Signaling Foreign Policy Interests

TYING HANDS VERSUS SINKING COSTS

JAMES D. FEARON
Department of Political Science
University of Chicago

The author distinguishes between two types of costly signals that state leaders might employ in trying to credibly communicate their foreign policy interests to other states, whether in the realm of grand strategy or crisis diplomacy. Leaders might either (a) tie hands by creating audience costs that they will suffer *ex post* if they do not follow through on their threat or commitment (i.e., costs arising from the actions of domestic political audiences) or (b) sink costs by taking actions such as mobilizing troops that are financially costly *ex ante*. Analysis of a game model depicting the essentials of each case yields two principal results. First, in the games' equilibria, leaders never bluff with either type of signal; they do not incur or create costs and then fail to respond if challenged. Second, leaders do better on average by tying hands, despite the fact that the ability to do so creates a greater *ex ante* risk of war than does the use of sunk-cost signals. These results and the logic behind them may help explain some empirical features of international signaling, such as many crises' appearance as competitions in creating domestic political audience costs. They also generate empirical puzzles, such as why the seemingly plausible logic of inference that undermines bluffing in the model does not operate in all empirical cases.

When a state's leaders threaten to use military force against another state, they generally would prefer not to carry out the threat, even if they would, in fact, be willing to. This is true not only in common cases of coercive diplomacy, such as that intermittently practiced by the Western powers in Bosnia, but also for would-be conquerors. As Clausewitz ([1830] 1984, 370) observed, "The aggressor is always peace-loving; he would prefer to take over our country unopposed." It seems quite likely that the main reason for this preference not to have to resort to force is that military operations are typically expensive and risky, obviously so for the soldiers who must be coerced or otherwise convinced to fight, but also for the leaders who order war.[1]

1. Fearon (1995) develops the general implications of this point for the problem of explaining why wars occur. See Bueno de Mesquita, Siverson, and Woller (1992) and Bueno de Mesquita and Siverson (1995) for some interesting evidence on the risks run by state leaders who engage in wars. Sometimes, of course, a leadership desires to use force simply to reveal its (privately known) military capabilities and strength to others, despite the immediate costs (Fearon 1995, 400-401).

AUTHOR'S NOTE: I wish to thank Lisa Martin, James D. Morrow, Barry O'Neill, Paul Papayoanou, Robert Pahre, Arthur Stein, and conference participants for valuable comments.

JOURNAL OF CONFLICT RESOLUTION, Vol. 41 No. 1, February 1997 68-90

Combined with the fact that leaders cannot directly observe each other's willingness to resort to force, this generally known disinclination creates one of the central dilemmas in international politics. Namely, how can a leader make a threat to use force credible when the leader would, in fact, be willing to use the military? The dilemma arises because (a) the target of the threat cannot directly observe the threatener's preferences and (b) the target knows the threatener has an incentive to pretend to be "resolved," even if this is not the case. In short, the dilemma concerns the problem of how a genuinely resolved state can threaten in such as way as to persuade the target that it is not bluffing.

In foreign policy, states confront this problem in two main contexts: grand strategy and specific international disputes or crises. In grand strategy, it appears as the problem of how to convey to other states what are one's "vital interests," which are precisely those interests over which a state is willing to fight if challenged. Although this was perceived as a crucial issue during the cold war, since 1991 the bigger problem in U.S. foreign policy has been to decide whether the United States *has* any vital interests abroad in this sense, rather than how to signal what they are to potential aggressors. By contrast, the other context in which signaling willingness to use force is an important problem—crisis diplomacy—remains significant in U.S. foreign policy, perhaps more so now than in the second half of the cold war. In Iraq, Bosnia, Haiti, and Somalia, U.S. administrations have at various times wanted military threats to be believed that they would rather not have carried out.

I have argued elsewhere that the main way that states attempt to resolve this dilemma is by making their threats *costly signals*. That is, a threat may be rendered credible when the act of sending it incurs or creates some cost that the sender would be disinclined to incur or create if he or she were in fact *not* willing to carry out the threat (Fearon 1990, 1992, 1994). For a threat to increase the target's belief that the sender would be willing to fight, it must be more likely that a resolved state would make the threat than an unresolved state. Thus, to be credible, a threat must have some cost or risk attached to it that might discourage an unresolved state from making it.[2]

How do state leaders make their threats into costly signals? I have argued that a principal way that a leader generates costly signals in crises is by creating *audience costs* that would be suffered if the leader backed down or backed away from a public threat or warning issued in a crisis. Audience costs arise chiefly from the reaction of domestic political audiences interested in whether foreign policy is being successfully or unsuccessfully handled by the leadership. Audience costs, however, are not the sole source of signaling costs in international disputes. Other means include taking financially costly mobilization or arming measures, engaging in limited conflicts, or running risks that the other side will opt for a first strike (Fearon 1992, chap. 3).

This article has three limited purposes. First, I propose a distinction between two "ideal types" of costly signals that leaders might use in either international disputes or grand strategy to signal their foreign policy interests to other states. Second, I analyze

2. Under some conditions, costless signals, or "cheap talk," may reliably communicate private information, although it remains unclear whether cheap talk is important in international disputes. Fearon (1995) analyzes a cheap-talk crisis bargaining game in which costless signals are shown to have no effect on either the probability of war or any agreement reached.

and contrast the strategic implications of these two different types of signals by using a game model. Third, I consider how these implications bear on a set of empirical intuitions that analysts of international disputes have advanced.

To summarize the main arguments, I propose that in trying to communicate willingness to fight or to respond forcefully to challenges on an issue, states can send signals that either tie their hands or sink costs. Tying hands means taking an action that increases the costs of backing down if the would-be challenger actually challenges but otherwise entails no costs if no challenge materializes. Tying-hands signals typically appear as public statements of intent by state leaders to the effect that national prestige is on the line in that case of *x*, *y*, or *z*. That is, a tying-hands signal typically works by creating audience costs that the leadership would suffer due to the reaction of domestic political audiences to a perceived failure in the management of foreign policy. Examples include statements such as "This will not stand" in a crisis, alliance treaties insofar as these work by engaging a state's domestic or international reputation for observing its commitments, and small "trip-wire" forces stationed in the threatened area.[3]

By contrast, sunk-cost signals are actions that are costly for the state to take in the first place but do not affect the relative value of fighting versus acquiescing in a challenge. There are few examples of the pure case here. Building arms or mobilizing troops entails costs no matter what the outcome, but they also may affect the state's expected value for fighting versus acquiescing in a challenge (which may have something like a tying-hands effect). It is important to see, however, that two distinct mechanisms are at work, and we need to analyze them separately as ideal types to understand the strategic logic of mixed cases.[4]

I consider a model with two states, a defender (*D*) and a potential challenger (*C*), where the defender may wish to signal resolve to defend a particular foreign policy interest. The defender is privately informed of its value for the international interest and then chooses a signal *m* (a number greater than or equal to zero). The challenger observes this signal and decides whether to challenge. If challenged, the defender decides whether to use force in response. In the sunk-cost case, the signal *m* is a cost the defender pays when sending the signal. Think of it as financial costs for arms or troops stationed on foreign soil. In the tying-hands case, the signal *m* is a cost that is incurred only if the defender backs down following a challenge.

There are two principal results. First, for either type of signal, when both states are uncertain about each other's value for the interest in question, no plausible equilibria involve bluffing by the defender.[5] Thus, if the defender sends the equilibrium signal $m^* > 0$, this means that the defender will fight with certainty if challenged, whereas any lesser signal implies that the defender surely will not fight. The logic behind the result is intriguing because it suggests that attempts to signal that "we *may* fight if

3. I think of audience costs as referring chiefly to costs imposed by a leader's domestic audience, although one can extend the concept to cover foreign audiences (international reputational costs) as well.

4. Sunk-cost signals are the standard case analyzed in economic theory since Spence (1973), who showed how costly (and sunk) investments in education might be used as a signal of employee quality to potential employers. One of the principal motivations for this article is to try to understand how tying-hands signals differ theoretically from the classical sunk-cost type analyzed by Spence and many other economists.

5. That is, there are no bluffing equilibria that survive Cho and Kreps's (1987) "intuitive criterion."

challenged" are unlikely to work. The problem with "part-way" signals is that the potential challenger is apt to conclude that "if they were truly serious, they would have signaled that they would *surely* fight."

This logic and result bear in an interesting way on several facts about international crises.[6] First, quite a few analysts have observed that state leaders very rarely make clear bluffs in international disputes—they rarely say, "We will do x, y, z if you do not back off," and then completely fail to carry out x, y, z if the threat fails.[7] Second, when a state leader, especially of a democracy, wants to make it clear that he or she "means business" and will follow through on a threat, it often seems that this is not so difficult. For example, President Clinton gave many halfhearted verbal signals of willingness to intervene or use limited force in Bosnia that were tested and ignored by the Bosnian Serbs. But on several occasions, he made it clear that a particular threat to intervene was serious (often by having the military fully and ostensibly plan air strikes, etc.), and these were believed. Likewise, Clinton was able to credibly signal a willingness to intervene in Haiti by pursuing a massive military mobilization that created very significant audience costs (not paid, of course, because the Haitian generals agreed to leave). The puzzle, suggested by this analysis, is why we sometimes observe halfhearted signals when convincing ones are possible. The equilibrium logic here, which does not seem implausible a priori, is that the possibility of sending convincing costly signals will make it impossible for a state to "partially convince" by sending less costly signals.

The second principal result of the theoretical analysis is that the signaling state does strictly better with tying-hands signals, despite the fact that tying-hands signals necessarily generate a greater risk of war. Empirically, this result may help explain both why states take actions in disputes that can raise the risk of war (i.e., they create audience costs) and why they prefer this sort of signal to alternative costly signals that would convey resolve without increasing the danger of violent conflict. I believe it is a defensible empirical intuition that international crises are characterized more by public contests to generate audience costs than by spending contests in which states sink costs to signal resolve. In contrast, however, signaling resolve to defend overseas interests in grand strategy more often is pursued with sunk-cost means, such as stationing troops abroad. This might be explained by the greater obstacles leaders face in generating credible audience costs far in advance of any challenge to the interest in question; leaders change over time, and alliance treaties can probably engage only so much reputational capital.[8]

A striking aspect of this analysis of foreign policy signaling is how a unitary rational actor question (how can states credibly signal their foreign policy intentions despite incentives to misrepresent?) proves to require an answer with a nonunitary conception of the state. In particular, the concept of audience costs inevitably forces us to bring domestic politics into our analysis of international disputes via the interaction between

6. I would grant that these are stylized facts; they are things about which it is difficult to collect systematic data, but a sense of them can be gained by reading about a range of cases. I believe that most analysts of crisis bargaining would concur with these empirical intuitions.

7. For an example, see Brodie (1959, 272), whose observations are quoted in the conclusion.

8. In other words, it may be harder to generate audience costs for purposes of general deterrence than for immediate deterrence.

principals (the domestic political audience) and agents (the political leadership), who perform on behalf and in front of the principals. This analysis suggests that different political structures linking audiences and leaderships can have important implications for the "high politics" of foreign policy making, and that it is difficult to understand the politics of foreign policy signaling at all without bringing in a domestic audience interested in foreign policy.[9] Future work on state signaling should model or conceptualize the links between leaders and domestic audiences more explicitly than I do in the simple game analyzed here.

In the main section of this article, I describe the game model and then analyze the sunk-cost and tying-hands cases in turn. A final section considers some empirical implications and limitations of the present model.

THEORETICAL ANALYSIS

For both types of signals, the sequence of actions is the same. First, nature informs both defender and challenger of their values for the issue in question, v_D and v_C, where v_D is the defender's value drawn from the cumulative distribution $F_D(v)$ on the positive reals and v_C is the challenger's value drawn from $F_C(v)$ on the real numbers.[10] Second, the defender chooses a signal $m \geq 0$, which is observed by the challenger. Third, the challenger chooses whether to challenge; if the state does not challenge, the game ends. Fourth, if challenged, the defender decides whether to fight.

Payoffs differ according to whether we are considering sunk-cost or tying-hands signals (see Figures 1 and 2, which depict complete information versions of both cases). In the sunk-cost case, if the challenger does not challenge, payoffs are $(v_D - m, 0)$. Thus the challenger's payoff for the status quo is normalized to be zero, whereas the defender receives its value for the status quo on the issue (v_D) less the sunk costs of the signal m. If the challenger challenges and the defender does not fight, payoffs are $(-m, v_C)$, indicating that the defender gets its value for having lost the interest in question (0) in addition to having lost the sunk costs of the signal. Finally, if the defender chooses to fight, I assume a conflict occurs in which the defender wins with probability $p \in (0, 1)$. Winning implies prevailing on this issue (e.g., controlling the territory). Thus expected payoffs for the "conflict outcome" are $(pv_D + (1-p)0 - c_D - m, (1-p)v_C + p0 - c_C)$ or $(pv_D - c_D - m, (1-p)v_C - c_C)$, where c_i is state i's costs for war relative to the possible benefits $(i = C, D)$. Notice that because signaling costs are sunk for the defender, they appear in the defender's value for the war outcome.

Because in the sunk-cost case I intend the signal m to represent some costly military investment (such as stationing NATO troops in Ukraine), it would be more realistic to have the probability that the defender wins in a conflict depend on m (thus $p(m)$ with $p'(m) \geq 0$). But this may introduce an element of tying hands because larger military

9. See also Fearon (1994) on this point as well as the other contributions to this issue.

10. I will need to assume that F_D and F_C have continuous and strictly positive density functions either on R^+ and R or on compact subsets of R^+ and R that include 0 and $c_D/p + \varepsilon$ and $c_C/(1-p) + \varepsilon$, respectively. Substantively, this will mean that there is positive *ex ante* probability that the defender prefers to fight rather than cede the issue at stake and that the challenger prefers fighting to the status quo.

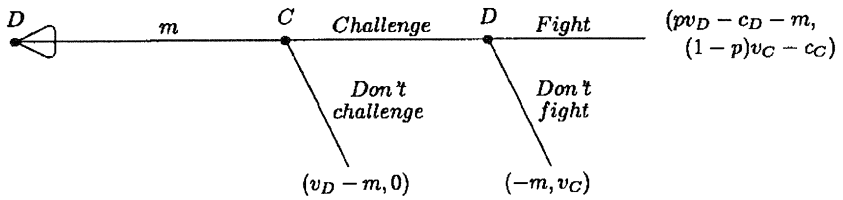

Figure 1: Sunk-Cost Signals

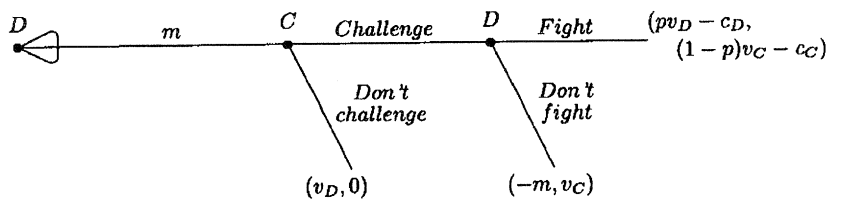

Figure 2: Tying-Hands Signals

investments make the option of fighting relatively more attractive compared to not fighting. As argued earlier, it makes sense first to consider the two ideal types to understand their different strategic logics.[11]

In the tying-hands case, payoffs are as follows. If the challenger does not challenge, the defender gets its value for the prize, v_D, and the challenger gets its value for the status quo, 0. If C challenges and the defender does not respond, the defender pays the price of the signal, m, and the challenger gets its value for prevailing on the issue, so payoffs are $(-m, v_C)$. If conflict occurs, payoffs are $(pv_D - c_D, (1-p)v_C - c_C)$. Note that in this case, the audience costs m are paid by the defender only if the defender backs down or backs away from a challenge.

ANALYSIS OF SUNK-COST SIGNALING

Complete information about the value of the prize. It is useful to begin by sketching what happens under complete information and then to examine the case where the defender has private information about v_D *but* v_C is common knowledge. With complete information, there are two major cases: either the defender's threat to fight if chal-

11. O'Neill (1989) considers a differently structured model of a nuclear arms race where spending on nuclear weapons is a costly signal of resolve and where the weapons bought are assumed to have no military value.

lenged is credible, or it is not. Because signaling costs are sunk here, the defender's decision whether to fight depends entirely on the military balance (p), its value for the "prize" in question (v_D), and its costs for fighting (c_D). If the threat is not credible ($pv_D - c_D < 0$), then the challenger will challenge, provided that $v_C > 0$, and the defender will not incur any signaling costs initially ($m^* = 0$). If the threat is credible ($pv_D - c_D \geq 0$), then the challenger will challenge only if it is so strongly motivated that it prefers war to the status quo, that is, $(1 - p)v_C - c_C > 0$. In either event, with complete information, the defender never incurs any signaling costs; there is no private information to try to signal, so why waste the costs?

Uncertainty about the defender's value for the prize. Now suppose that the challenger's value for the prize, v_C, is known but the defender's is not.[12] In this event, there is a unique "least-cost" semiseparating equilibrium in which the defender sends a signal of either $m^* > 0$ or $m = 0$, depending on whether v_D is greater than or less than a critical value $\hat{v}_D < c_D/p$. Notice that when it comes to deciding whether to fight, types $v_D > c_D/p$ will prefer to fight (simply by subgame perfection), but types with $v_D < c_D/p$ will prefer not to fight. In the least-cost semiseparating equilibrium, types $v_D \in [\hat{v}_D, c_D/p]$ mimic the behavior of more resolved types, incurring signaling costs $m^* > 0$. The critical value $\hat{v}_D$ is chosen so that the challenger just prefers not to challenge; the risk that the defender will fight in response is just large enough to deter. If r is the probability that the defender will fight in response to a challenge, then the challenger is indifferent between challenging and not challenging when

$$0 = r((1 - p)v_C - c_C) + (1 - r)v_C, \text{ or } r^* = \frac{v_C}{pv_C + c_C}. \tag{1}$$

The critical value $\hat{v}_D$ is chosen such that the probability that the defender will fight if challenged,

$$\frac{1 - F_D(c_D/p)}{1 - F_D(\hat{v}_D)}, \tag{2}$$

equals r^*.[13] The signal m^* is then chosen to equal $\hat{v}_D$ so that types $v_D > \hat{v}_D$ prefer to incur the signaling cost and get a payoff of $v_D - m^*$, whereas "weaker" types $v_D < \hat{v}_D$ prefer not to incur signaling costs (they send $m = 0$), in which case they receive 0, their value for conceding the issue.[14]

12. Also, assume that the challenger is not undeterrable—that is, the challenger does not prefer war to the status quo ($(1 - p)v_C - c_C < 0$). If the challenger is undeterrable, then signaling is pointless.

13. This is provided that there exists $\hat{v}_D > 0$, such that this is possible. If not, then there is a pooling equilibrium in which all types of the defender send the signal $m = 0$ and the challenger does not challenge.

14. State C's strategy is to challenge if the observed signal m is less than m^* and not to challenge otherwise. The challenger's beliefs in this equilibrium (on and off the path) may be specified as follows: if $m < m^*$ is observed, believe that v_D is distributed by F_D, truncated above at $\hat{v}_D$; if $m \geq m^*$ is observed, believe that v_D is distributed by F_D, truncated below at $\hat{v}_D$.

In this case of one-sided incomplete information, some seemingly natural comparative statics emerge regarding the cost of the equilibrium signal m^*. The higher the challenger's costs for war or the lower its value for the prize, the lower the m^*, meaning that the defender invests less in signaling resolve. Increasing the defender's relative power has the same effect even more powerfully. Finally, increasing the challenger's prior belief that the defender has a high value for the prize at stake means that the defender can spend less on signaling willingness to fight (m^*). The operative intuition for these results is that the less motivated the challenger is to fight for the prize, the more the defender can afford to allow the challenger to think that the defender might not actually fight for it. Thus it is less critical for the "high-value" types of the defender to distinguish themselves sharply from the "low-value" types by choosing a very costly signal m^*, and all types can save by having a lower m^*.

Uncertainty about both states' values for the issue. Introducing even a little bit of uncertainty about the challenger's motivation turns out to dramatically alter this picture. Although in the one-sided case the defender does not need to signal that it will certainly fight to deter a challenge, under broad conditions this is not true when the defender is uncertain about the challenger's motivation to challenge; bluffing then ceases to be feasible. The intuition is something like this: if the challenger might have a range of values v_C for the issue in question, then the probability of a challenge will be lower the more the defender can convince the challenger that the defender really would fight. In fact, the defender can minimize the risk of a challenge by choosing a sunk cost so high that the defender would choose this cost only if the defender was certain to be willing to fight. But then if the defender fails to choose such a costly signal, the challenger may conclude (under conditions specified later) that the defender is not so strongly resolved after all. Thus uncertainty about the challenger has the interesting effect of forcing the defender to signal "all or nothing," because signaling anything less than total commitment leads to the inference that the defender will surely not fight.

Although there exist perfect Bayesian equilibria in which the defender may bluff, they all require the challenger to draw inferences about the defender "off the path of play" that seem highly implausible. In particular, in these equilibria, the challenger must believe if it sees an unexpectedly costly signal m that it was sent by a type of the defender that could not possibly benefit from sending such a signal. I will first specify the range of these implausible separating equilibria and then show how the Cho and Kreps (1987) refinement argument rules out all but one.

Proposition 1: Consider the game with two-sided uncertainty. Choose any $\hat{v}_D$ such that $0 < \hat{v}_D \leq c_D/p$. Let

$$\hat{v}_C = \frac{c_C}{\dfrac{1 - F_D(\hat{v}_D)}{1 - F_D(c_D/p)} - p} \tag{3}$$

and let $m^* = F_C(\hat{v}_C)\hat{v}_D + (1 - F_C(\hat{v}_C))(p\hat{v}_D - c_D)$. The following strategies and beliefs form a perfect Bayesian equilibrium in the game.[15]

Strategies. State D plays m^* if $v_D \geq \hat{v}_D$ and $m = 0$ if $v_D < \hat{v}_D$; the defender fights if $v_D \geq c_D/p$ and does not fight otherwise. State C challenges if it observes $m \neq m^*$ or if $v_C > \hat{v}_C$; the challenger does not challenge if it observes m^* and $v_C \leq \hat{v}_C$.

Beliefs. State C believes that v_D is distributed by F_D, truncated below at $\hat{v}_D$, if it sees m^* and believes that v_D is distributed by F_D, truncated above at c_D/p, if it sees $m = 0$. Off the equilibrium path, if the challenger sees an $m \notin \{0, m^*\}$, then the challenger believes that v_D is distributed by any distribution that puts zero weight on values at or above c_D/p.

Behavior in these equilibria is basically similar to the one-sided case discussed earlier. If the defender cares strongly enough about the issue in question (i.e., has large enough v_D), the state chooses to incur the sunk signaling cost $m^* > 0$. However, in contrast to the preceding case where the defender knows the challenger's value for the prize, here incurring the sunk cost does not necessarily guarantee that the challenger will not challenge. Because v_C can be greater than $c_C/(1 - p)$, there is always some chance of facing such an "undeterrable" challenger who would challenge even if resistance were certain. But, on the other hand, if the defender chooses not to incur sunk costs, then the challenger will certainly challenge, so relatively "tough" types of the defender find it worthwhile to sink the costs to deter the deterrable types of challenger. As in the one-sided case, in these equilibria the defender may bluff. That is, for some levels of (privately known) resolve, the defender will incur the sunk costs but not actually fight if the challenger challenges. (As noted later, the exception is the "no-bluffing" case of $\hat{v}_D = c_D/p$.)

These bluffing equilibria are sustained, however, by a curious pattern of inferences on the part of the challenger. A bluffing equilibrium requires that if the challenger sees unexpectedly large sunk costs ($m > m^*$), then the challenger concludes that the defender must be weak and unwilling to fight. Note, however, that if the defender values the prize at v_D, it is a strictly dominated strategy for the defender to incur sunk costs $m > v_D$. That is, no matter what the challenger's strategy is, type v_D is better off choosing $m = 0$ than $m > v_D$. Thus, if the defender would be willing to fight in response to a challenge ($v_D > c_D/p$), then intuitively the defender ought to be able to signal this with confidence that the challenger will get the message by choosing $m = c_D/p$.[16] In other words, if D chooses to sink costs equal to the value of the prize for a type that is just willing to fight, then C should realize that doing this could not possibly be in D's interest if D were in fact not willing to fight and thus that D would in fact fight.

The question then becomes whether a very resolved (high v_D) type wants to try to signal that the challenger would surely face resistance if it challenged. There is a

15. A sketch of the proof, which is straightforward, is provided in the Appendix. Both here and in proposition 2, to save space I omit the defender's beliefs if it sees a challenge. These follow immediately from the challenger's strategy and Bayes's rule and are irrelevant in any event because optimal behavior for the defender after a challenge does not depend on beliefs about v_C in this model.

16. This is Cho and Kreps's (1987) "intuitive criterion" argument applied to this game.

trade-off here for the defender because to lower the risk of a challenge, the state must choose to sink greater costs. By sinking $m' = c_D/p > m^*$, the defender convinces the challenger that the defender will certainly fight (by the argument earlier), and this implies a payoff for the defender of[17]

$$F_C(\frac{c_C}{1-p})v_D + (1 - F_C(\frac{c_C}{1-p}))(pv_D - c_D) - c_D/p, \tag{4}$$

whereas by sticking with the proposed equilibrium level m^*, the defender gets $F_C(\hat{v}_C)v_D + (1 - F_C(\hat{v}_C))(pv_D - c_D) - m^*$. It is straightforward to show that for large enough v_D, the payoff for deviating to $m' = c_D/p$ is always better. Thus if the initial distribution of the defender's value for the prize, $F_D(v)$, puts positive weight on there being types that are sufficiently resolved in this sense, then we cannot support any of the equilibria described in proposition 1 in which bluffing may occur. The only equilibrium in which very resolved types of the defender do not have an incentive to deviate to signal that they will fight for sure are those in which sinking costs of m^* means that the defender *will* fight for sure.

Thus signaling dynamics drive the defender to signal "all or nothing." The state cannot signal that it might retaliate, even if this would be optimal for it,[18] because going only part-way would lead the challenger to conclude that the defender was not willing to fight at all. Thus, if a state could have sent a fully convincing signal, then a halfhearted costly signal will not convey resolve.[19]

In this equilibrium, the level of sunk costs m^* is

$$F_C(\frac{c_C}{1-p})\frac{c_D}{p}. \tag{5}$$

Comparative statics predictions on m^* differ from those of the case with one-sided incomplete information where the challenger's value for war was known. There, increasing the challenger's costs for fighting meant that the defender could deter successfully by choosing a lower level of sunk costs. Here the reverse holds: as the challenger's costs for war increase, the defender has to incur greater signaling costs to minimize the probability of attack. The reason is that the less resolved the challenger, the greater the incentive for the defender to bluff, so establishing that the defender would fight for sure requires a more costly display (note that in the one-sided uncertainty case, the defender is not driven to establish that it will fight for sure).

17. Note that $F_C(c_C/(1-p))$ is the probability that the challenger prefers not to challenge if it expects resistance for sure and that $1 - F_C(c_C/(1-p))$ is the probability that the challenger is undeterrable, that is, prefers to fight even if resistance is certain.

18. This is in light of the trade-off between the costs of signaling and reducing the probability of a challenge.

19. The one condition here is that the challenger initially believes that the defender might be sufficiently strongly resolved (i.e., put positive weight on large enough values of v_D). I think this is generally plausible because the challenger need only put an infinitesimal likelihood on this for the result to hold. The one case where it might not hold is nuclear weapons, where it may be close to common knowledge that no state could have a value for the prize that was worth the costs of a nuclear war. If so, then the model predicts more bluffing in nuclear crises than in conventional ones.

Changes in the balance of power p have two opposite effects on the costs necessary for the defender to signal resolve to fight. On the one hand, the stronger the defender is in observable military factors (p), the greater the incentive for types that have a low value for the prize to try to bluff, and thus the larger the sunk cost necessary for "high-value" defenders to distinguish themselves. On the other hand, with greater military strength, the defender's threat to fight becomes more credible for all types, in a way, which means that m^* can fall. Which effect predominates depends on the distribution of the challenger's types, $F_C(\bullet)$, but for many distributions, the defender will be able to signal resolve for less (lower m^*) when power is roughly balanced; when power is asymmetric, it takes a bigger signal.[20]

ANALYSIS OF TYING-HANDS SIGNALS

In the preceding case, signaling costs are completely sunk and thus have no effect on the defender's actual decision between resisting and acquiescing. In this section, I consider signals that incur no immediate costs but do affect the relative value of backing down versus fighting back for the leaders of the defending state. Specifically, they make backing down worse, perhaps by increasing the audience costs the leadership would suffer for the foreign policy defeat entailed by making a stand and then backing off. Again, it is useful to begin with the complete information case illustrated in Figure 2.

Complete information about values for the interest at stake. Notice that with a tying-hands signal, the defending state's leaders are in principle able to commit themselves to fight, regardless of their value for the prize;[21] they simply set $m = c_D - pv_D$, and they are credibly committed. If the challenger prefers the status quo to war (i.e., $(1 - p)v_C - c_C < 0$), then this is always worth doing because it ensures that the challenger will not challenge. If it happens that the challenger prefers war to the status quo, then the defender will want to commit to war only if the defender prefers war to ceding the issue, but then there is no need to tie hands anyway because this is known. If the challenger prefers war to the status quo and the defender prefers cession to fighting, then the defender creates no audience costs and the challenger simply takes the prize.

This complete information analysis reveals a significant difference between tying-hands and sunk-cost actions as signals. As Schelling (1960) stressed, tying hands can be valuable simply because it rearranges the incentives a person will face in the future, and the knowledge of this can help in bargaining. Moreover, as the analysis here clarifies, this effect can work even when tying hands plays no role whatsoever in signaling private information. By contrast, sinking costs does not rearrange incentives

20. I have used a computer to plot m^* as a function of p for normally distributed v_C and v_D, which for a great range of parameters yields this result.

21. One might want to put some restrictions on how large m can be, certainly for nondemocracies (because how much self-punishment can autocrats credibly commit to?) but probably for democracies as well. See Fearon (1994) for a different sort of model that allows restrictions on the audience costs that can be generated.

in the future and so depends entirely on information transmission for any strategic effects (at least in the specific international relations context analyzed here).

Uncertainty about the defender's value for the interest. The analysis barely changes if the challenger is uncertain about the defender's value for the prize, v_D. In this case, the defending state's leaders can guarantee the challenger that they will fight by setting $m = c_D$ (which means that even type $v_D = 0$ is committed to fight), and it is worth doing this provided that the challenger is known to prefer the status quo to war. If the challenger happens to be undeterrable, there is no point in "weak" types of the defender trying to deter by signaling a willingness to fight, so there is no point in using audience costs (although there is no harm in it for "tough" types of the defender, either). So in this case, the defender will commit to fight with certainty against a deterrable challenger, and tying hands is either not attempted or irrelevant if the challenger is known to be undeterrable.

Uncertainty about both sides' values for the prize. With uncertainty on both sides, matters again become more interesting. If the defender does not know the challenger's level of motivation, then a weak defender is running a risk by tying its own hands. The gambit might work, but it might not if the challenger is aggressive, in which case the defender may regret having committed itself not to back down. For example, the Clinton administration might like to extend security guarantees to various Eastern European countries, or even Ukraine, if it knew for sure that this would deter future aggression by an irredentist regime that might come to power in Moscow. But the danger is precisely that nature might draw a highly motivated type for a future Russian regime, one that would not be bothered by Western security guarantees. This would then put a U.S. administration in the unpleasant position of choosing between fighting in Ukraine or paying the audience costs for backing away from the commitment.

Proposition 2 characterizes the (essentially unique)[22] equilibrium in this case.

Proposition 2: Let

$$m^* = \frac{F_C(\frac{c_C}{1-p})c_D}{F_C(\frac{c_C}{1-p})(1-p) + p}, \tag{6}$$

and let $\hat{v}_D = (c_D - m^*)/p$. The following strategies and beliefs form a perfect Bayesian equilibrium of the game with tying-hands signals.

Strategies. If $v_D \geq \hat{v}_D$, then the defender sends the signal m^* and chooses to fight if challenged. If $v_D < \hat{v}_D$, then the defender sends $m = 0$ and does not fight if challenged.

22. There are some degrees of freedom in choosing the signals sent by tough types of the defender; we also can have equilibria where they choose arbitrary levels $m \geq m^*$. This has no substantive effects; probability distributions on outcomes remain the same.

State C chooses to challenge if $m < m^*$ or $v_C > c_C/(1-p)$ and chooses not to challenge if $m \geq m^*$ or $v_C \leq c_C/(1-p)$.

Beliefs. State C believes that v_D is distributed by F_D, truncated below at $(c_D - m)/p$, if it sees $m \geq m^*$ and believes that v_D is distributed by F_D, truncated above at $(c_D - m)/p$, if it sees $m < m^*$.

In this equilibrium, the defender ties its hands if it has a high enough value for the prize in question, and this reliably communicates that the defender would fight for sure if challenged. There is no bluffing in the sense that the defender never ties its hands and then backs down if challenged. Moreover, bluffing cannot be supported in any equilibrium with tying-hands signals due to a similar (but even more general) logic than in the sunk-costs case.[23] With no limits on the ability to tie hands, highly resolved types of the defender can always choose a very large cost for backing down, so that the challenger knows that the defender would want to fight regardless of the defender's value for the prize. Moreover, because highly resolved types of the defender will never actually pay the costs of backing down, they maximize their payoff simply by minimizing the risk of a challenge. The fact that they can always do this by creating large audience costs constrains less resolved types' ability to bluff.

It should be stressed that this result depends crucially on the assumption that the defender's leaders could, if they wished, generate costs for backing down after a challenge that would commit them to fight, regardless of their true value for the territory or issue at stake. If this is possible, then there cannot be an equilibrium in which one state, say the United States, offers a security guarantee to another, say Ukraine, that is understood to carry a significant risk of not being fulfilled in case of need. This cannot be an equilibrium because a "tough" type of United States—one that *would* be committed to fight by a given signal $m > 0$—could never be content with a signal that conveyed less than full commitment. Such a type would do better to signal unequivocal commitment by generating audience costs large enough to commit any type, provided this is possible.

Because it is obvious that commitments such as security guarantees do not always (or perhaps ever) carry with them a certain expectation of being fulfilled, one is naturally led to ask what difference or differences between the model and world account for this or what is wrong with the argument suggested by the model. One possibility is that states may not be able in all cases to generate arbitrarily large audience costs for backing down (Fearon 1994). Putting an upper bound on m creates the possibility of equilibria in which an equilibrium signal $m^* > 0$ does not convey certain commitment. Several other possibilities are briefly discussed in the conclusion.

A second valuable result compares the value of equilibrium for the defender in the case of tying-hands signals to the defender's value in the case of sunk-cost signals.

23. It is not even necessary to invoke the intuitive criterion here, as the argument in the text that follows indicates.

Proposition 3: For all types $v_D \geq 0$, the expected payoff for the defender in the tying-hands case is at least as great as that in the sunk-costs case, and the payoff is strictly greater for all types $v_D > (c_D - m^*)/p$ (where m^* is the equilibrium signal in the tying-hands case).

Thus, from both an *ex ante* and an "interim" perspective, tying-hands signals are more valuable and lead to better results for the defender (on average) than do sunk-cost signals. The reason is that tying-hands signals have the same effect on the challenger's behavior as do sunk-cost signals—they minimize the risk of a challenge by signaling that the defender will fight for sure—but they are not in and of themselves costly. Note that with sunk-cost signals, any type of the defender that chooses to signal pays the costs upfront for doing so, whereas with tying-hands signals the audience costs created are never paid in equilibrium because no type backs down after creating them. Thus highly resolved types of the defender with $v_D \geq c_D/p$ are strictly better off with tying-hands signals because they will get either v_D or $pv_D - c_D$ with tied hands as opposed to $v_D - m^*$ or $pv_D - c_D - m^*$ with sunk costs, with the same probability distribution either way.[24] Tying-hands signaling does imply a set of types of the defender

$$v_D \in \left(\frac{c_D - m^*}{p}, \frac{c_D}{p}\right),$$

which will be locked in by the audience costs created and will regret this if the challenger does decide to challenge. However, even these types do better on average by tying hands than by sinking costs because their expected payoff prior to the challenger's response is strictly positive under tying hands, but it is zero in the sunk-cost case (because in that case they choose not to signal and thus cede the prize).

The irony is that even though they are more attractive for the defender than sunk-cost signals, tying-hands signals invariably generate a higher *ex ante* probability of war by making the defender more likely to try to deter by committing itself to fight. While in the sunk-cost case, the defender signals willingness to fight if $v_D > c_D/p$, with a tying-hands signal, the defender commits if $v_D > (c_D - m^*)/p$.[25] Thus, because tying-hands signals are cheaper for the defender, relatively less resolved types are inclined to use them, despite the fact that this generates a risk that they will wind up committed to an unwanted conflict.

This result may help explain why in crises states take actions that in effect raise the risk of war. It also explains why states might prefer costly signals that have this effect to costly signals that also convey resolve but entail a lower risk of war.

Finally, the equilibrium in the tying-hands case has some interesting comparative statics. The equilibrium levels of audience costs created (m^*) are increasing in both states' costs for fighting, c_D and c_C. Raising c_D implies that the defender must create larger costs to convince the challenger that the defender will surely resist. Raising c_C

24. The m^* in this sentence refers to the m^* for equilibrium in the sunk-cost case. Incidentally, it is straightforward to verify that m^* in the sunk-cost case is always larger than m^* in the tying-hands case.

25. The *ex ante* risk of war is $(1 - F_D[c_D/p])\,(1 - F_C[c_C/(1-p)])$ in the sunk-cost case and $(1 - F_D[(c_D - m^*)/p])\,(1 - F_C[c_C/(1-p)])$ in the tying-hands case.

lowers the probability that the challenger is undeterrable, and because this is the only thing that makes creating audience costs potentially costly, greater audience costs must be generated to convince the challenger that the defender is not bluffing. As before, the effects of changing the balance of power p are indeterminate without exactly specifying $F_C(\bullet)$. But for many distributions of v_C, the defender has to incur greater audience costs the more observable indices of power (p) favor it.[26] This suggests an empirically testable prediction: we should observe states tying their hands more forcefully in confrontations with militarily weak adversaries than with strong ones.

EMPIRICAL IMPLICATIONS AND PUZZLES

I have proposed a distinction between two kinds of costly signals that state leaders might employ to try to signal their foreign policy interests to other states, whether in the realm of grand strategy or in crisis bargaining. Leaders might either (a) tie hands by creating audience costs that would be paid *ex post* if they fail to follow through on their threat or warning or (b) sink costs by taking actions such as mobilizing troops or stationing large numbers of them abroad that are financially costly *ex ante*. Analysis of a simple model depicting the essentials in each case yielded two principal results. First, in both cases, there is no bluffing in equilibrium; signaling states do not incur or create costs and then fail to respond if challenged. Second, leaders do better on average by tying their hands, despite the fact that the ability to do so creates a greater *ex ante* risk of war than would the use of sunk-cost signals.

These results emerge from a highly stylized model that omits many aspects of grand strategy or crisis bargaining that might affect the conclusions in specific cases. In particular, the results depend crucially on the assumption that leaders are able to generate arbitrarily large audience costs and so are able to tie their hands, no matter how great the expected costs of a military conflict.[27] This assumption is surely too strong. First, regime type may condition how easily a leader can generate audience costs; dictators may find it more difficult to commit credibly to self-punishment than can leaders in democracies who will face elections and other sanctions of public opinion (Fearon 1994). And even democratic leaders may find it impossible to generate arbitrarily large audience costs; there may just be an upper bound, given preferences and other parameters. Second, it may be more possible to generate audience costs in crisis bargaining than regarding security guarantees in grand strategy because of the difficulties for a leader of projecting tied hands into an uncertain and distant future against unknown adversaries. Leaders do try to stake national honor, prestige, and

26. Again, this statement rests on an examination of cases where v_C and v_D are normally distributed.

27. A second unrealistic assumption was that sunk-cost signals have no military impact. Insofar as sunk-cost signals are most naturally interpreted as money spent building arms, mobilizing troops, and/or stationing them abroad, this is implausible; the probability of winning a conflict, p, should increase with the size of the signal m. Although I have not been able to complete the analysis of this case (it is very complicated), I do not believe the first conclusion will be affected—that is, there should still be a no-bluffing result. The same logic undermining equilibria with bluffing should operate here as well. The second result, concerning the advantages of tying hands, also should go through, although here there is a problem in how to compare welfare across one case with $p(m)$ and one with just p.

reputation on the fulfillment of alliance or security guarantees, but insofar as such audience costs would become relevant only in a changeable future rather than an immediate crisis denouement, costly (and *ex ante*) military coordination or deployments of troops may be the only feasible options. Also, because leaders and circumstances change over time, audience costs created to signal the strength of alliances and security guarantees do not attach as directly to the person of the leader generating them as do efforts to tie hands in crisis bargaining.

In light of these considerations, two sets of predictions may be drawn from the theoretical analysis:

1. In cases where leaders could generate sufficiently large audience costs to make commitment certain, we should rarely observe bluffing (i.e., less than virtually certain commitment).
2. Leaders should generally prefer tying hands to sinking costs when the former is possible, despite the fact that doing so tends to "lock in" the leader and creates greater risks of war. We should expect tying hands to be more characteristic of signaling in crises than in grand strategy, where audience costs may be harder to generate. Furthermore, we should expect that sunk-cost signals will play a more prominent role in the efforts of authoritarian leaders to signal in crises than they will for leaders in democracies.

Regarding prediction 1, we have little in the way of systematic studies of bluffing in international disputes. It is surely the case that state leaders involved in crisis bargaining rarely make explicit public threats or warnings and then completely fail to carry them out. On the other hand, leaders often make statements or take actions that have ambiguous but potentially threatening implications concerning future performance and then back down later on. According to Brodie (1959, 272),

> In diplomatic correspondence, the statement that a specified kind of conduct would be deemed an "unfriendly act" was regarded as tantamount to an ultimatum and to be taken without question as seriously intended. Bluffing, in the sense of deliberately trying to sound more determined or bellicose than one actually felt, was by no means as common a phenomenon in diplomacy as latter-day journalistic interpretations of events would have one believe. In any case, it tended to be confined to the more implicit kinds of threat.

Brodie's empirical sense that explicit bluffing is uncommon in diplomatic practice is consistent with the theoretical results given here; bluffing does not occur in equilibrium in the signaling game. However, insofar as leaders do sometimes use implicit threats and subsequently back down or back away—as when they mobilize troops in response to a challenge but fail to use them when the challenger does not back off—this can be inconsistent with the model in an interesting way. The puzzle is this: why are "partial" threats and signals not invariably subject to the logic observed in the model, whereby the possibility of signaling full commitment will undermine any attempt to signal partial commitment?

An example may be helpful. Suppose that in an attempt to signal resolve in a crisis, a leader chooses to mobilize some troops and declare that "there is a significant danger of war" or some such thing. Why shouldn't the challenger reason as follows? "They *could have* committed themselves absolutely by declaring that they absolutely would not back down and that the prestige of the state was at stake in a fundamental way.

Moreover, if they were truly willing to fight, they would certainly wish to signal this rather than leave some doubt about it. Thus they must not in fact be willing to fight." If such thoughts were anticipated, then leaders attempting to deter would be compelled to commit themselves categorically in the first place (or not to try to deter at all), so we would not observe partial or ambiguous threats and commitments—but we do.

Several explanations, each of which adds complexities not comprehended in the simple model analyzed here, seem plausible at first glance. First, tying one's own hands may have the undesired side effect of provoking the other side.[28] For example, declaring that one's own reputation is at stake may engage the challenger's reputation in a way that it had not been previously. In terms of the model, sending the signal m might create costs that the challenging state's leaders will pay if they do not follow through on their challenge.[29] If so, this might conceivably explain why states sometimes send less than fully committing signals, even when they could.

Second, leaders may be signaling to multiple audiences, both domestic and international, rather than to just the other state, and this might sometimes favor partial signals of commitment. In the case of Bosnia, Clinton's halfhearted signals of resolve sometimes seemed designed to choose a middle line between the preferences of allies and domestic supporters and opponents of greater involvement. That is, the fact that the signals would not convey resolve to the Serbs may have mattered less than the effect achieved on various other audiences.[30]

Third, whether a leader wants to carry through on a threat may depend on factors such as domestic political support for doing so, which are known imperfectly at the time a threat is issued. For example, although President Bush could attempt in the fall of 1990 to commit the United States to fight against Iraq with certainty if Hussein did not withdraw from Kuwait by the January 15 deadline, the actual decision would be conditioned in part on U.S. public opinion in the second week of January. In other words, (partially) random factors may affect a leader's value for war v_D between the time a signal of interest is sent and the time of a decision whether to fight. If this is so, then it will be impossible for leaders to commit absolutely in advance, and this fact might in turn undermine the logic that militates against partial commitments.[31]

Finally, and perhaps most significantly, leaders often may shy away from absolute commitment due to the perverse effects this can have on the incentives of the state or group receiving the commitment. Sometimes called the problem of entrapment (Snyder 1984), it also might be called the problem of moral hazard in alliances and extended deterrence.[32] Historically, it has often happened that state Ego wishes to deter Other from attacking Friend and toward this end may contemplate an alliance with Friend

28. Using a different model, O'Neill (1992) studies this problem under the rubric of "the diplomacy of insults."

29. This argument applies more to the case of crisis bargaining than to grand strategy because in the latter case, there is no presumed prior challenge by state C.

30. See Papayoanou (1997 [this issue]) for this argument.

31. Whether this conjecture holds up could be assessed by examining a variant of the model studied here, in which state D observes the mean of the distribution from which v_D will be drawn in the event that state C challenges.

32. On extended deterrence, see Huth (1988).

or public statements of willingness to intervene on Friend's behalf in a crisis. But the reason Other might consider attacking Friend is that there is some set of issues in dispute over which Friend and Other are bargaining, and by committing strongly to Friend's defense or aid, state Ego may simply encourage Friend to take an intransigent position in the bargaining with Other. In fact, Ego's commitment to Friend might even bring on what it sought to avoid—war with Other—either by making Friend refuse to make the concessions necessary to gain agreement with Other or by actually leading Friend to provoke war with Other.[33]

The problem of moral hazard arises here as follows. In principle, Ego would like to offer support to Friend conditional on Friend behaving in a moderate way in the bargaining with Other. In fact, states frequently do try to condition their support for Friend-like states in just this fashion. For example, security guarantees and defense pacts typically are not expected to be obligatory if a guaranteed party attacks rather than being attacked by Other. But such conditions are problematic. Insofar as the details of the bargaining or the circumstances of attack between Friend and Other are not directly observable by Ego, and insofar as it often can be impossible to assign blame to the one who caused negotiations to fail and war to begin, there is a problem of moral hazard between Ego and Friend. In principle, this problem might lead Ego to make partial commitments rather than absolute ones, so as to balance deterring Other against restraining Friend. For example, during the July Crisis of 1914, a member of the British government argued against making a clear commitment to support Russia against Germany on the grounds that "if both sides do not know what we shall do, both will be less willing to run risks" (Joll 1984, 20).

Partially excepting the last argument, the preceding arguments concern bluffing or partial commitments in international crises. What about bluffing with alliance commitments and security guarantees in grand strategy? Here we might have a relatively clear test available; how often do states fail to honor alliance obligations to fight with their ally when the ally is attacked? The results from the model would predict that alliance reliability in case of war should be high, provided either that (a) alliance treaties can create arbitrarily large reputational costs for noncompliance (which is surely not true) or (b) states can sink costs to signal alliance commitment by stationing troops abroad or engaging in the costly coordination of military command structures (which is probably true, although whether they can sink large enough costs is unclear).

At first glance, the quantitative literature on alliance reliability would appear to disconfirm this. A number of studies has reported that states quite frequently do not fight alongside their allies when the allies are engaged in war (e.g., Siverson and King

33. My sense is that this problem is quite common historically. For example, in the July Crisis of 1914, Lord Grey initially held back from trying to deter Germany by making a strong declaration of support for Russia mainly because he was worried that doing so might bring on war by making Russia more intransigent (Joll 1984, 20). Similar concerns were evident with both British and French leaders regarding Czechoslovakia and later Poland versus Germany in 1938 and 1939 (Taylor 1961), and these concerns also can be found in the British deliberations as to whether to ally with Japan (whose leaders were bargaining with Russia over Korea) in 1902 (Bourne 1970, 177-78). Recently, the problem arose in force in the crisis over Taiwanese presidential elections. U.S. efforts to deter Chinese military moves on Taiwan ran the risk of encouraging Taiwanese leaders to be more provocative, making an attack more likely.

1980; Sabrosky 1980). The figure of 27% reliability, from Sabrosky (1980), is frequently cited.[34]

However, this figure must be interpreted with great caution. The empirical studies all use Singer and Small's (1969) alliance data, which distinguished between defense pacts, neutrality or nonaggression pacts, and ententes. The question we would really like to answer is how often a state that has a defense pact with another state will fight in case the latter is attacked and the terms of the defense pact apply. Sabrosky's (1980) 27% figure reports something very different—the proportion of cases in which at least one member of an alliance of any type fought alongside another member in a war, regardless of whether the specific terms of the alliance applied. So this includes three sets of cases that are not relevant to determining the rate at which states honor specific alliance commitments to fight with an ally: (a) cases involving ententes and neutrality pacts, which are not commitments to fight if the "ally" is attacked;[35] (b) defense pacts where the pact was never at issue because the war in question was started by one of the allies rather than beginning as a result of an attack on one of them; and (c) cases where the specific terms of the defense pact did not apply to the war in question (e.g., the terms of the Franco-Russian alliance did not oblige France to fight with Russia against Japan in 1905, so there was no question of an alliance commitment being reneged on).

To my knowledge, only one study has examined the rate at which states honor alliance commitments when the specific terms of the agreement apply. Holsti, Hopmann, and Sullivan (1973) found that of the 48 defense pacts, neutrality agreements, and ententes in place between 1815 and 1939 whose specific *casus foederis* was invoked, 42 (88%) were honored.[36] Sabrosky's (1980) data are also suggestive of higher rates of reliability if they are disaggregated by type of alliance. Of 85 defense pacts between 1815 and 1965 with at least one member involved in a war, in 37 cases at least one other member of the pact fought alongside, and in only 7 cases did members fight on opposite sides (in the remaining 41 cases, all other members remained neutral).[37]

34. It should be noted that the question of how reliable alliances are in observed cases of wars is very different from the question of how reliable the set of all alliances are, whether challenged or not. The cases we observe where an alliance commitment is tested is a nonrandom sample, and it is natural to think that the rate of reneging will be higher for allies that are challenged than for ones that are not due to a selection effect; aggressors will be more likely to attack states that have unreliable allies. See Morrow (1994) and Smith (1995). On selection effects and extended deterrence more generally, see Fearon (forthcoming).

35. Ententes simply require consultation or cooperation in some military contingency, whereas neutrality pacts are agreements to remain neutral in the event the other party is engaged in a (possibly specific) war.

36. Sabrosky (1980, 163) was well aware of the issue in question, but for reasons that are not made clear, he believes that Holsti et al. (1973) link "the honoring of alliance commitments too closely to the contingencies specified in a formal class of alliance."

37. These numbers have to be extracted via algebra from the data Sabrosky (1980) reports, which he aggregates in two different indexes of alliance reliability; for this reason, these numbers may be off by 1 due to rounding errors. The corresponding numbers for ententes are 9 (honored), 1 (violated), and 20 (neutral); for neutrality agreements, they are 2 (honored), 13 (violated), and 47 (neutral). (*Honored* for Sabrosky means that some or all members of the alliance fought together; *violated* means that some or all fought on opposite sides.) The relatively high rate at which signatories of nonaggression pacts fight each other is interesting. The most likely explanation is that if two states sign such a pact, this means that at least one of them is worried that the other may have a reason to attack it in the near future (e.g., most of these

Regarding the second set of predictions, I can at best offer some empirical generalizations that I believe most empirical analysts of crisis bargaining would agree with (e.g., George and Smoke 1974; Lebow 1981; Snyder and Diesing 1977). First, consistent with the theoretical finding that leaders will tend to do better with tying-hands signals, international crises in fact tend to be more characterized by the creation of audience costs that lower leaders' values for backing down than by competitions in spending via arms or troops (although the latter does sometimes occur). This is particularly interesting, given that leaders themselves surely understand that tying-hands signals are more provocative and more likely to lock themselves in than would signaling based more on sunk-cost actions. It is not implausible to think that the benefits of tying-hands signals discussed in the analysis of the model are also apparent to leaders engaged in coercive diplomacy.

By contrast, signaling foreign policy interests in grand strategy tends to be marked not only by efforts to tie hands by engaging reputation in alliance treaties but also by significant sunk-cost signaling in the form of troop deployments and military coordination. The costly U.S. investment in NATO is a case in point and one where the justification for permanently stationing troops in Europe rested in part on the idea that the domestic and international audience costs created by an alliance treaty alone would not convincingly commit the United States to fight. Because we do observe defense pacts and security guarantees that do not involve significant sunk-cost signaling, such cases must be regarded as either puzzles—how can partial commitment be sustained in equilibrium if sinking costs could convey full commitment?—or cases where the audience costs created by an alliance treaty are themselves sufficient to commit the defender to fight. Deciding which (and how to resolve the puzzle, in the former case) is an interesting empirical question to be asked of specific cases.[38]

The main contribution of the theoretical work undertaken here is the identification of a logic of inference that should tend to undermine states' ability to bluff in international disputes and in grand strategy. The logic holds that attempts to partially commit to a future course of action cannot be credible if the signaler could have taken a fully committing action. Because we do observe efforts to partially commit in international relations, even when it seems that stronger signals are possible, this observation poses a puzzle that future work, both empirical and theoretical, might usefully address. As an empirical matter, just how much bluffing is there, and why do leaders sometimes partially commit if they could do otherwise?

---

cases were in the 1930s, and many involved German or Russian agreements with the Baltic states). If this is correct, then the large quantitative literature on alliances and war is making a big mistake in using Singer and Small's (1969) 1-2-3 coding of defense pacts, neutrality agreements, and ententes as an ordinal variable that measures decreasing degree of commitment or commonality of interest.

38. The question of whether authoritarian leaders are more inclined to use sunk-cost signaling because their ability to tie hands is limited is impossible to answer at present, even impressionistically, because the (cold war-driven) case study literature never had in mind a comparison of authoritarian and democratic signaling strategies. I believe, however, that leaders of democracies are both more inclined and more able to create audience costs by making public statements in crises, whereas authoritarian leaders often resort to limited engagements, a form of sunk-cost signaling. Examples would be Mao's bombardments of Quemoy and Matsu, perhaps the limited interventions that prefigured Chinese entry into the Korean War, and Stalin's risky harassment of Western aircraft in the Berlin airlift.

APPENDIX

SKETCH OF THE PROOF FOR PROPOSITION 1

Given some $\hat{v}_D \in (0, c_D/p)$ and the proposed equilibrium strategy for the defender, the challenger expects that D will fight with probability

$$\frac{1 - F_D(c_D/p)}{1 - F_D(\hat{v}_D)}$$

if it observes m^*. (By subgame perfection, the defender fights if $v_D \geq c_D/p$.) $\hat{v}_C$ is chosen so that given this probability, type $\hat{v}_C$ is indifferent between challenging and not challenging if the challenger sees m^*. Thus $\hat{v}_C$ solves

$$0 = \frac{F_D(c_D/p) - F_D(\hat{v}_D)}{1 - F_D(\hat{v}_D)} \hat{v}_C + \frac{1 - F_D(c_D/p)}{1 - F_D(\hat{v}_D)} ((1-p)\hat{v}_C - c_C). \tag{A.1}$$

It is then straightforward to show that types $v_C > \hat{v}_C$ will strictly prefer to challenge if they see m^*, but types $v_C < \hat{v}_C$ will prefer not to challenge, implying that the probability of no challenge, given m^*, is $F_C(\hat{v}_C)$.

To sustain the equilibrium, it must then be the case that given $F_C(\hat{v}_C)$, type $\hat{v}_D$ is exactly indifferent between sending the messages $m = 0$ and $m = m^*$. If this is so, then types $v_D > \hat{v}_D$ will wish to send m^*, which will give rise to exactly the right probability of being resisted to make type $\hat{v}_C$ indifferent between challenging and not challenging, which in turn implies a probability of being challenged that makes types $v_D > \hat{v}_D$ wish to signal m^*, thus creating an equilibrium. We can make (the arbitrarily chosen) type $\hat{v}_D$ indifferent by choosing m^* such that the payoff for sending $m = 0$ equals the expected payoff for sending $m = m^*$. Formally, choose m^* so that the following equality holds:

$$0 = F_C(\hat{v}_C)(\hat{v}_D - m^*) + (1 - F_C(\hat{v}_C))(p\hat{v}_D - c_D - m^*). \tag{A.2}$$

Off the equilibrium path, perfect Bayesian equilibrium imposes no restrictions for the challenger's beliefs if it sees a signal m not equal to either 0 or m^*. Thus we are free to have the challenger believe that if it sees such an m, the defender is certainly not tough (i.e., $v_D < c_D/p$), which will induce weak types of the defender to prefer $m = 0$ to any other $m > 0$.

SKETCH OF THE PROOF FOR PROPOSITION 2

In the proposed equilibrium, the defender surely fights if it sends a signal $m \geq m^*$, implying that the challenger will challenge only if $v_C > c_C/(1-p)$. Thus committing by sending m^* will imply that the challenger challenges with the probability $1 - F_C(c_C/(1-p))$, from the defender's perspective. For equilibrium, we need to choose m^* such that the types of the defender that wish to signal resolve are precisely those that would in fact fight if challenged. Following a signal m^*, the defender will prefer to fight rather than back down if $v_D > (c_D - m^*)/p$. If we choose m^* such that type $\hat{v}_D = (c_D - m^*)/p$ is exactly indifferent between sending m^* and sending $m = 0$ (which yields a payoff of 0), then the proposed equilibrium will entail optimal behavior at every

information set by all types and Bayesian updating where possible. This indifference condition is

$$0 = F_C(\frac{c_C}{1-p})\hat{v}_D + (1 - F_C(\frac{c_C}{1-p}))(p\hat{v}_D - c_D), \tag{A.3}$$

which solves to yield the expression for m^*, given in proposition 2.

Types with $v_D > \hat{v}_D$ will do strictly better to send m^*, and types with $v_D < \hat{v}_D$ gain 0 by sending $m = 0$. Because the latter would receive a negative expected utility for sending some m such that $0 < m < m^*$, they choose optimally by choosing $m = 0$, in accord with the equilibrium strategy given in the text.

REFERENCES

Bourne, K. 1970. *The foreign policy of Victorian England, 1830-1902*. London: Clarendon.

Brodie, B. 1959. *Strategy in the missile age*. Princeton, NJ: Princeton University Press.

Bueno de Mesquita, B., and R. Siverson. 1995. War and the survival of political leaders: A comparative analysis of regime types and political accountability. *American Political Science Review* 89:841-55.

Bueno de Mesquita, B., R. Siverson, and G. Woller. 1992. War and the fate of regimes: A comparative analysis. *American Political Science Review* 86:638-46.

Cho, I., and D. M. Kreps. 1987. Signaling games and stable equilibria. *Quarterly Journal of Economics* 102:179-222.

Clausewitz, C. [1830] 1984. *On war*. Reprint, Princeton, NJ: Princeton University Press.

Fearon, J. D. 1990. Deterrence and the spiral model: The role of costly signals in crisis bargaining. Paper presented at the annual meeting of the American Political Science Association, August 31-September 2, San Francisco.

———. 1992. Threats to use force: Costly signals and bargaining in international crises. Ph.D. diss., University of California, Berkeley.

———. 1994. Domestic political audiences and the escalation of international disputes. *American Political Science Review* 88:577-92.

———. 1995. Rationalist explanations for war. *International Organization* 49:379-414.

———. Forthcoming. Selection effects and deterrence. In *Deterrence debates*, edited by K. Oye. Ann Arbor: University of Michigan Press.

George, A., and R. Smoke. 1974. *Deterrence in American foreign policy*. New York: Columbia University Press.

Holsti, O. R., P. T. Hopmann, and J. D. Sullivan. 1973. *Unity and disintegration in international alliances*. New York: John Wiley.

Huth, P. 1988. *Extended deterrence and the prevention of war*. New Haven, CT: Yale University Press.

Joll, J. 1984. *The origins of the first world war*. New York: Longman.

Lebow, R. N. 1981. *Between peace and war*. Baltimore, MD: Johns Hopkins University Press.

Morrow, J. D. 1994. Alliances, credibility, and peacetime costs. *Journal of Conflict Resolution* 38:270-97.

O'Neill, B. 1989. Game-theoretic approaches to the study of deterrence and wars. In *Perspectives on deterrence*, edited by P. Stern, R. Axelrod, R. Jervis, and R. Radner, 134-56. New York: Oxford University Press.

———. 1992. The diplomacy of insults. Unpublished manuscript, Yale University.

Papayoanou, P. A. 1997. Intra-alliance bargaining and U.S. Bosnia policy. *Journal of Conflict Resolution* 41:91-116.

Sabrosky, A. N. 1980. Interstate alliances: Their reliability and the expansion of war. In *The correlates of war II: Testing some realpolitik models*, edited by J. D. Singer, 161-98. New York: Free Press.

Schelling, T. 1960. *The strategy of conflict.* Cambridge, MA: Harvard University Press.
Singer, J. D., and M. Small. 1969. Formal alliances, 1816-1965: An extension of the basic data. *Journal of Peace Research* 3(1):257-82.
Siverson, R., and J. King. 1980. Attributes of national alliance membership and war participation, 1815-1965. *American Journal of Political Science* 24:1-15.
Smith, A. 1995. Alliance formation and war. *International Studies Quarterly* 39:405-26.
Snyder, G. 1984. The security dilemma in alliance politics. *World Politics* 36:461-96.
Snyder, G., and P. Diesing. 1977. *Conflict among nations.* Princeton, NJ: Princeton University Press.
Spence, A. M. 1973. Job market signaling. *Quarterly Journal of Economics* 87:355-74.
Taylor, A.J.P. 1961. *The origins of the second world war.* London: Hamilton.