

[2] The crazy type should not be thought of as "irrational"; he simply maximizes a different payoff function than the normal type.

benefit δ_t. Intuitively, the crazy type may be more likely to 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 (e.g., 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: $\delta_z < \gamma$. To simplify the exposition, we assume $\delta_n = 0$.[3] We also assume that if A attacks, then he eliminates the threat posed by B.[4]

To summarize, we assume

$$0 < d_n < d_z$$

and

$$0 = \delta_n < \delta_z < \gamma < \alpha.$$

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. Bayes' rule does not apply to out-of-equilibrium "surprises." All we require in this case is that all types of the player who received the surprise update in the same way.

B. 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.

[3] Our main results still hold if $0 < \delta_n < \delta_z$.

[4] More generally, the threat from B could be reduced but not completely eliminated when A attacks. This would complicate the exposition without adding any new insights.

- Time 3: If B invested, then he privately learns whether or not he has acquired weapons.
- Time 4: Player B decides whether or not to allow inspections. If inspections take place, then player B incurs the cost ε and the inspectors publicly reveal whether or not B is armed.
- Time 5: Player A decides whether or not to attack.

C. Parameter Restrictions

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

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

The first inequality in assumption 1 guarantees that opportunistic types can exist. It can be rewritten as

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

The left-hand side of (1) is A's net benefit from attacking an unarmed B. The right-hand side of (1) is A's expected net benefit of attacking an armed B if B is crazy with probability τ (i.e., the payoff from attacking minus the payoff from not attacking). The inequality (1) says that given A's prior τ, A is less inclined to attack when B is armed than when B is unarmed. Because of this, there can be a deterrence motive for acquiring weapons.[5]

The second inequality of assumption 1 can be rewritten as

$$a - c - (-d_z) > a. \tag{2}$$

The left-hand side of (2) is A's net benefit from attacking an armed B if B is thought to be crazy for sure. Thus, (2) says that if A is convinced that B is crazy, then A is more inclined to attack when B is armed than when B is unarmed. If (2) did not hold, 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 (as in Grossman [1981] and Sobel [1992]). To support an equilibrium with

[5] In reality, deterrence is often claimed to be the justification for weapons programs. An Iranian newspaper editorial cited by Takeyh (2005) states that "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" (23).

ambiguity, A's disutility from advanced weapons in the hands of crazy types must be sufficiently big.[6]

Next, we assume that the cost of inspections is small but positive.

ASSUMPTION 2. $0 < \varepsilon < (1 - H - D)(\alpha - \gamma)$.

The first inequality in assumption 2 does not play a significant role in the argument in the text. However, with $\varepsilon > 0$, player B allows inspections only if they strictly reduce the probability of attack; this intuitive property will be used in the Appendix.

Notice that $1 - H - D$ is the probability that A is opportunistic. If B is attacked when he is armed, then he suffers a *disutility* of $\alpha - \gamma > 0$ (i.e., his payoff is $-\alpha + \gamma$). Thus, the second inequality in assumption 2 guarantees that B is willing to incur the cost of inspection if this deters the opportunistic types from attacking. If this inequality did not hold, then inspections might be refused simply because the resource cost of inspections dominates the value of deterrence. In contrast to the literature on costly state verification, we are not interested in the tradeoff between the real resource cost of inspections and the value of information. Instead, we want to focus on the "pure" incentives to demand and supply information, and to this end, we assume that ε is small. In our theory, the benefit of strategic ambiguity is not simply to economize on the cost of revealing hard evidence (but in a real scenario, this might of course be an additional benefit).

To motivate our final assumption, suppose for the moment that B cannot hide his military capability from A. Then if B is unarmed, he expects the payoff $\pi = [1 - F(0)](-\alpha)$ because (knowing B is unarmed) A attacks whenever $a > 0$. Suppose A thinks that only crazy types arm. Then if B is armed, he will be attacked whenever $a - c > -d_z$, yielding the expected payoff

$$\pi' = [1 - F(c - d_z)](-\alpha + \gamma) + F(c - d_z)\delta_n \tag{3}$$

for the normal type of B. The investment costs k and succeeds with probability σ, so in this scenario the normal type of B prefers not to invest if

$$\sigma\pi' + (1 - \sigma)\pi - k \leq \pi. \tag{4}$$

With our simplifying assumption $\delta_n = 0$, the inequality (4) can be re-

[6] Nuclear weapons in a "rogue nation" may not have the range to reach the United States, but in the hands of terrorists they could destabilize world security. It seems plausible that a U.S. leader could see the net benefit of eliminating this threat as positive or negative, depending on his or her assessment of the "type" of the "rogue leader." This is exactly the scenario generated by the two inequalities of assumption 1.

written as

$$\kappa \geq \frac{\alpha}{\gamma}[1 - F(0)] + \left(1 - \frac{\alpha}{\gamma}\right)[1 - F(c - d_z)], \tag{5}$$

where $\kappa \equiv k/(\sigma\gamma)$. If (5) holds, then the normal type of B does not try to arm himself, knowing he cannot hide his weapons from A. Clearly, this inequality holds if B's normalized cost of investment κ is high. In this case, the big power A has nothing to gain by allowing ambiguity, which agrees with the conventional wisdom discussed in Section I. Our main objective is to show that there are parameter regions in which there is a trade-off between ambiguity and arms proliferation and in which the conventional wisdom is incorrect. To focus on the main case of interest, we assume that (5) is not satisfied.

Assumption 3.

$$\kappa < \frac{\alpha}{\gamma}[1 - F(0)] + \left(1 - \frac{\alpha}{\gamma}\right)[1 - F(c - d_z)].$$

As we will see, assumption 3 guarantees that the normal type of B invests with positive probability in every equilibrium, generating a trade-off between the probability of investment and the probability of inspections. In contrast to conventional wisdom, ambiguity may then benefit A as well as B. As the normal type invests with positive probability and the crazy type attaches a higher value to acquiring weapons ($\delta_z > \delta_n$), we get the following result (the proof is in the Appendix).

Proposition 1. In every equilibrium, the crazy type of player B invests with probability one.

III. Equilibria without Communication

In this section we study two equilibria, "full disclosure" and "full ambiguity," that do not involve any communication at time 1. In Section IV, we will study a third equilibrium with cheap talk. In the Appendix, we show that these three equilibria are focused on without loss of generality if ε is sufficiently small.

A. *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).

Proposition 2. There is an equilibrium with full disclosure. Full disclosure implies that both types of player B invest with probability one.

Proof. Suppose in equilibrium that 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 in which 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 $a < c - d_z$). If the investment fails, B behaves exactly as he would have done had he not invested. The gain from this deviation is at least

$$\sigma\{[1 - F(c - d_z)](-\alpha + \gamma) + F(c - d_z)\delta_t - [1 - F(0)](-\alpha)\} - k.$$

This expression is strictly positive, by assumption 3. Therefore, in any full disclosure equilibrium, both types of player B invest with probability one.

It remains to show that an equilibrium with full disclosure exists. Let the equilibrium strategy specify that 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 he is a hawk, that is, if $a - c > \tau d_z + (1 - \tau)d_n$ (A thinks that the armed B is crazy with probability τ since both types invest). If B refuses inspections, then A infers that B is unarmed; so A attacks if he is not a dove, that is, 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 that B deviates by refusing inspections when he is armed. This raises the probability of attack from H (the probability that A is a hawk) to $1 - D$ (the probability that A is not a dove), which has a cost $(1 - D - H)(\alpha - \gamma + \delta_t)$. The gain from the deviation is only ε. Assumption 2 guarantees that the cost exceeds the benefit, so B prefers to reveal that he is armed. Finally, given full disclosure, assumption 3 guarantees that B prefers to invest. QED

Notice that with full disclosure the opportunistic type benefits from finding out if B is armed (he attacks if and only if B is unarmed). However, this information does not benefit hawks and doves because they will not use it: in the full disclosure equilibrium, the hawk attacks whether or not B is armed and the dove never attacks.

B. 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. 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

$$\kappa > 1 - F(\sigma[c - \tau d_z - (1 - \tau)d_n]), \tag{6}$$

and $\tilde{x} = 1$ otherwise.

Proof. Proposition 1 implies that player B's crazy type always invests. Let $\tilde{x}$ denote the probability that the normal type of B invests. Thus, B is armed with probability $\sigma[\tau + (1 - \tau)\tilde{x}]$.

If A attacks with probability one, then B will surely set $\tilde{x} = 1$; but then doves prefer not to attack. If A attacks with probability zero, then B will surely set $\tilde{x} = 0$; but then hawks prefer to attack. This shows that in equilibrium A must attack with probability strictly between zero and one. Therefore, the equilibrium must satisfy a cutoff property: if there is no inspection, then there is a type $\tilde{a} \in (a_0, a_1)$ such that A attacks if $a > \tilde{a}$ but not if $a < \tilde{a}$. Type $\tilde{a}$ must be indifferent between attacking and not attacking. He expects $\tilde{a} - \sigma[\tau + (1 - \tau)\tilde{x}]c$ by attacking and $-\sigma[\tau d_z + (1 - \tau)\tilde{x}d_n]$ by not attacking. Therefore, the indifference condition is

$$\tilde{a} = \sigma\tau(c - d_z) + \sigma(1 - \tau)\tilde{x}(c - d_n). \tag{7}$$

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 that 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

$$-[1 - F(\tilde{a})]\alpha = -[1 - F(\tilde{a})](\alpha - \sigma\gamma) - k, \tag{8}$$

which is the same as

$$\kappa - [1 - F(\tilde{a})] = 0. \tag{9}$$

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 in which $0 < \tilde{x} < 1$ requires that both (9) and (7) hold, which implies $\Gamma(\tilde{x}) = 0$. Now

$$\frac{k}{\sigma} < [F(0) - F(c - d_z)](-\alpha) + [1 - F(c - d_z)]\gamma < [1 - F(0)]\gamma$$
$$< [1 - F(\sigma\tau(c - d_z))]\gamma, \qquad (10)$$

which implies $\kappa < 1 - F(\sigma\tau(c - d_z))$. (The first inequality in [10] follows from assumption 3, the second follows from $c - d_z < 0$ and $\alpha > \gamma$, and 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 (6). Thus, there are two possible cases.

Case i: Equation (6) 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 the normal type invests with probability zero or one.) It is indeed an equilibrium because $\tilde{a} > 0$. This follows from

$$\kappa - [1 - F(0)] < 0 = \kappa - [1 - F(\tilde{a})],$$

where the inequality is due to (10) and the equality to (9).

Case ii: Equation (6) is violated. Then $\Gamma(1) \leq 0$, so we must have $\tilde{x} = 1$. It is indeed an equilibrium because (7) and assumption 1 yield

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

QED

It is intuitively clear that an unarmed B has an incentive to preserve ambiguity (deterrence by doubt): he does not want to invite attack from opportunists. In the equilibrium constructed in the proof of proposition 3, B also prefers to preserve ambiguity when armed, because if he reveals that he is armed, then A updates his beliefs and thinks that B is the crazy type. This out-of-equilibrium belief updating passes the D1 test of Banks and Sobel (1987). Indeed, a deviation in which B reveals that he is armed can make B better off only if it reduces the probability of attack, making it more likely that B gets δ_t instead of $-\alpha + \gamma$. But since $\delta_z > \delta_n$, it is the crazy type who would benefit more and hence be more inclined to deviate. So the D1 criterion implies that A should think that B is crazy after the deviation. Since the equilibrium satisfies D1, it passes every weaker test, such as the intuitive criterion of Cho and Kreps (1987).

In a more complex model, the armed B might not want to reveal his weapons, even if for some reason A would not update his beliefs. For example, suppose that we were to add a type of A who is a dove ($a <$

0) but whose payoff when attacking an armed B is $a - c'$, where

$$a - c' > -[\tau d_z + (1 - \tau)d_n]. \qquad (11)$$

This "tough dove" does not want to attack an unarmed B (since $a < 0$), but because of (11), he would attack an armed B when he assigns the prior probability to B being crazy. In this more complex (but possibly quite realistic) scenario, an armed B fears that revealing his weapons would invite attack from tough doves, even if their beliefs are not updated. In our simpler model, he fears attack because of the updated beliefs. As we have shown, the updating is quite reasonable and satisfies the standard refinements.

C. *Welfare Analysis*

In this subsection, we compare the payoffs under full disclosure with the payoffs under full ambiguity. With full disclosure B invests with probability one. If (6) holds, then ambiguity strictly reduces the risk of arms proliferation, in the sense that B invests with probability strictly less than one. Thus, if (6) holds, then hawks and doves strictly prefer full ambiguity to full disclosure. Even if (6) is violated, they weakly prefer ambiguity (as mentioned at the end of subsec. A, they gain nothing from inspections). However, opportunistic types who are uninformed will sometimes make "mistakes" and so may prefer disclosure.

It is useful to define

$$a^* \equiv \frac{\sigma d_n[c - \tau d_z - (1 - \tau)d_n]}{(1 - \sigma)c + \sigma d_n}. \qquad (12)$$

Assumption 1 implies

$$0 < a^* < \sigma\{c - [\tau d_z + (1 - \tau)d_n]\}, \qquad (13)$$

so type a^* is an opportunist. We distinguish two cases.

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

$$\kappa > 1 - F(a^*), \qquad (14)$$

where a^* is defined by (12). The inequalities (13) and (14) imply that (6) holds, so B invests with probability $\tilde{x} < 1$ behind the veil of ambiguity. It is not hard to see that the opportunistic type most likely to prefer disclosure is type $\tilde{a}$, defined by (7). The smaller $\tilde{x}$ is, the more likely it is that type $\tilde{a}$ prefers full ambiguity. Type $\tilde{a}$'s expected utility under full ambiguity is $\tilde{a} - \sigma[\tau + (1 - \tau)\tilde{x}]c$. From (7), type $\tilde{a}$ is an opportunist. Thus, with full disclosure he attacks if and only if B is unarmed, which happens with probability $1 - \sigma$; so type $\tilde{a}$'s expected payoff is

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

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

$$(1-\sigma)\tilde{a} - \sigma[\tau d_z + (1-\tau)d_n] < \tilde{a} - \sigma[\tau + (1-\tau)\tilde{x}]c. \qquad (15)$$

From the definition of $\tilde{a}$, (15) 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 (14). Thus, in case 1, ambiguity reduces the risk of arms proliferation sufficiently to make all of A's types strictly better off.

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

$$\kappa < 1 - F(a^*). \qquad (16)$$

If (6) is violated, then B invests with probability one under full ambiguity; so the opportunistic types strictly prefer full disclosure because it allows them to make better decisions. If (6) holds, then B invests with probability $\tilde{x} < 1$. However, by reasoning similar to that in case 1, we find that (16) implies that type $\tilde{a}$ strictly prefers full disclosure (inequality [15] is reversed). Thus, in case 2, whether or not (6) holds, some opportunistic types of A strictly prefer full disclosure to full ambiguity. Here ambiguity does not reduce the risk of arms proliferation by enough to make the opportunists better off.

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 changes by an amount

$$\sigma(\alpha - \gamma - \delta_t)[F(c - \tau d_z - (1-\tau)d_n) - F(\tilde{a})] \\ - (1-\sigma)\alpha[F(\tilde{a}) - F(0)] - \sigma\varepsilon. \qquad (17)$$

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 make the mistake of attacking B even though he is armed. 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 even though he is unarmed. Disclosure deters tough opportunists when B is armed, but ambiguity deters weak opportunists when B is unarmed. The third term is the expected cost of inspections. Without making further assumptions on the distribution of A's types, we cannot sign the expression in (17).

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 (14) holds. Player B prefers full ambiguity to full disclosure if and only if the expression in (17) is negative.

Suppose that case 2 applies. Some opportunistic types strictly prefer disclosure, whereas if (6) holds, then hawks and doves strictly prefer ambiguity. There is a conflict of interest among A's types. This suggests that there may be cheap-talk equilibria in which A demands inspections only when he is an opportunistic type who prefers disclosure. We consider this issue next.

IV. Communication Equilibrium

In this section we consider the role of communication in the form of cheap talk. At time 1, player A sends a message $m \in M$, where M is the set of feasible messages. We are interested in equilibria in which A's message is informative about his type, and B uses this information when deciding whether to invest and whether to reveal his military capability. (Since B has no incentive to reveal that he is crazy, we can assume without loss of generality that B sends no message.) Information about B's capabilities is useful to intermediate (opportunistic) types of A since they are ambivalent about whether or not to attack. Extreme types (hawks and doves) do not benefit from information revelation; they simply want to persuade B not to invest. Intermediate types also prefer if B remains unarmed, but they are willing to trade off an increased risk that B arms against a higher probability of information revelation.

These arguments lead us to the following type of cheap-talk equilibrium, called a *mixed inspections equilibrium*.[7] Only two messages are sent, a tough message and a conciliatory message. Intermediate types send the tough message, which induces B to invest and to reveal any weapons he acquires. Extreme types of A send the conciliatory message, following which B invests with probability strictly less than one but never reveals whether or not the investment succeeded.

More formally, in the mixed inspections equilibrium, there is a' and a'' such that player A sends the tough message if $a \in (a', a'')$ and the

[7] In the Appendix, we show that this is the only form a communications equilibrium can take when ε is small.

conciliatory message otherwise. It holds that

$$0 < a' < a'' < c - \tau d_z - (1 - \tau)d_n, \tag{18}$$

so all types in (a', a'') are opportunists. Player B's crazy type always invests. The normal type invests with probability one after the tough message but with probability $x \in (0, 1)$ after the conciliatory message. Player B allows inspections if and only if he hears the tough message and is armed. If an inspection reveals that B is unarmed, then A attacks if and only if $a > 0$. If after a tough message B refuses to allow inspections, then A attacks if and only if $a > 0$. If after a tough message B reveals that he is armed, then A attacks if and only if A is a hawk. If inspections are refused following a conciliatory message, then A attacks if $a \geq a''$ but not if $a \leq a'$.[8] Finally, if after a conciliatory message B reveals that he 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 and intuitive criteria).

It turns out that the mixed inspections equilibrium exists if and only if two conditions are satisfied. First, we must be in case 2 of Section III.B. Otherwise, all of A's types would prefer ambiguity, and no one would send the tough message. Thus, the first condition is that (16) must hold. Second, A should not be much more likely to be a hawk than a dove. Otherwise, B's fear of attack would induce him to invest for sure after the conciliatory message since it is sent by both hawks and doves. It is intuitively clear that if A is highly likely to be a hawk, then ambiguity will not prevent B from investing. Specifically, the second condition turns out to be

$$\frac{H}{H + D} < \kappa. \tag{19}$$

PROPOSITION 5. The mixed inspections equilibrium exists if and only if

$$\frac{H}{H + D} < \kappa < 1 - F(a^*). \tag{20}$$

Proof. Consider a mixed inspections equilibrium as defined above. To show that such an equilibrium exists, we must show that there exist a', a'', and x such that the given strategies are optimal for each player. (It is easy to see that a', a'', and x will be uniquely defined if they exist, so there can be only one mixed inspections equilibrium.) We will define a' and a'' in such a way that (18) holds. Notice that types in (a', a'') who send the tough message are opportunistic and will attack if inspections

[8] If a type in (a', a'') should "tremble" and send a conciliatory message, then he (off the equilibrium path) takes whatever action maximizes his payoff.

are refused or if inspections reveal that B is unarmed. If they discover that B is armed, then they will not attack.

The cutoff types a' and a'' must be indifferent between the two messages. Suppose that type a'' sends the tough message. Then B invests for sure, and type a'' will find out if the investment succeeds. With probability σ, B acquires weapons and type a'', who prefers not to attack because he is an opportunist, gets $-\tau d_z - (1-\tau)d_n > a'' - c$. With probability $1-\sigma$, B remains unarmed and type a'', who prefers to attack because he is an opportunist, 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 $\sigma[\tau + (1-\tau)x]$, but B will not reveal his weapons. 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

$$a'' = [\tau + (1-\tau)x]c - \tau d_z - (1-\tau)d_n < c - \tau d_z - (1-\tau)d_n, \qquad (21)$$

where the inequality holds as long as $x < 1$.

Similarly, 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

$$a' = \frac{(1-x)\sigma(1-\tau)d_n}{1-\sigma} > 0. \qquad (22)$$

Define

$$x^* \equiv \frac{(1-\sigma)\tau(d_z - c) + (1-\tau)d_n}{(1-\sigma)(1-\tau)c + \sigma(1-\tau)d_n} < 1, \qquad (23)$$

where the inequality follows from the first inequality in assumption 1. If $x = x^*$ is substituted into (21) and (22), we get $a' = a'' = a^*$, as defined in (12). Now (21) and (22) 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 that 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 $\delta_t - \varepsilon$, whereas 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 would increase the probability of an attack.

Next, consider the normal type's decision to invest. If he hears the tough message, then his expected payoff from investing is $\sigma(-\varepsilon) + (1-\sigma)(-\alpha) - k$. His expected payoff from not investing is $-\alpha$. He prefers to invest because

$$\sigma(\alpha - \varepsilon) - k > \sigma\gamma - k > 0,$$

where the first inequality follows from assumption 2 and the second from assumption 3.

Now consider the normal type's investment decision following the conciliatory message. If B hears the conciliatory message, then he thinks that 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

$$[1 - F(a'') + F(a')]\kappa - [1 - F(a'')] = 0. \tag{24}$$

We can use (21) and (22) to substitute for a' and a'' in (24). Define

$$\begin{aligned}\Psi(x) \equiv \Bigg\{&1 - F([\tau + (1-\tau)x]c - \tau d_z - (1-\tau)d_n)\\ &+ F\left(\frac{(1-x)\sigma(1-\tau)d_n}{1-\sigma}\right)\Bigg\}\kappa\\ &- \{1 - F([\tau + (1-\tau)x]c - \tau d_z - (1-\tau)d_n)\}.\end{aligned}$$

Notice that $\Psi(x^*) = \kappa - [1 - F(a^*)]$ and

$$\begin{aligned}\Psi(1) = {}&[1 - F(c - \tau d_z - (1-\tau)d_n) + F(0)]\kappa\\ &- [1 - F(c - \tau d_z - (1-\tau)d_n)].\end{aligned}$$

The indifference condition (24) is verified, together with (21) and (22), if x is chosen such that $\Psi(x) = 0$. Now (20) 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

$$\begin{aligned} a' - \sigma[\tau + (1-\tau)x]c &= a' - \sigma[a'' + \sigma\tau d_z + \sigma(1-\tau)d_n] \\ &< (1-\sigma)a' - \sigma[\tau d_z + (1-\tau)d_n], \end{aligned}$$

where the equality uses (21), 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

$$\begin{aligned} -\sigma\tau d_z - \sigma(1-\tau)xd_n &= -\sigma\tau d_z - \sigma(1-\tau)d_n + (1-\sigma)a' \\ &< -\sigma\tau d_z - \sigma(1-\tau)d_n + (1-\sigma)a'', \end{aligned}$$

where the equality uses (22), and the inequality is due to $a'' > a'$. The right-hand-side expression is what type a'' gets in equilibrium. QED

All of A's types weakly prefer the mixed inspections equilibrium to full disclosure, and there is strict preference for some. Indeed, any type of A can induce the same outcome 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 mixed inspections equilibrium.

Player B faces a dilemma with mixed inspections similar to that with full ambiguity. Consider his expected payoff. 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 mixed inspections 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 mixed inspections to full disclosure, B's expected payoff changes by

$$\begin{aligned} &\sigma(\alpha - \gamma + \delta_t)[F(c - \tau d_z - (1-\tau)d_n) - F(a'')] - \\ &(1-\sigma)[F(a') - F(0)]\alpha - \{1 - [F(a'') - F(a')]\}\varepsilon. \qquad (25) \end{aligned}$$

The interpretation is similar to (17). 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, even though B is armed with some probability, generating costly mistakes. Under full disclosure, they would have been 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 unarmed. They are deterred by doubt. Under full disclosure, they would always attack the unarmed B. The final term is the extra cost of inspections that must be borne in a full disclosure equilibrium.

Without making further assumptions on the distribution F, we cannot sign the expression in (25). But $\delta_z > \delta_n$ implies that if the crazy type prefers the mixed inspections equilibrium, then so does the normal type.

PROPOSITION 6. All of A's types prefer the mixed inspections equilibrium to full disclosure. Player B faces a trade-off. Both types of B prefer the mixed inspections equilibrium to full disclosure if and only if the expression (25) is negative for $t = z$.

V. Concluding Comments

In policy debates, it is often argued that the U.S. objective should be to prevent smaller powers from using strategic ambiguity (e.g., Schrage 2003). But if weak nations can practise "deterrence by doubt," then they have less incentive to acquire advanced weapons. Eliminating ambiguity may actually increase the risk of arms proliferation, contrary to conventional arguments.

If the leader of the big power (A) is an opportunist, he may prefer that the small power (B) submits to arms inspections even if this increases the risk of arms proliferation. Thus, in a communication equilibrium the opportunist sends a "tough" message that can be interpreted as requiring B to sign the NPT. Dovish types instead send "conciliatory" messages. Player B is less likely to acquire advanced weapons if he receives a conciliatory message. Unfortunately, hawks have an incentive to send a conciliatory message as well. If $H/(H + D)$ is large, then the conciliatory message will not reassure B, who suspects a "false dove," and the equilibrium breaks down. However, $H/(H + D)$ will be small if the advanced weapons are very powerful. Unless there is strong evidence that it is about to share its weapons with terrorists, a state armed with nuclear weapons is unlikely to be attacked, suggesting that H is small or even zero. In this case, the conciliatory message convinces B that A is friendly, making B willing to refrain from investing, as required for the equilibrium to exist.

Another important parameter is the normalized cost of investing, κ (which is higher the bigger the cost k of investing, the smaller the probability σ that investment succeeds, and the smaller the value of advanced weapons, γ). If κ is high enough that assumption 3 is violated, then B will not try to acquire new weapons, even if there is no ambiguity. In this case, the "conventional wisdom" is correct: ambiguity cannot benefit A. If assumption 3 holds but κ is so small that inequality (16) holds, then B very likely will try to acquire new weapons, whether there is ambiguity or not, so at least opportunistic types prefer inspections. But if κ is in the intermediate range, then ambiguity can make A strictly better off regardless of type.

In order for ambiguity to be a deterrence in our model, B has to invest with positive probability. It is sometimes argued that such mixed strategies are unlikely to be employed in the real world. Suppose, however, that the true κ is B's private information. This "purifies" the equilibrium in the sense that B will invest if and only if κ is below some cutoff value (which can depend on the message). If A cannot observe κ, he will not know if B has invested or not, which creates deterrence by doubt.

In reality, weapons programs require the completion of several steps. The small power may create ambiguity about how many steps, if any, it has completed. A state that has completed some but not all required steps may be vulnerable to a preemptive attack, but (if the weapons are sufficiently powerful) an attack may not be likely once all steps are completed. We plan to study this dynamic problem in future work. Another interesting possibility is that several big powers may interact with the same small power. Then communication involves multiple senders as in Krishna and Morgan (2001) or Battaglini (2002). A dovish big power wants to minimize the risk of arms proliferation, whereas a more opportunistic big power may want inspections. An analysis of such many-player interactions is also left for future work.

Appendix

A. *Proof of Proposition 1*

Fix any equilibrium. Let $x(m, t)$ denote the probability that type $t \in \{z, n\}$ invests after player A has sent message $m \in M$, where M is the message space. 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 $x(m, n) = 0$.

We will prove proposition 1 by contradiction. Suppose, contrary to the proposition, that the crazy type of player B invests with probability strictly less than one. That means $x(m, z) < 1$ for some m. Let $M^* \subseteq M$ be the set of messages that minimize $x(m, z)$. If $m^* \in M^*$, then $x(m^*, z) < 1$ and $x(m^*, n) = 0$. Four claims will be needed to derive the contradiction.

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

Proof. 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^*, and player A attacks if and only if $a \geq 0$, which happens with probability $1 - F(0)$. Suppose that B deviates to a strategy in which he invests and refuses inspections if the investment succeeds. If the investment fails, he behaves exactly as he would have done had he not invested. 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

$$\sigma\{[1 - F(c - d_z)](-\alpha + \gamma) - [1 - F(0)](-\alpha)\} - k.$$

This is strictly positive by assumption 3, a contradiction. QED

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

Proof. If $a - c > -d_n$, then A always attacks; and if $a - c < -d_z$, then A never attacks. In either case, he wants to minimize the probability that B invests and therefore sends a message in M^*. Finally, suppose $0 < a \leq c - d_n$. 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 that 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 wants to attack (since $a > 0$). If B is armed, then A wants 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

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

since $a > 0$, $d_n > 0$, and $x(m^*, z) < x(m', z)$. So sending $m' \notin M^*$ is strictly worse than sending $m^* \in M^*$. QED

Notice that claim 2 establishes that $M^* \neq \emptyset$. In other words, $\inf\{x(m, z) : m \in M\}$ must be attained by some m, otherwise A would not have a best response.

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

Proof. To obtain a contradiction, suppose that, following $m^* \in M^*$, there is a positive probability that inspections reveal that B is armed. Since only crazy types invest ($x(m^*, z) > 0 = x(m^*, n)$), B will be known to be crazy. 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 it would cost $\varepsilon > 0$ but would not reduce the probability of attack. This is a contradiction. QED

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

Proof. Suppose that there exists $m^* \in M^*$ such that if B receives m^* and is unarmed, then he allows inspections with probability one. By claim 3, it follows that if B refuses inspections, he will be known to be armed. Therefore, following message m^*, there is no ambiguity about B's weapons. In this case, all of A's types must send a message in M^*, and since $x(m^*, n) = 0$ for $m^* \in M^*$, the normal type never invests. Since there is no deterrence by doubt, the unarmed B is attacked whenever $a > 0$. Suppose that B's normal type deviates from the equilibrium and invests. If the investment fails, he behaves just as he would have done had he not invested. If the investment succeeds, he refuses inspections. Now, in no circumstance would A ever attack if $a < c - d_z$, so the probability that B is attacked is at most $1 - F(c - d_z)$. Therefore, B's payoff is increased by at least

$$\sigma\{[1 - F(c - d_z)](-\alpha + \gamma) - [1 - F(0)](-\alpha)\} - k,$$

which, by assumption 3, is strictly positive. This is a contradiction. QED

Now we can establish the contradiction that proves proposition 1. Let π^* denote the expected equilibrium payoff for the normal type of B, conditional on the event that some message in M^* is sent. Following any $m^* \in M^*$, B's normal type will remain unarmed ($x(m^*, n) = 0$), and the unarmed B weakly prefers to refuse inspections (by claim 4). Thus, to calculate π^*, we may assume that B remains unarmed and refuses inspections. Then B is attacked with some probability, which we denote ζ, so $\pi^* = -\zeta\alpha$. By claim 2, A sends a message in M^* if $a > 0$ and if $a < c - d_z$. Following any $m^* \in M^*$, $x(m^*, z) > 0 = x(m^*, n)$, so B will be either unarmed or armed and crazy. In either case, type $a > 0$ prefers to attack. Type $a < c - d_z$ never attacks. It follows that

$$1 - F(0) \leq \zeta \leq 1 - F(c - d_z).$$

We aim to show that a deviation in which, after each $m^* \in M^*$, B's normal type invests and then refuses inspections makes him strictly better off. Conditional on the event that some message in M^* is sent, the normal type's expected payoff if he deviates in this way is

$$\begin{aligned}\sigma\zeta(-\alpha + \gamma) + (1 - \sigma)\zeta(-\alpha) - k &\geq \sigma[1 - F(c - d_z)](-\alpha + \gamma) + (1 - \sigma)\zeta(-\alpha) - k\\ &> -\sigma[1 - F(0)]\alpha - (1 - \sigma)\zeta\alpha \geq -\zeta\alpha = \pi^*,\end{aligned}$$

where the first inequality is due to $\zeta \leq 1 - F(c - d_z)$, the second to assumption 3, and the third to $1 - F(0) \leq \zeta$. Therefore, the deviation makes him strictly better off, a contradiction.

B. The Set of All Equilibria of the Game

We will show that when ε is small, we are justified in focusing on the three equilibria discussed in the text: full disclosure, full ambiguity, and mixed inspections. Our method of proof is to fix any arbitrary strategy profile. We show that if this is an equilibrium, then if ε is small enough, in terms of all payoff-significant variables (including investment and attack probabilities), it must be arbitrarily close to one of the three equilibria discussed in the text. It may differ in the probability of inspections, but this difference becomes payoff irrelevant when ε is small. Thus, in the idealized case in which inspections do not consume significant real resources, nothing is lost by focusing on the full disclosure, full ambiguity, and mixed inspections regimes.

To see how there can exist equilibria that are almost the same as one of the three discussed in the text, but with a different probability of inspections, consider modifying the full disclosure equilibrium from Section III.A in the following way. When A's type is indifferent with respect to inspections, because they would not influence his decision to attack anyway, let him send a message that allows B to cancel inspections in those situations in which they would not have altered the outcome anyway. The probability of B arming and A attacking will be the same as in the full disclosure equilibrium of Section III.A. There will be fewer inspections, but this is payoff irrelevant for small ε. Similarly, we can modify the other two equilibria discussed in the text in a "payoff-irrelevant" way (for small ε). We will show that no other equilibria can exist.

The arguments require some notation. If B is unarmed, then whether he is normal or crazy is payoff irrelevant to both A and B. Therefore, there is no reason to distinguish the unarmed normal type from the unarmed crazy type. Abusing terminology, then, 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 that B is *unarmed*.

Fix an equilibrium. Let $A(m)$ be the set of types that send message m in this equilibrium. (That is, $a \in A(m)$ if type a sends m.) We can assume, without loss of generality, that every message in M is sent by some type (other messages can simply be dropped). Thus, $A(m) \neq \emptyset$ for all m. Let $I(m, t)$ be the probability that player B allows inspections following message m when his ex post type is $t \in \{z, n, u\}$.

Proposition 1 showed that the crazy type always invests, $x(m, z) = 1$. To simplify notation, we let $x(m)$ (rather than $x(m, n)$) denote the probability that the normal type of player B invests when player A sends $m \in M$.

Lemma 1. For any $m \in M$, $I(m, z) \geq I(m, n)$ (and equality can hold only if $I(m, z) = I(m, n) = 0$ or $I(m, z) = I(m, n) = 1$).

Proof. For B to be willing to pay $\varepsilon > 0$ to allow inspections, they must strictly reduce the probability of attack. Type z earns $\delta_z > \delta_n$ when not attacked, so he is strictly more willing to allow inspections than type n. In particular, if (given m and ε) type n is indifferent between allowing and refusing inspections, then type z strictly prefers inspections. QED

Let M^A be the set of messages such that, if $m \in M^A$ is sent, then there is positive probability that inspections reveal that B is armed. Since the crazy type always invests (proposition 1) and is more likely to allow inspections (lemma 1),

$$M^A = \{m \in M : I(m, z) > 0\}.$$

Let $M^U = \{m \in M : I(m, u) > 0\}$ be the set of messages such that, if $m \in M^U$ is sent, then there is positive probability that inspections reveal that B is unarmed. Let M^R be the set of messages such that, if $m \in M^R$ is sent, then there is positive probability that inspections are refused. Notice that $M \backslash (M^A \cup M^U)$ is the set of messages that cause B to refuse inspection with probability one.

If B receives message $m \in M$ and allows inspections, and these reveal that B is armed, then the probability that B is crazy is

$$\tau(m, \text{allow, armed}) \equiv \frac{\tau I(m, z)}{\tau I(m, z) + (1-\tau)x(m)I(m, n)}. \tag{A1}$$

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

$$\tau(m, \text{refuse, armed}) \equiv \frac{\tau[1 - I(m, z)]}{\tau[1 - I(m, z)] + (1-\tau)x(m)[1 - I(m, n)]}. \tag{A2}$$

The expressions (A1) and (A2) are well defined only along the equilibrium path (i.e., as long as the denominator is nonzero).

Lemma 2. For any $m \in M$, the following conditions hold: (i) $\tau(m,$ allow, armed$) \geq \tau \geq \tau(m,$ refuse, armed$)$ whenever (A1) and (A2) are well defined; (ii) $M^A \cap M^U \cap M^R = \emptyset$; (iii) if type $a \in A(m)$ sends message $m \in M^A$ (respectively $m \in M^U$) and then attacks when inspections reveal that B is armed (respectively unarmed), then he must also attack if inspections are refused.

Proof. Part i: follows from lemma 1 and equations (A1) and (A2).

Part ii: To obtain a contradiction, suppose $m \in M^A \cap M^U \cap M^R$. Thus, when m is sent, both armed and unarmed types allow inspections with positive probability, yet there is also a positive probability that inspections are refused. If the unarmed B allows inspections, then he is attacked if and only if A is not a dove. For him to be willing to incur the inspection cost, the probability of attack must be strictly higher if there is no inspection; so some dove must send m and then attack if B refuses inspections. Thus, if a^* is the lowest type in $A(m)$ who attacks when there is no inspection, we must have $a^* < 0$. Type $a^* < 0$ has only one motive for attack: to eliminate the threat from type z. He wishes to attack neither type u nor type n. But type z is more likely to allow inspections than type n, by lemma 1. So if type $a^* < 0$ is willing to attack when there is no inspection, then he is certainly willing to attack when inspections reveal weapons; the same is true for any higher type. Therefore, if inspections reveal weapons, the probability of attack is no less than if inspections are refused, so the armed B would never allow inspections, a contradiction.

Part iii: Since the benefit of attacking is increasing in a, at time 5 there is a cutoff type such that A attacks if and only if his type is above this cutoff. To give B an incentive to pay the inspections cost, the cutoff type must be lower when inspections are refused than when they are allowed, which implies part iii. QED

Subsection 1 characterizes equilibria in which A's message influences the probability that B invests. Any such equilibrium must have the same structure as the mixed inspections regime and yield virtually the same expected payoff when ε is small. Subsection 2 considers equilibria in which A's message does not influence the probability that B invests (although it may influence the probability of inspections). Such equilibria must be essentially equivalent to either full disclosure or full ambiguity when ε is small. The proofs are somewhat involved because, as mentioned above, if A does not intend to base his attack decision on the outcome of the inspections, then he is indifferent with respect to inspections. He may then ask for inspections with some probability, generating equilibria that are identical except that this probability is different. However, if ε is small, these equilibria must generate approximately the same payoff.

1. Equilibria in Which A's Message Influences B's Investment Decision

In this subsection, we consider equilibria in which $x(m)$ is not constant across M; that is, B's investment decision depends on A's message. We will first show (proposition 7) that any such equilibrium must have the same structure as the mixed inspections equilibrium of Section IV. In particular, only two messages are sent, one tough and one conciliatory.[9] Proposition 8 will confirm that, for small ε, the probabilities of investment, attack, and expected payoffs must be virtually the same as in the mixed inspections equilibrium.

Proposition 7. Consider any equilibrium such that A's message influences B's investment decision. There exist a' and a'' such that player A sends a tough message if $a \in (a', a'')$ and a conciliatory message otherwise, where

$$0 \leq a' < a'' \leq c - \tau d_z - (1 - \tau) d_n.$$

After hearing the tough message, B invests and allows inspections if and only if he is armed. After hearing the conciliatory message, the normal type of B arms with positive probability and refuses inspections with positive probability if he is armed. Player B never allows inspections when unarmed.

Proof. Recall that the crazy type of B invests with probability one, and $x(m)$ denotes the probability that the normal type invests following message $m \in M$. By Bayes' rule, the probability that B is crazy conditional on being armed and message m having been sent is

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

The set of messages that minimize $x(m)$ is denoted M^C and is interpreted as the set of conciliatory 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. (Thus, the minimum must be attained.) We interpret $M^T \equiv M \backslash M^C$ as the set of tough messages. By definition, if $m^C \in M^C$ and $m^T \in M^T$, then $x(m^C) < x(m^T) \leq 1$. By hypothesis, $x(m)$ is not constant, so $M^C \neq \emptyset$ and $M^T \neq \emptyset$. Recall that $A(m) \neq \emptyset$ for all $m \in M$ (since unused messages are dropped). We will show that a tough message induces B to invest and to remove ambiguity, whereas a conciliatory message leads to less investment but more ambiguity.

The proof has nine steps. The first three steps characterize what happens following a tough message.

Step 1: Suppose that type a sends $m \in M^T$. With positive probability, events at time 4 cause type a to attack at time 5; with positive probability, events at time 4 cause him to refrain from attacking.

[9] The labeling of a message is arbitrary: a sentence could be said in English or in French. Therefore, it is more precise to say that A's message space is divided into two "equivalence classes": one containing tough messages and the other containing conciliatory messages. Within each equivalence class, each message leads to exactly the same outcome.

Proof. The decision to attack must depend on what happens at the inspection stage, for otherwise type a could increase his expected payoff by sending a message in M^C (which would reduce the probability that B invests). This proves step 1.

Step 2: $M^T \subseteq M^A \backslash M^U$.

Proof. For A to have any incentive to send a message in M^T, these messages must trigger inspections with positive probability. That is, $M^T \subseteq M^A \cup M^U$. We now show that $M^T \subseteq M^A$. Suppose not, so there is $m \in M^T$ such that $m \in M^U \backslash M^A$. Now part iii of lemma 2 and step 1 imply that every type in $A(m)$ must attack when there is no inspection, but not when inspections reveal that B is unarmed. But then, by assumption 2, the best response to m for a normal type of B is to refrain from investing and reveal that he is unarmed. So $x(m) = 0$, which certainly must mean that $m \in M^C$, a contradiction. So we must have $M^T \subseteq M^A$ after all. Suppose $m \in M^A \cap M^U$. Then $m \notin M^R$ by part ii of lemma 2, so inspections are never refused following m. To induce B to allow inspections when both armed and unarmed, there must be a type $a' \in A(m)$ that does not attack when inspections reveal weapons and a type $a'' \in A(m)$ that does not attack when inspections reveal that B is unarmed. Since there is always an inspection, type $a^* = \min\{a', a''\}$ never attacks, contradicting step 1. This completes the proof of step 2.

Step 3: For all $m^T \in M^T$, the following conditions hold: (i) $I(m^T, u) = 0$; (ii) no $a \in A(m^T)$ will attack if inspections reveal that B is armed; (iii) each $a \in A(m^T)$ will attack if B refuses inspections; (iv) $I(m^T, z) = I(m^T, n) = 1$; (v) if $a \in A(m^T)$, then a is an opportunistic type; and (vi) $x(m^T) = 1$.

Proof. Part i: This follows from $M^T \subseteq M^A \backslash M^U$.

Part ii: If type $a \in A(m^T)$ attacks when inspections reveal weapons, by part iii of lemma 2, he always attacks, which contradicts step 1.

Part iii: Step 1 and part ii imply that inspections must be refused with positive probability; when this happens, each $a \in A(m^T)$ must attack.

Part iv: By assumption 2 and parts ii and iii, B prefers inspections when armed.

Part v: A dove would prefer not to attack when an inspection is refused, since B is unarmed in this case, by parts i and iv. So part iii implies that no dove is in $A(m^T)$. A hawk would surely prefer to attack if inspections reveal that B is armed. So part ii implies that no hawk is in $A(m^T)$.

Part vi: 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 $\sigma(\delta_t - \varepsilon) - (1 - \sigma)\alpha - k$. By not investing he gets $-\alpha$ for sure (by parts iii and v). Assumptions 2 and 3 imply that the former expression is strictly greater, so he strictly prefers to invest. This completes the proof of step 3.

The next four steps characterize what happens following a conciliatory message.

Step 4: $M^C \subseteq M^R$.

Proof. Message $m^C \in M^C$ must generate some ambiguity about B's weapons, or else A would never want to send a message in M^T. This proves step 4.

Step 5: $M^C \cap M^U = \varnothing$.

Proof. To obtain a contradiction, suppose that there is $m^C \in M^C \cap M^U \subseteq M^R \cap M^U$ (where the inclusion is due to step 4). By part ii of lemma 2, $m^C \notin M^A$. Now, if inspections reveal that B is unarmed, then all types with $a > 0$ attack.

Since $m^C \in M^U$, the unarmed must be willing to allow inspections, so the inspections must strictly reduce the probability of attack. This implies that some doves in $A(m^C)$ 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 (since $m^C \notin M^A$). If instead they send $m^T \in M^T$ and never attack, they get $-\sigma[\tau d_z + (1 - \tau)d_n]$. Thus, for them to prefer m^C, we need

$$-\sigma[\tau + (1 - \tau)x(m^C)]c \geq -\sigma[\tau d_z + (1 - \tau)d_n]. \tag{A3}$$

If type $a > 0$ sends $m^T \in M^T$, then step 3 implies that 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 (A3). 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, contradicting part v of step 3. This proves step 5.

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

Proof. Message $m^C \in M^C$ must generate some ambiguity. Since B never allows inspections when unarmed (step 5), he must sometimes also refuse inspections when armed, otherwise there is no ambiguity. In view of lemma 1, $I(m^C, n) < 1$. This proves step 6.

Step 7: If $m^C \in M^C$, then $0 < x(m^C) < 1$.

Proof. By definition of M^C, if $m^C \in M^C$ and $m^T \in M^T$, then $x(m^C) < x(m^T) = 1$. Suppose $x(m^C) = 0$. That is, after message m^C, only crazy types invest. Clearly, any type $a > 0$ prefers to send m^C rather than m^T if $x(m^C) = 0$ and $x(m^T) = 1$. This contradicts part v of step 3. Thus, $x(m^C) > 0$. This proves step 7.

The next step shows that there is no need to consider more than one tough and one conciliatory message.

Step 8: Without loss of generality, we can assume $M^T = \{m^T\}$ and $M^C = \{m^C\}$.

Proof. Step 3 implies that all messages in M^T generate exactly the same behavior for B, so we can also assume $M^T = \{m^T\}$. Next, we show that the same is true for M^C. By definition, each message in M^C induces B to invest with the same probability, say $x(m^C) = x^*$ for all $m^C \in M^C$. By step 5, following m^C, B never allows inspections when he is unarmed. In this case, if both the crazy type and the normal type never allow inspections after all messages in M^C, we are done. Thus, if all messages in M^C do not produce exactly the same outcome, then there must exist two messages in M^C, say m' and m'', that differ only in the sense that m' induces a higher probability of inspections when B is armed. Now, if an armed B is willing to allow inspections, they must reduce the probability of attack. So, there must be types who send messages in M^C and attack if and only if inspections are refused. These types will send the message m' rather than the message m''. 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. This proves step 8.

Finally, we can complete the proof of the proposition.

Step 9: There exist cutoff points a' and a'', where $0 \leq a' \leq a'' \leq c - \tau d_z - (1 - \tau)d_n$, such that A sends m^T if $a \in (a', a'')$ and m^C if $a < a'$ or $a > a''$.

Proof. Regardless of message, B never allows inspections when unarmed (part i of step 3 and step 5). If he is armed, he always allows inspections following a tough message (part iv of step 3) but refuses with some probability following a conciliatory message (step 6). The probability that B allows inspections is summarized in the following matrix:

	Armed	Unarmed
m^T	1	0
m^C	< 1	0

Thus, following the tough message there will be no ambiguity about B's military capability, but following the conciliatory message there will be ambiguity. However, $0 < x(m^C) < x(m^T) = 1$ (by part vi of step 3 and step 7), so B is more likely to invest following the tough message. Thus, from A's point of view, the tradeoff is between more investment but less ambiguity on the one hand (message m^T) and less investment but more ambiguity on the other (message m^C). Types of A who take the same action regardless of what happens at time 4 do not suffer from ambiguity; they only want to encourage B not to invest. Thus, they send m^C. This includes low a types who never attack as well as high a types who always attack (the "nonconvexity"). The types who send m^T, however, must be opportunistic types who base their decision on what happens at time 4 (parts ii, iii, and v of step 3).

More formally, let $a' = \inf A(m^T)$ and $a'' = \sup A(m^T)$. Part v of step 3 established that $0 \leq a' \leq a'' \leq c - \tau d_z - (1 - \tau)d_n$. Types a' and a'' are indifferent between m^T and m^C. Types between a' and a'' are more ambivalent about attacking than types either below a' or above a'', so they send m^T in order to maximize the probability of inspection. Types below a' and above a'' send m^C by definition of a' and a''. This proves step 9. QED

We can now show the following proposition.

Proposition 8. Consider any equilibrium such that A's message influences B's decision to invest. If ε is small, then the probability of investment and attack, and the expected payoff for each player, must be almost the same as in the mixed inspections equilibrium, the difference vanishing as $\varepsilon \rightarrow 0$.

Proof. By proposition 5 and lemma 1, either

$$I(m^C, z) = I(m^C, n) = 0 \tag{A4}$$

or

$$1 > I(m^C, z) > I(m^C, n) = 0 \tag{A5}$$

or

$$1 = I(m^C, z) > I(m^C, n) > 0, \tag{A6}$$

where m^C is the conciliatory message.

If (A4) holds, then there is never an inspection following the conciliatory message, and the equilibrium is precisely the mixed inspections equilibrium from Section IV.

If (A5) holds, then by allowing inspections, B reveals that he is not only armed but crazy. This cannot reduce the probability of attack; hence B has no reason to allow inspections in this case, contradicting $I(m^C, z) > 0$. So (A5) cannot hold.

Finally, suppose that (A6) holds. To give B a reason to allow inspections, some type $\tilde{a} \in A(m^C)$ must attack when there is no inspection, but not when inspections reveal weapons. The probability that inspections reveal weapons is strictly less than σ because the normal type does not always invest $(x(m^C) < 1)$; and even when he is armed, he sometimes refuses inspections $(I(m^C, n) < 1)$. Therefore, type $\tilde{a}$ attacks with probability strictly greater than $1 - \sigma$. Types in the interval (a', a'') send m^T and attack with probability $1 - \sigma$. Incentive compatibility requires that the probability that type a attacks is nondecreasing in a. Therefore, $\tilde{a} \geq a''$. In other words, types below a' send m^C but they do not attack, whether or not there is an inspection. There must be $\hat{a} > a''$ such that if inspections reveal that B is armed, type $a \in A(m^C)$ attacks if $a > \hat{a}$, but not if $a < \hat{a}$. In particular, types between a'' and $\hat{a}$ refrain from attacking if inspections reveal that B is armed, but they attack if there is no inspection. (This is what gives B the incentive to allow inspections when armed.)

Since $0 < I(m^C, n) < 1$, the normal type is indifferent between allowing and refusing inspections when he is armed and receives message m^C. If the cost of inspections is small, then this indifference requires that the probability of attack is reduced very slightly by inspections. This requires that $\hat{a}$ and a'' are very close (for small ε). Type $\hat{a}$ must be indifferent between attacking and not attacking. Since a'' is very close to $\hat{a}$, type a'' must be almost indifferent as well. Therefore, to compute the approximate expected payoff for a'', we can assume that a'' attacks when the inspections reveal that B is armed, as well as when there is no inspection. This implies that the conditions for a'' to be indifferent between m^C and m^T must be approximately equation (21). For a', the indifference condition is equation (22).

Recall that B's normal type is indifferent between allowing and not allowing inspections when he is armed and receives m^C. To calculate his expected payoff, we may assume that he refuses inspections. If he refuses inspections, he is attacked approximately when $a > a''$ but not when $a < a'$; so the condition for him to be indifferent between investing and not investing is approximately $\Psi(x) = 0$, where the function Ψ was defined in the proof of proposition 5.

In summary, the equations that determine a', a'', and x are similar to the equations in the proof of proposition 5, and the similarity becomes exact if $\varepsilon \rightarrow 0$. It is clear from the equations that the probability of investment and attack, and the expected payoff for each player, must be almost the same as in the mixed inspections equilibrium, the difference vanishing as ε vanishes. QED

2. Equilibria in Which A's Messages Do Not Influence B's Investment Decision

Consider an equilibrium in which A's message does not influence B's investment decision. Player A faces no trade-off between the probability of investment and the probability of inspections since the former probability is independent of the message. Either A does not care about inspections, in which case he is indifferent with respect to all messages, or else he simply wants to maximize the probability of inspections.

There are three possibilities for this kind of equilibrium. The first possibility is that there exists a message such that, if A sends it, then there is no ambiguity about B's military capability: with probability one, B will reveal whether he is armed or not (either by always allowing inspections or by always allowing inspections when armed or always when unarmed). In this case, the equilibrium must be essentially equivalent to the full disclosure equilibrium of Section III.A. To see this, notice that any type of A who wants to base his decision to attack on whether or not B is armed will send a message that eliminates all ambiguity. Therefore, the probability of investment and attack must be exactly as described in Section III.A. However, the probability of inspections may be different. The reason is that if A does not plan to base his attack decision on whether or not B is armed, then A is indifferent with respect to inspections and might send a message that triggers inspections only with some probability. But when ε is small, the probability of inspections does not significantly influence payoffs. What matters is the probability of investment and the probability of attack, which are the same as in Section III.A.

The second possibility is that, regardless of the message, B refuses inspections with probability one. In this case, messages play no role and the equilibrium must be the full ambiguity equilibrium of Section III.B.

It remains only to consider the third and final possibility, namely, an equilibrium that satisfies the following three properties:

E1. A's message never influences B's investment decision.

E2. There is no message that eliminates all ambiguity with probability one.

E3. There is some message that triggers inspections with positive probability.

Thus, fix an equilibrium satisfying properties E1, E2, and E3. We will make four claims about this equilibrium. We first claim that no message will induce both armed and unarmed types of B to allow inspections.

Claim 1. $M^R = M$ and $M^A \cap M^U = \emptyset$.

Proof. Since some ambiguity must always exist by property E2, we have $M = M^R$. Part ii of lemma 2 now implies that $M^A \cap M^U = \emptyset$. QED

Claim 1 implies that the three sets M^A, M^U, and $M \backslash (M^A \cup M^U)$ partition M.

Claim 2. If $M^U \neq \emptyset$, then each opportunistic type attacks with positive probability. Each opportunistic type who sends a message in $M \backslash (M^A \cup M^U)$ attacks with probability one.

Proof. If the messages do not influence the probability that B invests, then an opportunistic type who never attacks can do strictly better by sending a message in M^U and attacking when inspections reveal that B is unarmed. This proves that he must attack with positive probability. Now suppose that an opportunist sends a message in $M \backslash (M^A \cup M^U)$. Since there is no inspection, he must attack or not, with no information. If he does not attack, then as before, he would have been better off sending a message in M^U and attacking when inspections reveal that B is unarmed. QED

Claim 3. $M^U = \emptyset$.

Proof. Assume that $M^U \neq \emptyset$ in order to derive a contradiction. Claim 1 implies that inspections following $m \in M^U$ always reveal that B is unarmed, so B is attacked whenever $a > 0$. Now, for unarmed B to be willing to incur the cost of inspections, some $a < 0$ must attack if inspections are refused. But a dove's only reason to attack is to eliminate the threat from the crazy type. Specifically, if there is no inspection following message m, then type $a < 0$ is willing to attack only if $\tau(m, \text{refuse, armed}) > \tau$. This requires $x(m) < 1$ (from the formula for $\tau(m, \text{refuse, armed})$ and the fact that $I(m, n) = I(m, z) = 0$ by claim 1).

Consider what happens when B is unarmed. If $m \in M^U$ is sent, then in order to calculate B's expected payoff, we may suppose that the unarmed B allows inspections (since this is a best response); so he is attacked by any type $a \in A(m)$ with $a > 0$. If $m \in M \backslash (M^A \cup M^U)$ is sent then by claim 2, any opportunist in $A(m)$ will attack (and certainly hawks as well). Finally, if $m \in M^A$ is sent, then the unarmed B will reject inspections (by claim 1). Suppose that for every $m \in M^A$ every type $a \in A(m)$ with $a > 0$ attacks when there is no inspection. In this case, we conclude that for any m (whether in M^U, M^A, or $M \backslash (M^A \cup M^U)$), all types $a \in A(m)$ such that $a > 0$ attack the unarmed B. But then (under assumptions 2 and 3) B strictly prefers to invest, which contradicts $x(m) < 1$. Therefore, the supposition was wrong: there must be $m \in M^A$ and $a \in A(m)$ with $a > 0$ such that type a does not attack if there is no inspection. If there is an inspection that reveals weapons, then this type must attack, for otherwise he never attacks, which contradicts claim 2. However, this behavior contradicts part iii of lemma 2. QED.

Claim 4. There must be an interval $[a', a'']$ such that all types in this interval send the same message $m^A \in M^A$, attack when inspections are refused, but do not attack when inspections reveal weapons. Types below a' never attack (and are indifferent across all messages), and types above a'' always attack (and are also indifferent across all the messages).

Proof. By claim 3, $M^U = \emptyset$. By property E3, some message triggers inspections with positive probability, so $M^A \neq \emptyset$. By claim 1, $M = M^R$. Lemma 1 implies that for $m \in M^A$, either (i) $I(m, z) > I(m, n) = 0$ or (ii) $1 = I(m, z) > I(m, n) > 0$.

If condition i applies, then inspections reveal not only that B is armed but also that he is crazy. However, the armed and crazy type is the type A most wants to attack, so revealing that you are armed and crazy cannot reduce the probability of attack. Therefore, B would not allow inspections, so condition i cannot happen.

Suppose instead that condition ii applies. To induce the armed B to allow inspections, there must be a type $a^* \in A(m^A)$ who attacks when there is no inspection, but not when inspections reveal weapons. Since type a^* prefers not to attack when inspections reveal weapons, not knowing if A is crazy or not, he certainly would like to avoid attacking a normal and armed B. Therefore, for a^* to be willing to send m^A, it must maximize $I(m, n)$ among all messages in M^A (since this reduces the chance that the normal and armed type is attacked). But, since m^A was an arbitrary message in M^A, all messages in M^A must induce the same probability of inspection. They are therefore equivalent (up to labeling), and we can assume $M^A = \{m^A\}$. Moreover, it is clear that the types who send the message m^A and then attack when there is no inspection, but not when inspections reveal weapons, must lie in some interval (a', a''). Higher types always attack, and lower types never attack. QED

Finally, we can state a proposition.

Proposition 9. Consider any equilibrium such that A's message does not influence B's decision to invest. If ε is small, then the probability of investment and attack, and the expected payoff for each player, must be almost the same as in either full disclosure or full ambiguity, the difference vanishing as $\varepsilon \rightarrow 0$.

Proof. As argued above, it suffices to consider an equilibrium satisfying properties E1, E2, and E3 and thus characterized in claim 4. Following $m^A \in M^A$, the armed B refuses inspections with probability strictly between zero and one (recall that $M = M^R$ by claim 1, so there must always be a chance of refusal). Player B's indifference condition requires that if the inspections cost is small, the probability of attack must be approximately the same, whether or not inspections occur. Thus, a' and a'' must be very close, and B is attacked with probability approximately $1 - F(a')$, whether or not there is an inspection. Types below a' never attack, and (disregarding the very small interval $[a', a'']$) types above a' always attack. The outcome is therefore approximately the full ambiguity equilibrium of Section III.B. In particular, B must invest with probability close to $\tilde{x}$, and both a' and a'' must be close to $\tilde{a}$, where $\tilde{x}$ and $\tilde{a}$ are as defined in the proof of proposition 3. An attack occurs approximately with probability $1 - F(\tilde{a})$, just as in the full ambiguity equilibrium of Section III.B. As $\varepsilon \rightarrow 0$, each player's expected payoff must converge to the payoff of the full ambiguity equilibrium. QED

References

Ayres, Ian, and Steven D. Levitt. 1998. "Measuring Positive Externalities from Unobservable Victim Precaution: An Empirical Analysis of Lojack." *Q.J.E.* 113 (February): 43–77.

Baliga, Sandeep, and Tomas Sjöström. 2004. "Arms Races and Negotiations." *Rev. Econ. Studies* 71 (April): 351–69.

Banks, Jeffrey S., and Joel Sobel. 1987. "Equilibrium Selection in Signaling Games." *Econometrica* 55 (May): 647–62.

Battaglini, Marco. 2002. "Multiple Referrals and Multidimensional Cheap Talk." *Econometrica* 70 (July): 1379–1401.

Bond, Philip. 2004. "Bank and Nonbank Financial Intermediation." *J. Finance* 59 (December): 2489–2530.

Cho, In-Koo, and David M. Kreps. 1987. "Signaling Games and Stable Equilibria." *Q.J.E.* 102 (May): 179–221.

Crawford, Vincent P., and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica* 50 (November): 1431–51.

Gordon, Michael R., and Bernard E. Trainor. 2006. *Cobra II: The Inside Story of the Invasion and Occupation of Iraq.* New York: Pantheon.

Green, Jerry R., and Nancy L. Stokey. 2007. "A Two-Person Game of Information Transmission." *J. Econ. Theory* 135 (July): 90–104.

Grossman, Sanford J. 1981. "The Informational Role of Warranties and Private Disclosure about Product Quality." *J. Law and Econ.* 24 (December): 461–83

Hecker, Siegfried S. 2004. "Prepared Statement." In *An Update on North Korean Nuclear Developments: Hearing before the Committee on Foreign Relations,* U.S. Senate. 108th Cong., 2nd sess., January 21.

Krishna, Vijay, and John Morgan. 2001. “A Model of Expertise.” *Q.J.E.* 116 (May): 747–75.

Lott, John R., and David B. Mustard. 1997. “Crime, Deterrence, and Right-to-Carry Concealed Handguns.” *J. Legal Studies* 26 (January): 1–68.

Matthews, Steven A. 1989. “Veto Threats: Rhetoric in a Bargaining Game.” *Q.J.E.* 104 (May): 347–69.

Myre, Gregory. 2006. “In a Slip, Israel’s Leader Seems to Confirm Its Nuclear Arsenal.” *New York Times* (December 12), A5.

Norris, Robert S., William M. Arkin, Hans M. Kristensen, and Joshua Handler. 2002. “Pakistan’s Nuclear Forces, 2001.” *Bull. Atomic Scientists* 58 (January/February): 70–71

Norris, Robert S., and Hans M. Kristensen. 2005. “India’s Nuclear Forces, 2005.” *Bull. Atomic Scientists* 61 (September/October): 73–75

Schrage, Michael. 2003. “No Weapons, No Matter. We Called Saddam’s Bluff.” *Washington Post* (May 11), B2.

Sobel, Joel. 1992. “How (and When) to Communicate with Enemies.” In *Equilibrium and Dynamics*, edited by Mukul Majumdar. London: Macmillan.

Takeyh, Ray. 2005. “Prepared Statement.” In *WMD Terrorism and Proliferent States: Hearing before the Subcommittee on Prevention of Nuclear and Biological Attack of the Committee on Homeland Security,* U.S. House of Representatives. 109th Cong., 1st sess., September 8.

Townsend, Robert M. 1979. “Optimal Contracts and Competitive Markets with Costly State Verification.” *J. Econ. Theory* 21 (October): 265–93.

Woods, Kevin, James Lacey, and Williamson Murray. 2006. “Saddam’s Delusions: The View from the Inside.” *Foreign Affairs* 85 (May/June): 2–26.