

# Honest Threats

## THE INTERACTION OF REPUTATION AND POLITICAL INSTITUTIONS IN INTERNATIONAL CRISES

ALEXANDRA GUISINGER

ALASTAIR SMITH

*Department of Political Science*

*Yale University*

---

---

Traditional arguments that link credibility to a reputation for resolve, power, or strength are contrasted with a model that posits that credibility arises from the expectation of future, continued gains from retaining an honest record. Diplomatic statements are believed only if a country's or leader's credibility is unmarred. Leaders keep their word so that they are believed in later crises. Two environments are contrasted: one in which a country's record for honesty resides within the country as a whole and another in which reputation resides with individual leaders. In this latter case, citizens have an incentive to remove leaders caught bluffing. More robust than previous reputation theories, this model also offers comparative statics for when diplomacy will be more effective—namely, when leaders are domestically accountable.

---

---

Though they carried on the mysteries of secret diplomacy, there were few real secrets in the diplomatic world, and all diplomatists were honest, according to their moral code. No ambassador said "No" when the true answer should have been "Yes." . . . Many diplomatists were ambitious, some vain or stupid, but they had something like a common aim—to preserve the peace of Europe without endangering the interest or security of their country.

—A.J.P. Taylor ([1954] 1980, xxxiii) on 19th-century diplomats

What made the diplomats of the late 1800s so carefully select their words? Were they honest by nature? If diplomats were simply selected from a more honorable breed of man, it is difficult to explain the occasions in which they did choose to deceive. Instead, we suggest that diplomats protected their reputation for honest statements to retain the benefits of credible diplomatic communication in the future. As the many pages of Taylor's (1980) history note, during this period of shifting European borders and expanding colonial lands, crises were numerous. To resort to war at each crisis would have been highly expensive. Diplomacy functioned as a way to determine whether the issue at hand was critical enough to fight for. The diplomats needed no moral code to follow; the benefits of deterring war through the use of diplomatic statements created its own value for a reputation for honesty.

War is expensive and risky. Both sides would prefer to avoid conflict while gaining the most beneficial outcome. The problem for both threatening and threatened countries during a crisis is that they lack information about each other's interests. This uncertainty leads to states being unsure as to what actions the other is prepared to undertake. Although some threatening countries value an issue so highly that they will attack regardless, others would be deterred if they knew with certainty that the defender would resist. In these latter cases, war is inefficient and avoidable.

A country considering an attack could of course ask its target how highly it values the issue at hand by first issuing a threat, but whether and how a diplomatic response is meaningful is debatable. In conveying information about its position, the target country has an incentive to overemphasize its value for the issue in the hope of deterring the aggressor from attacking.<sup>1</sup> If both types of target countries—countries that would resist and countries that would not resist if attacked—have an incentive to signal “will resist,” then signaling during a crisis becomes meaningless.

Deterrence literature has focused on “reputational effects” to explain when threats to resist are believed. These arguments rely on perceptions of the underlying characteristics of the defending states—whether they are resolved, powerful, or capable of resisting. Formal game theory models using reputation in terms of deterrence theory first arose within the field of economics to explain the behavior of monopolies and their potential rivals (Selten 1978; Kreps and Wilson 1982) but found resonance among political scientists in the ability to explain diplomacy and deterrence in the sphere of international conflict (e.g., Alt, Calvert, and Humes 1988; Nalebuff 1991; Wagner 1992; O'Neill 1989). These models have focused on reputation as a property that can be invested in and built up; countries make and follow through on threats not for the immediate gains but to achieve a reputation for a certain trait, typically labeled aggressiveness, resolve, or toughness. Countries anointed with a strong reputation are expected to encounter fewer threats and are more likely to be believed when they say they will resist than those branded as weak or irresolute. Reputation, according to Schelling (1966) in his oft-quoted *Arms and Influence*, is “one of the few things worth fighting over.”

These types of reputation arguments have been challenged on several fronts. Jervis (1984) has argued that a reputational paradox prevents reputation from creating credible signals: both weak and strong countries have an incentive to signal resolve. More recently, Mercer (1996) has questioned the assumed universal meanings of reputation by introducing psychological research demonstrating that people ascribe either situational or dispositional explanations to behavior depending on their relationship with the actor. In the realm of international relations, Mercer contends that reputations are predetermined by whether the acting country is an ally or a rival: in the case of a rival, even “weak” behavior is typically perceived as a strategic move rather than a demonstration of a lack of resolve. A reputation for resolve becomes meaningless because

1. Schelling (1960) posits a distinction between messages to be conveyed: some are information in that they provide details about the current situation or environment, whereas others, such as a threat to resist, are commitments. Because the provisions of facts need not necessarily be objective, we treat these types of messages together.

perceptions contort the signals being sent. In addition, reputation models have been criticized for requiring an interdependence of crises; the government in each crisis is expected to feel the exact influence in all past as well as all future crises, a static influence not found in international relations (Morrow 1994; Snyder 1972).

We distinguish ourselves from these earlier models and their critiques by defining reputation not in terms of a country's or an individual's behavioral traits but simply the past record of diplomatic honesty. Countries and individuals start with a clean record, and their diplomatic statements are initially believed. Actors mar their record if they renege on diplomatic statements (e.g., by failing to follow through on threats) and lose their credibility. In contrast to theories of interdependence, it is expected that the intensity of interests will vary across issues and that countries will consequently vary their degree of commitment. In terms of reputation, what is observed is a country's honesty in signaling these varying degrees of commitment rather than the levels of commitment themselves. Thus, unlike a reputation for "strength," a reputation for "honesty" cannot be developed but is instead defended.

This distinction emerges within a debate between Queen Victoria and her minister of foreign affairs, Lord Clarendon, over defining English obligations to support fellow European states during the tumultuous period of the late 1860s. Victoria, fearing England was perceived as weak by both European rivals and allies, demanded greater intervention in European affairs. She claimed that a lack of action on the part of England had encouraged its rivals in Europe to believe that "the aggressive Power may dismiss all fears of England across its path."<sup>2</sup> In contrast, Clarendon expressed less fear of being perceived as weak than of being caught bluffing: "It would seem more honest and dignified on the part of England not to menace, if she is not sure of being able to strike, and not to promise more than she may be able to perform" (qtd. in Baldelli 1998, 32).

It is precisely this latter concern for reputation—concern for having been caught lying rather than concern for not having been seen to act aggressively—that we seek to formalize. In doing so, we demonstrate why, for Lord Clarendon and so many other leaders and diplomats, it has been important not only to carefully select which commitments are made but also to follow through once a commitment is made. In our first model (the country-contingent reputation model), the country as a whole is held accountable for false diplomatic statements. Once found to renege on a commitment, its diplomatic statements are no longer credible. In the second (the agent-contingent reputation model), a country's leader is held accountable for false diplomatic statements. The country carries the leader's reputation for as long as she or he remains in power, creating an incentive to remove leaders caught being deceptive in their diplomatic claims.

Both models provide a more robust explanation for the credibility of diplomatic statements than previous reputational arguments—an explanation that neither requires interdependence between crises nor depends on behavioral characteristics of the state such as resolve. Furthermore, the comparison of the two models provides comparative

2. Queen Victoria to Lord Clarendon, April 16, 1869, as quoted in Baldelli (1998).

statics on when states will be most credible. We show that diplomatic communications have greater credibility and are effective under a broader range of conditions when leaders are domestically accountable.

The format of this article is as follows. First, we distinguish our model from earlier models of credibility, particularly those of Fearon (1997) and Sartori (1998, forthcoming). Second, we analyze a conflict undertaken without communication to form a base from which to compare both the country-contingent reputation (CCR) and agent-contingent reputation (ACR) models. Finally, we discuss expected behavior resulting from the agent-specific model.

### CREDIBILITY AND SIGNALING

Throughout the course of history, countries have invested much time and energy into diplomacy, suggesting that signals between countries do have value. In modeling these signals, we follow in the tradition of formal models proposing mechanisms for making signals credible (Powell 1990; Morrow 1989; Wagner 1989) and use Fearon's (1994) concept of domestic audience costs to expand on recent work by Sartori (1998, forthcoming).

As noted previously, credibility is important because it allows countries to avoid "inefficient" wars or those that would not be undertaken if the aggressor knew for certain that the target would resist. In peacetime diplomacy, countries are generally assumed to share a common interest that allows them to share values or make the concessions necessary to produce a positive and peaceful outcome (Snyder 1972; Crawford and Sobel 1982; Farrell and Gibbons 1989; Austen-Smith 1992). Variations in the prisoner's dilemma game have shown the benefits of cooperation even in a state of anarchy (Taylor 1976; Axelrod 1984; Oye 1988). However, in the midst of a crisis, countries are assumed to have lost such a common interest. Why should the conditions of peacetime or war create a different incentive structure? In a 1995 paper, James Fearon revived the question of why rational leaders would not find negotiated settlements, knowing that due to the costs of war, there exists at all times a settlement that all sides would prefer to the risky outcome of war. One reason is that in the crisis period preceding conflict, both parties have an incentive to posture or claim more than their true interest. Both countries would benefit if they could trust the other's statement of intent, but to do so, the signals themselves must be costly.<sup>3</sup> Although various mechanisms have been proposed to make signals costly (mobilization, increased arms spending, limited conflicts, and alliance formation, to name a few), we look specifically at two recent arguments: Fearon's (1997) domestic audience costs and Sartori's (1998, forthcoming) national reputation model.

Fearon (1997) posits that the domestic audience serves as a lie detector for leaders making threats. Citing the historic norm that domestic audiences punish or criticize

3. Crawford and Sobel (1982) show that when preferences are antagonistic, cheap talk signals are uninformative.

leaders more for backing down after escalating a crisis than for not escalating at all, Fearon suggests that public sentiment creates costs that leaders would not risk unless their international threats were credible. Audience costs are paid only if the leader backs down after having made a threat to attack or a claim to resist. This implies that the domestic audience accepts the initial overcommitment (perhaps due to acceptance that the government holds private information) but not a retreat to the actual interest of the state. By assuming that the ability of democratic populaces to punish their leaders is in general greater than that of autocratic populaces, Fearon proposes that democracies are better able to signal commitment.<sup>4</sup>

Fearon's (1997) argument, however, lacks a rational underpinning as to why the domestic audience should punish a political leader who attempted to bluff to achieve a better deal for the country but then backed down rather than pay unwarranted costs of war (Smith 1998a, 1998b; Schultz 1998). If we accept Fearon's argument that it is always beneficial for a state to mislead about its capabilities and interests during a crisis, then why should domestic audiences with rational expectations not expect leaders to bluff? Furthermore, why would the public impose domestic audience costs that "trap" a state into an action that the populace in general does not support? We propose that the domestic audience punishes the leader for destroying the country's honest record and thus for putting in jeopardy the future benefits of being able to communicate during a crisis.

The benefit of a reputation for honesty or integrity arose early in the discussion of costly signals. Snyder (1972) contended that threats become credible when they put at stake a nation's reputation and thus its future bargaining position. Outside of political science, honesty is a common theme in French, British, and American manuals of diplomacy in every era (Nicolson [1939] 1964; Bailey 1968; Berridge 1995; de Callières [1716] 1919; Cambon 1931). In common among these writers is the belief that honesty was not only a moral trait but also a necessary one. de Callières noted ([1716] 1919) that "a lie always leaves in its wake a drop of poison" and that to be effective in the business of diplomacy, a negotiator must have "a reputation for straight and honest dealing." Bailey (1968) created an analogy between the diplomatist and the banker: for each, a lie could bring a profitable coup, but success will be a one-time phenomenon because they will be blackballed from their respective communities. Cambon (1931) claimed that "the most persuasive method at the disposal of a government is the word of an honest man." For Nicolson ([1939] 1964), the entire realm of diplomacy rests on integrity, and thus he observed that "national honour" must be interpreted as "national honesty." According to Nicolson, as soon as countries begin to repudiate their promises, "anarchy follows."<sup>5</sup>

4. As has commonly been recognized (e.g., Gowa 1995; Goemans 1995), nondemocratic leaders potentially face much greater and/or fatal "audience costs" in the face of a coup than do democratic leaders. However, it is assumed that democracies provide a lower cost mechanism for accountability.

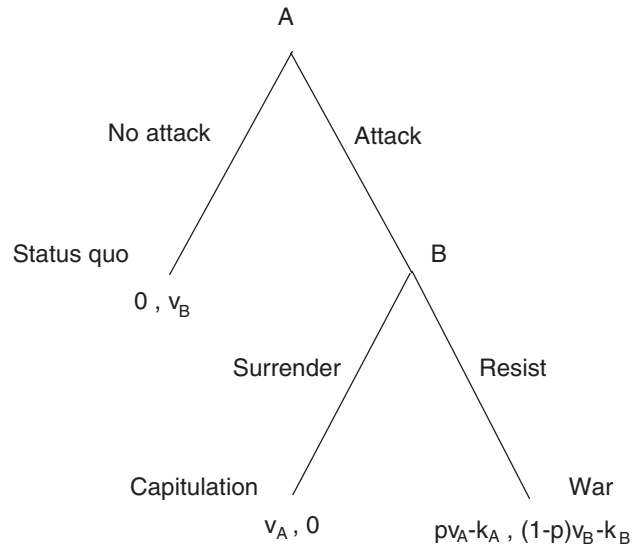
5. These statements are in contrast to the well-known quip by Sir Henry Wotton in the early 1600s: "An ambassador is an honest man sent abroad to lie for the commonwealth." However, as Nicolson ([1939] 1964) points out, not only was this statement a self-proclaimed jest, but, perhaps more importantly, King James never employed Wotton again.

Sartori (1998, forthcoming) formalizes this role of honesty. In her model, two games are played out. First the challenger threatens, and the defender has an opportunity to communicate whether the country will resist. Given a past history of honest statements (all countries start the game as credible), the challenging country will believe the statement of the defender. Sometimes, a statement to resist will effectively deter a challenger, but in other cases (where the challenger highly values the issue), the challenger will still choose to attack. If the defending country had claimed that it would resist but fails to do so at this point, it is punished for dishonesty by the international community. For a finite future period, a country is branded as dishonest, during which diplomatic statements lack credibility, and the country loses the possible deterrence benefits of credible communication in future crises. The potential benefits in future crises of being able to deter through diplomatic statements create a value for a reputation of honesty and, under certain conditions, restrain countries from bluffing.

We seek to expand beyond Sartori's (1998, forthcoming) national reputation model for several reasons. First, Sartori is forced to rely on exogenous reinstatement of credibility for nations; Sartori assumes that the length of punishment for dishonesty is exogenously determined by the members of the international community. Second, by focusing on national reputation, Sartori concludes that states do not require certain domestic conditions to create costly signals. In other words, the unitary actor model of realists need not be broken to explain the value of diplomacy. However, in focusing purely on national reputation, Sartori ignores the additional constraints and, hence, the additional credibility that certain domestic institutions may create. Drawing on recent work by McGillivray and Smith (2000), we propose an agent-contingent reputation model (ACR) to resolve the problems with both Sartori's (1998, forthcoming) and Fearon's (1997) models.<sup>6</sup> Countries and their agents initially begin with a clean record, and thus their statements are initially believed. Countries benefit from retaining this clean record by being able to communicate over a series of crises. Although political leaders might attempt to reap benefits from bluffing, if caught, they risk being labeled dishonest and thus risk being prevented from credibly communicating in future crises. If reputation resides with individuals rather than nations as a whole, citizens can remove leaders caught lying to restore the benefits of diplomatic communication. This possibility of removal creates audience costs for leaders who make threats but who subsequently back down. The greater the benefits for holding office and the easier it is to remove leaders, the greater the incentive for domestically accountable leaders to carry out the threats they make. Hence, in comparison to their more autocratic counterparts, democratic leaders are more credible. This ability to clearly signal intention has been offered as a theoretical explanation for the empirical findings of a democratic peace (Fearon 1994; Schultz 1998; Martin 1993).<sup>7</sup>

6. McGillivray and Smith (2000) examine cooperation in the context of the infinitely repeated prisoner's dilemma. They show that when politically accountable, office-seeking leaders condition future cooperation on the past good behavior of leaders, and cooperation is more robust and attainable under a far wider range of conditions than in unitary actor scenarios.

7. For other explanations, see Bremer (1992), Bueno de Mesquita et al. (1999), Dixon (1994), Lake (1992), Levy (1988), Maoz and Abdolali (1989), Maoz and Russett (1993), Ray (1995), and Rousseau et al. (1996).



**Figure 1: The Crisis Game**

Although it draws on earlier theories of credible commitments, our model distinguishes itself by

1. reconceptualizing reputation in the form of integrity, freeing it from the problems inherent in resolve-based arguments;
2. creating a microfoundational explanation for domestic audience costs;
3. explaining how domestic political institutions shape the credibility of leaders; and
4. proposing an endogenous explanation for the reinstatement of communication between governments and the nature of these communications.

### CRISIS INTERACTION

We start by introducing a model of conflict between two nations. This serves as a background against which to examine the properties of reputation. We assume a crisis exists between two nations, *A* and *B*. To keep the model as simple as possible, we look at the case where the status quo represents *B*'s ideal position. *A* can potentially alter the status quo by challenging *B*. Should it do so, country *B* must decide whether to resist. Figure 1 represents this game. The status quo prevails if *A* does not attack. The values for this outcome are 0 for country *A* and  $v_B$  for country *B*. Should *A* attack and *B* resist, war occurs. Consistent with the extant literature,<sup>8</sup> we model conflict as a simple lottery where *A* wins with probability  $p$ . Should *A* win, it alters the status quo to its favored

8. For a discussion, see Wagner (2000) and Smith (1998c).

outcome, which we normalize to a value of  $v_A$  for country  $A$  and 0 for country  $B$ . If  $B$  prevails in the war (which occurs with probability  $1 - p$ ),  $B$  retains its favored position. The payoffs associated with this policy outcome are 0 and  $v_B$  for  $A$  and  $B$ , respectively. In addition, both nations  $A$  and  $B$  pay costs associated with conflict,  $k_A$  and  $k_B$ , respectively. If  $B$  fails to resist  $A$ 's challenge, then  $A$  changes the status quo policy to its favored outcome, which is worth  $v_A$  to  $A$  and 0 to  $B$ .

Before country  $A$  initiates its attack, it is reasonable to assume that it calculates the expected value of attacking. With some probability, the country attacked,  $B$ , will acquiesce and provide  $A$  with its full value of the contested issue,  $v_A$ . With the complementary probability, war will break out, providing  $A$  with a diminished expected value dependent on the probability of victory and the cost of war,  $pv_A - k_A$ . When undertaking this calculation,  $A$  knows its own value for the issue under dispute and can gauge through observing military capabilities, alliances, and other variables its probability of victory in the case of war. However, without knowing  $B$ 's value for the issue under dispute,  $v_B$ , country  $A$  is uncertain whether its opponent will resist or acquiesce. Although country  $A$  will value some issues so highly as to always attack regardless of country  $B$ 's intentions to resist, for some portion of lesser valued issues,  $A$  attacks only because it believes, perhaps erroneously, that the probability of  $B$  resisting is low. In these latter cases, war occurs needlessly: if  $A$  knew that  $B$  would resist, then it would not attack, saving both  $A$  and  $B$  the cost of war. To formalize these arguments, we now analyze the decision-making calculus.

We start by analyzing  $B$ 's decision to resist. If attacked,  $B$ 's payoff from surrendering is 0 ( $U_B(\text{surrender}|v_B) = 0$ ). If  $B$  resists and war breaks out,  $B$ 's expected payoff is the following:

$$E[U_B(\text{resist}|v_B)] = (1 - p)v_B - k_B.$$

Hence,  $B$  resists when  $(1 - p)v_B - k_B \geq 0$ , alternatively expressed as  $v_B \geq \frac{k_B}{1-p}$ .

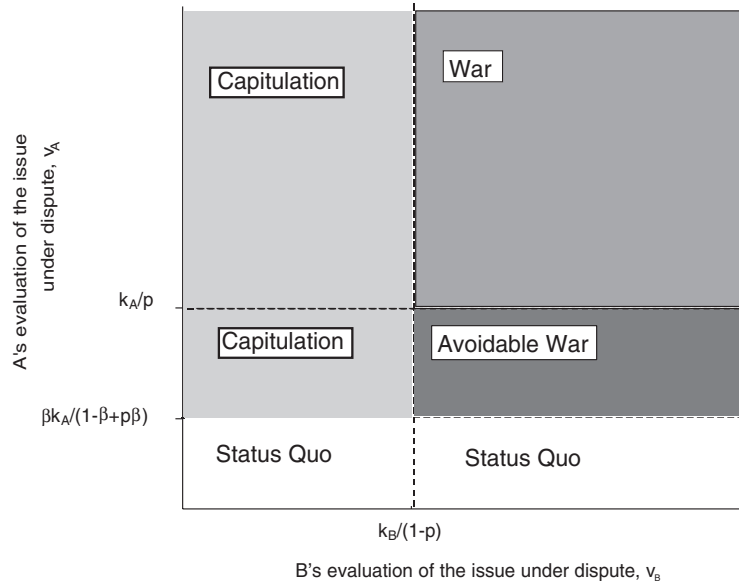
Because  $A$  does not know  $B$ 's valuation for sure, it is uncertain exactly whether  $B$  will resist. However, using its beliefs about  $v_B$ , it can estimate the probability that  $B$  resists. We assume the distribution of  $v_B$  is  $F_B(x)$ ; that is,  $\Pr(v_B \leq x)$  is  $F_B(x)$ . We similarly assume the distribution of  $A$ 's valuation,  $v_A$ , is  $F_A$ .<sup>9</sup>

Given these assumptions, the ex ante probability of  $B$  resisting,  $\beta$ , is  $\Pr(v_B \geq \underline{v}_B)$ , which equals  $1 - F_B(\underline{v}_B)$ , where  $\underline{v}_B = \frac{k_B}{1-p}$ . The special case of the uniform distribution, which we will use for all our examples, is  $\beta = 1 - F_B(\underline{v}_B) = 1 - \left(\frac{k_B}{1-p}\right)$ .<sup>10</sup>

With this estimate of the probability that  $B$  resists,  $A$  can calculate the expected value of attacking. With probability  $\beta$ , an attack leads to war for which  $A$ 's payoff is  $(pv_A - k_A)$ , and with probability  $1 - \beta$ ,  $B$  capitulates, giving  $A$  its desired policy without a fight. Formally,

9. We will let  $F_i$  have the standard nice properties of continuity, differentiability, and full support.

10. Strictly,  $\beta = 1 - F_B(\text{median}\{0, \underline{v}_B, 1\})$  rather than simply  $1 - F_B(\text{median}\{0, \underline{v}_B, 1\})$ . If  $\frac{k_B}{1-p}$  is greater than 1, then  $\underline{v}_B$  refers to the appropriate corner solution. To avoid additional notation, we shall assume an interior solution. Similarly, when specifying  $A$ 's decision to attack, we shall ignore the corner solutions in the notation.



**Figure 2: The Outcome in the Crisis Game Given the Type of Each Player**

NOTE: *B* resists only if  $v_B$  is greater than  $k_B/(1 - p)$ . *A* attacks, provided that  $v_A$  is greater than  $\beta k_A/(1 - \beta + p\beta)$ . However, if *A* were certain *B* would resist, then *A* would only attack if  $v_A$  were greater than  $k_A/p$ .

$$\begin{aligned}
 E[U_A(\text{attack}|v_A)] &= \Pr(B \text{ resists})(pv_A - k_A) + (1 - \Pr(B \text{ resists}))v_A \\
 &= \beta(pv_A - k_A) + (1 - \beta)v_A = \left(1 - F_B\left(\frac{k_B}{1-p}\right)\right)(pv_A - k_A) + F_B\left(\frac{k_B}{1-p}\right)v_A.
 \end{aligned}$$

If *A* chooses not to attack, the status quo prevails:  $U_A(\text{no\_attack}|v_A) = 0$ .

Therefore, *A* attacks if  $v_A \geq v_A = \left(1 - F_B\left(\frac{k_B}{1-p}\right)\right) \frac{k_A}{p + (1-p)F_B\left(\frac{k_B}{1-p}\right)} = \frac{\beta k_A}{1 - \beta + p\beta}$ .

These calculations tell us about the distribution of outcomes for any given crisis. Although *A*'s and *B*'s actions are ex ante optimal, they are ex post inefficient in the sense that sometimes *A* initiates an attack that leads to a war that *A* would prefer not to fight. Figure 2 depicts the strategies given the parameters  $p = .5$  and  $k_A = k_B = 0.2$ . When *A*'s valuation is small ( $v_A < \frac{\beta k_A}{1 - \beta + p\beta}$ ), *A* does not challenge the status quo. When *A* values the issue more highly, it attacks. Some of these attacks result in capitulation and some in war. Yet, if *A* knew *B* would resist, *A* would be more reluctant to attack, initiating conflict only when  $v_A \geq \frac{k_A}{p}$ .

In this numerical example, country *A* attacks 83% of the time and country *B* resists 60% of the time. Yet, if *A* were certain that *B* would resist, it would only attack if  $v_A \geq \frac{k_A}{p}$ , which occurs only 60% of the time. Therefore, if *B* could credibly communicate its intentions to resist, it could deter approximately 28% of all possible attacks

$\left(\frac{83\% - 60\%}{83\%}\right)$ . Although small in this example, the proportion of deterrable attacks rapidly increases as the cost of fighting,  $k_A$ , increases. For example, if the cost of war were to increase to .3 for both parties, 45% of all attacks could be deterred.

The revelation of  $B$ 's intentions to resist also benefits  $A$ . If they are certain that  $B$  would resist, only  $A$  types above  $0.4 \left(v_A > \frac{k_A}{p}\right)$  attack; yet when uncertain, all types above  $0.17 \left(v_A > \frac{\beta k_A}{1 - \beta + p\beta}\right)$  attack. The types falling between these two cut points benefit from knowing  $B$ 's intentions because they avoid initiating conflicts they do not wish to undertake. Hence, it is in both sides' interests to use diplomatic communications when possible.

Unfortunately,  $B$  has no incentive to remain honest. If  $A$  believes signals at face value, and therefore types between 0.17 and 0.4 are deterred, then  $B$  always benefits from a declaration of intent to fight, regardless of its true intention. Consider  $B$  types with  $v_B$  below 0.4, those who would not fight in our numerical example. An honest declaration would yield a zero payoff because  $A$ , knowing  $B$  would not resist, would always attack. These low  $B$  types would have no incentive to reveal their weakness and would instead declare a willingness to fight. If  $A$  believed such claims,  $B$  would gain from deterring  $A$  types between 0.17 and 0.4. Thus, nonresolved types would be better off imitating resolved types. Unfortunately, this would destroy the value of the signal: all types would declare a willingness to fight, and  $A$  would no longer be able discriminate between types without attacking. Attempts to deter  $A$  through statements of foreign policy would come to naught!

## REPUTATION

If in a crisis like this, we run away from these obligations of honour and interest as regards the Belgian Treaty, I doubt whether, whatever material force we might have at the end, it would be of very much value in face of the respect that we should have lost. (Sir Edward Grey to Parliament, 1914, qtd. in Robbins 1994, 179)

What is the value of a reputation? In his speech to Parliament, Sir Edward Grey supports resistance to Germany's incursion into Belgium not because of the value of an independent Belgium but because of the value of Britain's reputation. In doing so, he calculates a price for reputation that is above the expected cost of war and above the loss of life of British citizens. A reputation for honesty is valuable because it allows countries to determine each other's intent and hence avoid unnecessary wars, as discussed above. The willingness to go to war to protect a country's reputation gives weight to diplomatic statements, rendering them meaningful. As previously discussed, we define reputation as a past record of honesty in threats to challenge and resist. We assume the existence of common knowledge about previous crises and the behavior of each player.<sup>11</sup>

Does reputation reside with the country or the leader, and does it matter? We develop parallel concepts of reputation: one in which the reputation resides with the country as a whole and a second in which a country's reputation is embodied by the

political leader. Using a common framework of precrisis communication, we compare the properties of the two reputation mechanisms. First, we look at country-contingent reputation (CCR), in which we assume that the country as a whole carries the burden of a loss of reputation. In this setting, once one leader loses the nation's reputation for honesty, it is lost until the time other nations choose to forgive it. In this case, the country is functionally equivalent to a unitary actor. To make comparisons to the agent-contingent mechanism, we consider the limiting case in which nations never forgive previously observed bluffs. This so-called grim-trigger strategy creates the strongest incentive for states to remain honest because if caught bluffing, a country loses indefinitely the ability to communicate its intentions to others. Anne Sartori (1998, forthcoming) analyzes the consequences of finite punishments.

In the second case, we assume that leaders and the public are separate actors and that the reputation for honesty within crisis situations resides with the leader and not the people he or she represents. In this second scenario, which we call agent-contingent reputation (ACR), integrity is restored with the removal of a leader.<sup>12</sup> This institutional structure has two effects. First, the policy statements of domestically accountable leaders are more credible than their autocratic counterparts. Realizing that their citizens have an incentive to remove them from office if they are caught bluffing, democratic leaders do not make threats they are not prepared to carry out. Second, the agent-contingent reputation provides a mechanism through which informative communication can be restored after a leader has been caught bluffing.

#### PRECRISIS COMMUNICATION AND DOMESTIC POLITICS

To incorporate the role of reputation, we now extend the crisis game to allow for the possibility of precrisis communication and present two variations in line with the two models of domestic politics described above: a unitary actor model (CCR) and a principal-agent model (ACR). The crisis interaction forms the basis of the stage game for an infinitely repeated game. In each period, nations *A* and *B* find themselves in dispute. The structure of the interaction is identical in every period.<sup>13</sup> Yet because the issues under dispute are often different between crises, we assume that each nation's evaluation,  $v_A$  and  $v_B$ , of the issue is redrawn from the distributions  $F_A$  and  $F_B$  at the start of each period. Hence, unlike extant reputation models, nations have no underlying trait. Each country's evaluation of the issues at stake is independent of its evaluation in previous crises.

11. Milgrom, North, and Weingast (1990) express concern about assuming common information about the behavior of all players. In the context they consider (trade between a large number of individuals), these concerns are justified. Yet here we are looking at the high politics of a relatively smaller set of players (i.e., nations). In addition, the only relevant history we need to keep about players' past behavior is whether they failed to live up to a pledge in the past.

12. For simplicity, we use the term *political leader*, but this entity could vary depending on the political system. In a parliamentary system, political leadership may lie in the entire party rather than just the prime minister.

We introduce the possibility of precrisis communication by allowing  $B$  to announce whether it intends to resist should  $A$  attack. In terms of modeling,  $B$  announces either the message “resist” ( $R$ ) or the message “surrender” ( $S$ ). Obviously, in reality, there is a huge variety of messages that  $B$  could send, but these all come down to a declaration as to whether  $B$  will resist. Having observed this message, country  $A$  decides whether to attack, and if an attack occurs, then  $B$  decides whether to resist.

We introduce domestic politics by dropping the assumption that country  $B$  is a unitary actor. Instead, country  $B$  has a leader to whom we refer as leader  $B$  and an electorate. Very much as Fearon (1994) informally laid out, following the conclusion of the crisis, the electorate in country  $B$  can depose the incumbent leader and replace him or her with another leader. Leaders of political regimes differ in accountability. In many systems, such as democracies, institutions and laws provide for a relatively inexpensive way to replace leaders. In others, overturning the incumbent is much harder, often requiring civil unrest or even civil war. To model these differences, we assume that the electorate, whom we treat throughout as a unitary actor, pays a cost  $\epsilon$  to replace the incumbent. With respect to crises, all leaders and their potential replacements have the same preferences as the citizenry. Specifically, all actors in country  $B$  receive a payoff of  $v_B$  if country  $B$  prevails in the crisis. However, in addition to the value of outcomes, leader  $B$  receives a payoff of  $\Psi$  associated with the benefits of holding office.

The extended crisis interaction that makes up the stage game is as follows:

1.  $A$ 's evaluation of the issue under dispute,  $v_A > 0$ , is randomly drawn from the distribution  $F_A(v)$ .  $A$  learns this value, but the members of country  $B$  do not, knowing only the distribution  $F_A(v)$  from which it was drawn. The members of country  $B$ , the leader and electorate, simultaneously learn their valuation of the issue under dispute,  $v_B > 0$ . Again, country  $A$  knows only the distribution  $F_B(v)$  from which  $v_B$  is drawn.
2. Leader  $B$  announces either the message  $R$  (an intention to resist) or  $S$  (no intention to resist).
3. Having observed the message  $R$  or  $S$ , country  $A$  decides whether to attack.<sup>14</sup>
4. If  $A$  attacks, then leader  $B$  decides whether to resist.
5. In the agent-contingent reputation (ACR) model, the citizens observe the outcome of the crisis and decide whether to retain their incumbent leader or replace him or her at a cost of  $\epsilon$ .

13. The model might be extended to allow nations to be randomly matched in each period or allow for the possibility that the state is destroyed. Although the latter possibility might be interesting and perhaps more realistic, empirically few states completely fall. Furthermore, this does not change the fundamental structure of the game. In each period, a nation is engaged in a dispute and realizes that its actions today affect its reputation and hence its ability to influence crises tomorrow.

14. We should perhaps add some qualifiers. We restrict our attention to equilibria in which messages are fully informative. This restriction has several advantages. First, from a practical perspective, the mathematics of these equilibria is considerably simpler than that in which messages only partly reveal intentions. Second, the fully informative message equilibria represent the limiting case to partially informative messages. Although we characterize these later equilibria in the appendix, their properties are strongly related to the limiting case. This leads to the third advantage: when contrasting the country-contingent and agent-contingent reputation, restricting attention to this single class of equilibria sharpens the comparison between the two mechanisms.

The past history of the game plays an important role in the equilibria we characterize. Formally, we let the history of past play be  $h'$ . Yet for the purposes of our analysis, we are interested in only one aspect of past play: the honesty of  $B$ . Given this, we reduce the history of previous play into two categories: *Honest'* or *Cheat'*. Thus, at time  $t$ , if  $B$  has always followed through on all threats to resist, we say  $B$ 's reputation is honest, or  $h' \in \text{Honest}'$ . Alternatively, if in a past crisis  $B$  ever claimed it would resist (message  $R$ ) but failed to do so when actually attacked, then we say that  $B$  has lost its reputation for honesty or that it has cheated in the past ( $h' \in \text{Cheat}'$ ). The key distinction between the reputation mechanisms we propose is whether by  $B$ 's reputation we mean any current or previous leader of country  $B$  or the specific incumbent leader currently in office.

#### COUNTRY-CONTINGENT REPUTATION

In the days of absolute monarchy the personal honour of the King was involved in the maintenance of the contracts or treaties which had been signed and ratified in his name. Monarchs were not invariably very sensitive to this obligation, but they were at least aware (and Louis XIV was constantly aware) that their own reputation for integrity was directly and personally at stake. (Nicolson [1939] 1964, 47)

We will argue that this concern for a reputation for integrity was similarly invoked in diplomatic communications; a monarch's proclamation could serve as a powerful statement of commitment. The problem with kings and other entrenched rulers, however, is that once this reputation for integrity is lost, the country as a whole suffers the consequences for as long as the ruler remains in power.<sup>15</sup>

In the country-contingent reputation strategy (CCR), by an honest reputation we mean that country  $B$ , or its executive agent, has never announced message  $R$  and then failed to resist. Under CCR, providing  $B$  has an honest reputation,  $A$  believes  $B$  whenever its leader claims he or she is prepared to fight over an issue. In equilibrium, when the value of the issue is above some level,  $v_B^*$ , which we will define in a moment, leader  $B$  declares an intention to resist (message  $R$ ) and does so if  $A$  attacks.  $B$  types that value the issue below this level send message  $S$ , effectively conceding the issue diplomatically. These types do not resist when  $A$  attacks. Because  $B$  always follows through on its declaration to resist,  $A$  believes  $B$ 's messages and hence conditions its decision to attack on  $B$ 's diplomatic statement. Using notation parallel to that used in the uninformative scenario, we let  $\alpha(R)$  represent the probability that  $A$  attacks given that  $B$  states it will resist. Under the CCR, in equilibrium, threats to resist are credible so  $A$  attacks only when it prefers the war outcome to the status quo (i.e.,  $pv_A - k_A \geq 0$ ). Hence,  $\alpha(R) = \Pr(v_A \geq \frac{k_A}{p})$ . Alternatively, when  $B$  states it is unwilling to resist (message  $S$ ),  $A$  always attacks,  $\alpha(S) = 1$ . Once  $B$  has lost its reputation for honesty,  $A$  ignores all of  $B$ 's policy declarations, and play is equivalent to the earlier described uninformative case. Under the CCR, the electorate's decision to remove the leader is irrelevant. Given the

15. A second problem is that of incentive; as pointed out by a reviewer, individual diplomats may be tempted to bluff to further their own career prospects.

strategy of the other players, any voting strategy is optimal. Therefore, we ignore this aspect of the game and assume electorates always retain their leaders.

Given this behavior, we can calculate the value for playing the infinitely repeated game starting with an honest reputation,  $h' \in \text{Honest}'$ . We think of a country (here country  $B$ ) looking into the future, speculating about the type of disputes it is likely to encounter, and thus considering the payoffs from future crises. The expected rewards of playing, often referred to as the continuation value, are calculated as follows. For all future crises, we assume the country,  $B$ , knows only the distribution of values but from these can calculate expected payoffs. With probability  $F_B(v_B^{\frac{*}{2}})$ —the probability that in the future crisis  $v_B < v_B^{\frac{*}{2}}$ — $B$  will concede defeat diplomatically, admitting that it does not value the issue sufficiently to resist. Knowing  $B$  will not resist,  $A$  will always opportunistically advance its claims. Although under this contingency,  $B$ 's payoff in the current period is zero, it preserves its reputation for honesty. With probability  $1 - F_B(v_B^{\frac{*}{2}})$ ,  $B$  will value the issue sufficiently to declare a willingness to resist if challenged. For these types, their payoff in the current period is  $\alpha(R)((1-p)v_B - k_B) + (1 - \alpha(R))v_B$ , where the first term corresponds to the probability of  $A$  attacking multiplied by the expected payoff of conflict, and the second term is the value of the status quo,  $v_B$ , multiplied by the probability  $A$  does not attack. Given that  $B$  only declares an intention to resist when it is willing to do so,  $A$  only attacks when it prefers war to the status quo,  $\left(\alpha(R) = \Pr\left(v_A \geq \frac{k_A}{p}\right)\right)$ .

We now have the components to calculate the average value of playing this game starting with an honest reputation. At the end of the current period, provided that  $B$ 's reputation remains intact, the game looks structurally identical. Hence, the value for playing today is the immediate payoff in this period plus the discounted value of playing the games in the future.  $\delta$  represents the discount factor (i.e., the extent to which nations value future payoffs relative to payoffs today).

If  $B$  follows the CCR described above, it maintains its reputation, and hence the continuation value is given by a combination of the probability and the expected values of the three outcomes—surrender, war, and status quo:

$$W_h = F_B(v_B^{\frac{*}{2}})(0 + \delta W_h) + (1 - F_B(v_B^{\frac{*}{2}})) \left( \alpha(R)((1-p)E[v_B | v_B \geq v_B^{\frac{*}{2}}] - k_B + \delta W_h) + (1 - \alpha(R))(E[v_B | v_B \geq v_B^{\frac{*}{2}}] + \delta W_h) \right)$$

Given this recursive definition, for the special case of the uniform distribution  $W_h = \frac{1}{1-\delta} (1 - v_B^{\frac{*}{2}}) \left( \frac{1+v_B^{\frac{*}{2}}}{2} + \alpha(R) \left( -p \frac{1+v_B^{\frac{*}{2}}}{2} - k_B \right) \right)$ , where  $\alpha(R) = \left(1 - \frac{k_A}{p}\right)$ .

Alternatively, if  $B$  plays the infinitely repeated game without a reputation for honesty,  $h' \in \text{Cheat}'$ , then its foreign policy statements are ignored. Hence, in each period,  $B$  expects to receive the ex ante average value of uninformative play. Because without a reputation for honesty, behavior is identical to the initially analyzed uninformative case, the continuation value is simply the discounted expected value of all future crises:

$$W_c = \sum_{\tau=0}^{\infty} \delta^\tau E[U_B(\text{crisis})] = \frac{E[U_B(\text{crisis})]}{1-\delta},$$

which for the special case of the uniform distribution is

$$W_c = \frac{1}{1-\delta} \left( (1-\alpha) \frac{1}{2} + \alpha \beta \left( (1-p) \frac{1+v_B}{2} - k_B \right) \right),$$

$$\text{where } v_B = \frac{k_B}{1-p}, \beta = 1 - \frac{k_B}{1-p}, \alpha = 1 - \left( \frac{\beta k_A}{1-\beta + p\beta} \right).$$

Knowing the values for playing the game with an honest and a dishonest reputation, we now calculate the incentives to threaten to resist. Above we stated that only types whose valuation of the issue is greater than  $v_B^{\ddagger} = \alpha(R) \frac{k_B}{1-\alpha(R)p}$  send message  $R$ . We now show the origins of this value. If  $B$  declares itself willing to resist (and does so if attacked), then its expected payoff is  $\alpha(R)((1-p)v_B - k_B) + (1-\alpha(R))v_B + \delta W_h$ , where  $\alpha(R)((1-p)v_B - k_B)$  corresponds to the value of war multiplied by the probability that  $A$  attacks;  $(1-\alpha(R))v_B$  refers to the status quo payoff multiplied by the probability of deterring  $A$ ; and  $\delta W_h$  is  $B$ 's expected payoff from playing in the future given it maintains its honest reputation. If instead of threatening to resist,  $B$  sends message  $S$  and surrenders when attacked, its payoff is  $0 + \delta W_h$ . Type  $v_B^{\ddagger} = \alpha(R) \frac{k_B}{1-p\alpha(R)}$  is indifferent between sending these messages; types above  $v_B^{\ddagger}$  prefer to send  $R$  and resist, whereas those that value the issue less admit diplomatic defeat, sending message  $S$ .

We now assess the credibility of  $B$ 's messages. As previously shown, types that value the issue greater than  $v_B^{\ddagger} = \alpha(R) \frac{k_B}{1-p\alpha(R)}$  declare a willingness to resist. Yet, myopically,  $B$  only wants to resist if  $v_B \geq \frac{k_B}{1-p}$ . Thus, some types that send threats do not really want to carry them out. However, once a threat has been made, reputation is on the line, and its value must be entered into the equation. Provided that the benefits of maintaining an honest reputation offset the cost of fighting an undesirable conflict,  $B$  can credibly commit to resist, and the CCR strategy is equilibrium behavior.

If a previously honest  $B$  is attacked following a declaration of intent to resist (message  $R$ ), then its payoff for resisting is  $(1-p)v_B - k_B + \delta W_h$ , where  $W_h$  is the continuation value for playing the infinitely repeated game with an honest reputation. Alternatively, if this same  $B$  chooses surrender, then its payoff is  $0 + \delta W_c$ , where  $W_c$  is the continuation value from playing the game with a reputation for cheating. The difference between these two payoffs depends not only on payoffs for the current crisis but also on the (discounted) difference between playing the infinitely repeated game with an honest and dishonest reputation,  $\delta(W_h - W_c)$ . Hence, types valuing the current issue more than  $\hat{v}_B = \frac{k_B}{1-p} - \frac{\delta(W_h - W_c)}{1-p}$  carry out their threats to resist, in part, to maintain an honest reputation.

The country-contingent reputation mechanism constitutes a perfect Bayesian equilibrium in stationary strategies providing  $v_B^{\ddagger} \geq \hat{v}_B$ . This condition ensures that all types that threaten to resist (those above  $v_B^{\ddagger}$ ) actually carry out their threat (those types above  $\hat{v}_B$ ). If this condition does not hold, then  $B$  cannot credibly commit to follow through on its threats, and hence the credibility of messages degrades. To ensure  $v_B^{\ddagger} \geq \hat{v}_B$ , nations need to be sufficiently patient. For the special case of the uniform distribution

this implies,  $\delta \geq k_A \frac{k_B}{p(1-p+k_A)(W_h - W_c)}$ , which for our running numerical example ( $k_A = 0.2, k_B = 0.2, p = \frac{1}{2}$ ) reduces to  $\delta \geq .58824$ . We postpone discussion of the substantive meaning of this expression until we derive the corresponding condition for agent-contingent reputation.

We provide a formal statement of the country-contingent reputation strategy in an appendix, which is located online at [www.yale.edu/plsc506a/](http://www.yale.edu/plsc506a/). Here we focused only on the key substantive features. Just as these features explain why Sir Grey supported a war he did not want (i.e., to protect Britain's reputation), they also help explain when diplomacy will not be effective and, in doing so, may shed light on old historical arguments.

The extent to which the Russia's capitulation over the annexation of Bosnia in 1909 led to World War I has been broadly debated (see Mercer 1996; see also Huth 1988; Snyder and Diesing 1977; Geiss 1976; Joll 1984). Although many scholars have argued that Russia's demonstration of irresoluteness or weakness in 1909 led Germany and Austria to 5 years later ignore Russian threats to resist a similar incursion into Serbian territory, Mercer (1996) has criticized this reasoning, arguing in part that the Germans did not question Russians' general resolve.<sup>16</sup> As an alternative explanation to why Germany ignored Russian threats to resist, Mercer proposed that it might be that the Germans thought the Russians lacked the capacity to fight. However, this uncertainty brings into question why the Russians' diplomatic statements were no longer credible in 1914.

As suggested above, part of the answer may lie in Russia's loss of integrity. In 1909, despite depicting itself as the protector of the Serbs, Russia was in no position to prevent Austria's annexation of Bosnia. Russia's foreign minister Izvolsky clearly wished to avoid conflict: "To strain our relations with Austria (and hence with Germany too) and to risk a war on account of Bosnia and the Herzegovina would be madness" (qtd. in Mercer 1996, 114-15). Yet, seeking to extract some advantage, Izvolsky secretly negotiated a pledge from the Austrians in which Russia would initially resist but later accept the annexation in return for Austrian support of Russian control of the Bosphorus Straits. The plan backfired, however, due in part to Izvolsky's misreading of internal politics. The czar disowned the agreement, leaving Russia with the announced policy to resist. Six months later, Russia faced a Germany ultimatum: Russia could agree to Austrian proposals or risk war. Russia backed down.

What did Russia lose in this confrontation? As Izvolsky and his military leaders knew, war had never been an option. The secret agreement that included the public announcement of resistance had offered a small chance to extract some gain from the crisis. Russia had publicly backed its allies, the Serbs; in addition, according to Mercer (1996), its general resolve remained unsullied. However, by threatening to resist and then backing down, Russia lost far more than a payoff from the Austrians. The situation in 1909 led others to ignore later Russian diplomatic statements. The German undersecretary of state for foreign affairs, Alfred Zimmermann, commented that

16. Mercer (1996) quotes German Chancellor Bethmann Hollweg, who in 1912 commented that "if it were in the Russian interest to make war on us tomorrow, they would do so in cold blood."

“bluffing constitutes one of the favorite weapons of Russian policy, and while the Russian likes to threaten with the sword, yet he does not willingly draw it for the sake of others at the critical moment” (qtd. in Huth 1988, 186). Thus, Russia’s bluff in 1909 cost Russia dearly; it diminished the likelihood that Russia’s claims to resist would deter German aggression in 1914 and forced Russia to enter what would become a lengthy and expensive war to demonstrate its intent.

#### AGENT-CONTINGENT REPUTATION

The agent-contingent reputation mechanism uses a trigger mechanism similar to the country-contingent reputation mechanism except that the reputation resides with an individual leader or agent rather than with the country as a whole. New leaders start with a clean record. Thus, de facto, the national reputation can be restored by replacing any leader with a tarnished reputation. As shown above, a national reputation is valuable because it enables informative communication and thus the possible avoidance of unnecessary wars. A country carries the reputation of its leader, so citizens have an incentive to replace leaders who fail to follow through on their threats, creating a domestic audience cost.<sup>17</sup>

The ACR mechanism works in much the same manner as that of the CCR. Without an honest reputation, play under the ACR is equivalent to that of the noncommunicative single-shot game. However, given an honest reputation,  $h^t \in \text{Honest}^t$ ,  $B$  sends message  $R$  if and only if  $v_B \geq v_B^{\frac{\delta}{1-\delta}} = \alpha(R) \frac{k_B}{1-p\alpha(R)}$ , that is, when the issue under dispute is sufficiently valuable.  $A$  believes threats to resist and so only attacks when it prefers war to the status quo:  $v_A \geq \frac{k_A}{p}$ , which implies  $\alpha(R) = 1 - F_A\left(\frac{k_A}{p}\right)$ . If  $B$  threatens to resist (message  $R$ ) but is still subsequently attacked, then the leader chooses to resist, provided that  $v_B \geq \hat{v}_B = \frac{k_B}{(1-p)} - \frac{\delta v}{(1-\delta)(1-p)}$ . Provided that  $\hat{v}_B \geq v_B^{\frac{\delta}{1-\delta}}$ , all messages sent are credible. The ACR mechanism differs from the CCR in the final stage: the electorate can select to remove any leader who has damaged the country’s history of honest threats.

As before, an honest reputation is valuable because it allows  $B$  to deter aggression in crises. As before, having declared an intent to resist (message  $R$ ), leader  $B$  loses his or her reputation for honesty unless this leader backs up a threat when attacked. However, the cost of losing one’s reputation differs from the CCR. Under the CCR, a loss of reputation means that country  $B$  loses the ability to communicate in future crises. Yet, because new leaders are absolved of their predecessor’s loss of integrity, the electorate can restore the country’s reputation by replacing the incumbent with a new leader. In many cases, the electorate will want to punish a leader caught lying rather than incur the cost of not being able to communicate in the international system. The threat of such a punishment also should limit the incentive for leaders, as the electorates’ agents, to be deceitful to reap short-term career benefits. As has been previously noted, a lie in the diplomatic sphere can be seen to have the same consequence as a lie in the business

17. The threat of dismissal in the agent-contingent reputation (ACR) mechanism limits the incentives of diplomats to bluff for short-term personal career gains.

world (Bailey 1968). Although the initial temptation of a bluff might be high, a banker caught lying loses the trust of the community and is thus fired. For an incumbent caught lying, the cost is not losing the ability to communicate in future crises but is rather the loss of his or her job and associated benefits,  $\Psi$ .

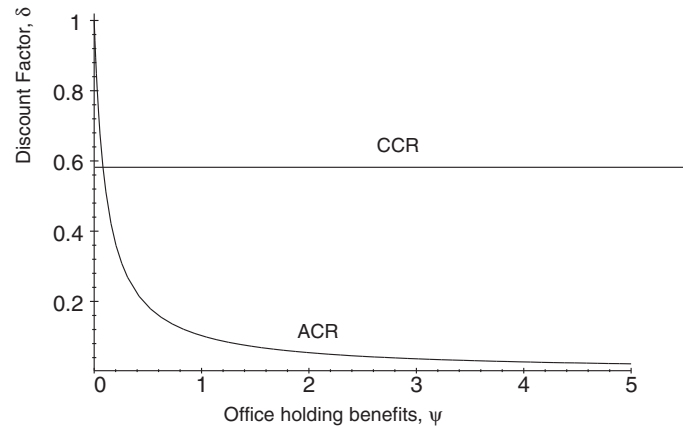
The credibility of the mechanism relies on two features. First, leaders must care sufficiently about their jobs. Second, the benefits of restoring communication in future crises must offset the cost of ousting an incumbent with tarnished integrity. More formally, for the ACR strategy to be a perfect Bayesian equilibrium, we require that  $\hat{v}_B \geq v_B^{\frac{\delta}{1-\delta}}$ , a condition that implies  $\delta \geq k_B \frac{1-\alpha(R)}{k_B(1-\alpha(R)) + \Psi(1-p\alpha(R))}$ , and  $\varepsilon \leq \delta(U_{e_B}(\text{crisis}|h^t \in \text{Honest}^t) - U_{e_B}(\text{crisis}|h^t \in \text{Cheat}^t))$ , where  $U_{e_B}(\text{crisis}|h^t \in \text{Honest}^t) = (1-\delta)W_h$  is the expected electorate's value of playing a single crisis when their leader is honest, and  $U_{e_B}(\text{crisis}|h^t \in \text{Cheat}^t) = (1-\delta)W_c$  is the corresponding payoff associated with a dishonest leader, in order.

Under ACR, the electorate removes leaders caught lying to restore the benefits of communication during crises. The threat of this audience cost makes leaders who value office holding honor their commitments. The force of this mechanism can be seen in Clinton's 1994 follow-through on his threat to use force to overthrow Haiti's military regime. Having committed to support democratic reform, Clinton found little active support among the American populace. However, Clinton followed through with the policy, risking the lives of 20,000 troops and incurring a cost of an estimated \$1.5 billion, rather than step down from his threat ("Has U.S. Wasted Effort on Haiti?" 1999). Our analysis suggests that despite these high costs of invasion, Clinton's personal value for maintaining foreign policy integrity and its consequent risk to his reelection offset these costs. If Clinton had been in his second presidential term, the invasion may have been less likely, given the lowered accountability of a nonrenewable president.

Credibility in the ACR depends on the value of office holding and the discount factor. If office holding has no value,  $\Psi = 0$ , the decision to resist collapses to the myopic case ( $B$  resisting only when  $v_B \geq \frac{k_B}{(1-p)}$ ).<sup>18</sup> However, as the value of office holding rises, more types can credibly commit to resist. Once  $v_B^{\frac{\delta}{1-\delta}} \geq \frac{k_B}{(1-p)} - \frac{\delta\Psi}{(1-\delta)(1-p)}$ , then all types who send message  $R$  subsequently resist. As the value of office holding increases, even relatively impatient leaders can credibly commit to action. In the country-contingent reputation model, for our running numerical example ( $k_A = 0.2$ ,  $k_B = 0.2$ ,  $p = \frac{1}{2}$ ), the minimum discount factor sufficient to ensure fully credible foreign policy signals was  $\delta \geq .58824$ . As demonstrated in Figure 3, the ACR supports credible policy signals at this value as long as office holding is in excess of 0.08. As the value of office holding rises further, credibility can be sustained even with much smaller discount factors, allowing credible communications under a much greater range of conditions than the CCR.

### COMPARISON OF THE MECHANISMS

Both the CCR and the ACR models suggest that we should expect to see greater efficacy in diplomacy than that predicted by traditional realist models. When countries sufficiently value the future benefits of informative communication in later crises, then



**Figure 3: The Minimum Discount Factor to Ensure Fully Credible Foreign Policy Statements under the Country-Contingent Reputation and the Agent-Contingent Reputation Strategies**

NOTE: Parameters:  $k_A = 0.2$ ,  $k_B = 0.2$ ,  $p = .5$  and  $v_A$  and  $v_B$  uniformly distributed over the unit interval.

diplomatic communication is credible. The expected behavior of countries within the CCR and ACR models is also radically different from that of previous reputation models. Reputation models based on countries creating a facade of “resolve” or “strength” suggest that due to the interdependence of commitments, countries must make and carry through commitments at low levels to avoid sending a signal of “weakness” or “irresolute.” In contrast, within both the CCR and ACR models, countries with a reputation for honesty will avoid committing themselves when they place little value on the issue under dispute. Following Lord Clarendon, countries should avoid making such commitments in fear of being forced into an expensive unwanted war or losing the ability to communicate interests in the future. In this light, Defense Secretary William Cohen announced the U.S. decision not to lead a humanitarian effort in East Timor:

We have to be selective where we commit our forces and, under the circumstances, this is not an area that we are prepared to commit forces. . . . As I have indicated before, the United States cannot be—and should not be—viewed as the policeman of the world. (“Cohen Says No U.S. Troops Planned for East Timor” 1999)

Queen Victoria and other reputation theorists would suggest that Cohen is signaling weakness, but we suggest that he is retaining America’s ability to credibly commit in future incidents.

The ACR combines features of both the traditional realist reputation arguments and the role of domestic institutions in credible commitments. By doing so, the model

18. However, in this case, we would expect to observe country-contingent reputation behavior.

integrates these two approaches, resolves the paradox's in Fearon's domestic audience cost model as to why the domestic audience would ex post punish a leader caught bluffing, and derives a series of testable implications. It is to the testable implications that we now turn.

1. Provided that leaders care about office holding, the foreign policy statements of politically accountable leaders are credible under a wider range of conditions than the declarations of their autocratic counterparts.

CCR functions adequately for small crises but unfortunately falls apart if crises are larger in magnitude. To demonstrate this comparative instability, we reconsider the interpretations of the discount factor. Above we characterized the minimum discount factor,  $\delta$ , sufficient to support credible signaling under each mechanism. We showed that if leaders value office holding, then the credibility of messages can be maintained for a wider range of conditions (i.e., a smaller discount factor) under the ACR than the CCR mechanism. The discount factor is a convenient way of assessing the relative importance of current events versus future crises. Recall that maintaining credibility requires the discounted value of an honest reputation in future crises to outweigh the short-term incentive to bluff:  $v_B^{\frac{\delta}{1-p}}(1-p) - k_B \leq \delta(W_c - W_h)$ , where  $v_B^{\frac{\delta}{1-p}} = \alpha(R) \frac{k_B}{1-p\alpha(R)}$ . One way to consider this equation is in terms of patience,  $\delta$ . Yet a more meaningful application lies in comparing the magnitude of crises. Suppose the magnitude (salience) of the current crisis is particularly large relative to most crises that nations encounter (by *large*, we mean that some multiplier is placed in front of both the payoff [ $v_B$ ] and the cost of fighting [ $k_B$ ] associated with the current crisis). Although this leaves incentives unchanged within the current crisis, it makes receiving a favorable outcome in the current crisis extremely desirable relative to success in future crises.

Both the CCR and the ACR mechanisms rely on the relative importance of today versus tomorrow to maintain credibility. As seen in the equation above, this ratio of future to current crises interacts with the discount factor. When the minimum discount factor required to maintain credibility is low, then credible communications can be supported, even when the ratio of the importance of the current crisis relative to future crises is high. Because our model shows that ACR supports credibility at much smaller discount factors, the ACR mechanism allows credible communication to continue in larger one-time crises than the CCR mechanism.

One might speculate that the stakes of a possible nuclear confrontation make it impossible for leaders to credibly commit to a policy given that the magnitude of such a crisis is likely to dwarf subsequent disputes. However, our comparative statics suggest that if any leader is able to credibly commit, it would be the democrat, not the autocrat.<sup>19</sup>

2. Domestically accountable leaders are more likely to carry out any threats they make and hence are more careful to avoid making threats they are not prepared to carry out.

All leaders weigh the costs of war against the costs of losing an honest reputation. Democratic leaders face the additional personal cost of losing the benefits of office

holding. We assume that these benefits are relatively large in magnitude. As such, the risk of being removed from office means democratic leaders are prepared to follow through on threats when the potential loss of reputation by itself would be insufficient to ensure credibility.

In 1994, President Clinton pledged U.S. involvement to restore democratic governance in Haiti, only to find general apathy among the public. However, once committed, the personal cost for Clinton of backing away from this commitment was large enough to cause him to carry through with his pledge. This illustrates an important difference from Fearon's (1994) original conception of audience costs. For Fearon, Clinton carried out his pledge purely for the selfish reason of maintaining office, raising the question of whether citizens would choose to threaten punishment given this outcome. In contrast, in the mechanism we propose, Clinton has the dual motives of staying in office and preserving national integrity, and thus the citizens benefit from following through on the threat to punish.<sup>20</sup>

Our theory suggests that accountable leaders who renege on commitments are punished domestically. Yet, we rarely observe this phenomenon in practice. This is consistent with the expectations of equilibrium behavior (Schultz 1999a). If elected leaders foresee being punished, they are unlikely to make the commitment in the first place. Of course, in reality, nations do back down. Some of these retreats are better conceptualized as arising from changing conditions rather than predetermined bluffs: commitments to resist can be inherited (such as in the Vietnam War), a country's preferences can alter (such as in U.S. intervention in Somalia), or costs of war can appear to escalate. In other cases, fluctuations in political accountability, even within democratic systems, make retreat more likely. In the United States, a second-term president might be concerned with a loss of power, prestige, or party reputation but not the specific loss of office-holding benefits. Indeed as Smith (1996, 1998a, 1998b) shows, the *ex ante* probability of reelection affects the extent to which a leader can credibly commit. If a leader is perceived to have few or no prospects of winning the next election, then he or she has little personally at stake from backing down. This undermines the credibility of his or her statements.<sup>21</sup>

Although autocratic leaders caught bluffing risk losing the benefits of communicating in future crises, their office-holding benefits are less at stake. For example, mainland China has made numerous threats to forcibly reclaim Taiwan in the event that Taiwan proclaims its independence. However, we do not expect Taiwan's recent denial of the "one China" policy to lead to government action purely due to audience costs ("Taiwan's High-Stakes Game" 1999). Although China's leaders have the advantage

19. Allison and Zelikow (1999) discuss how in 1962, President Kennedy, having declared he would not allow "offensive" weapons to be stationed on Cuba, felt compelled to take an aggressive stance during the Cuban missile crisis, believing he would be impeached otherwise. One of the many compelling reasons according to Allison and Zelikow was that failure to do so would "create public distrust of his word and his will."

20. In scanning through the *Department of State Dispatch*, it is interesting to note the difference in the stated source of the commitment between the Bosnia and Haiti ultimatums. In the case of Haiti, Clinton and his spokesmen clearly attributed the commitment to the United States and to the president himself (see, in particular, Warren 1993). In the case of Bosnia, the commitment discussed was NATO and thus any U.S. action contingent on NATO's decision to back its threats (see Clinton 1994a, 1994b).

of being able to make threats, the lack of accountability serves to remove meaning from their verbal threats. Although China has made numerous threats to resist Western intervention in Southeast Asia, including Korea, Vietnam, and Taiwan, the threats become news only when mobilization occurs.

A further implication that follows from the enhanced clarity of the democratic leaders' diplomatic statements is that countries may benefit from targeting countries that use agent-contingent rather than country-contingent reputations. Alternatively, states may seek to create conditions that support a transition to an agent-specific reputation mechanism by the target country. For example, as recommended by Smith (2000), the country issuing the demand may choose leader-specific policies such as the personalization of the crisis that provides incentives for the overthrow of the ruler or regime.

3. The arena in which diplomatic communications take place depends on domestic accountability. Domestically accountable leaders use the public forum of press conferences, international summits, and direct public addresses to signal commitment to a foreign policy. In contrast, public communiqués by autocrats are unnecessary.

During the 20th century, the nature of diplomatic communiqués has radically altered. Diplomatic histories of the 18th and 19th centuries depict the forging of agreements taking place in smoke-filled rooms between diplomatic elites and career politicians. Yet increasingly in the 20th century, foreign policy declarations were broadcast nationally or even universally. This has been particularly true for democratic countries. The different requirements of the CCR and the ACR mechanisms provide an explanation for this transition. For the country-contingent reputation, communication need only reside within the diplomatic and political elites. The ACR provides greater levels of credibility, but it does so through accountability. For policy pledges to bind, citizens must know what commitments their leaders have undertaken. Thus, diplomatic statements must be more widely communicated. In contrast, when the credibility of messages is not supported by domestic accountability, the wider distribution of information is unnecessary and, as a consequence, likely to be rare. This suggests that in closed negotiations, such as are often used for economic trade agreements, regime type is less important in signaling credibility, but threats, unless expressed publicly, are also less credible.

4. In general, the domestic accountability of democratic leaders means that they can more reliably signal their intentions, resulting in democracies being attacked less frequently, participating in fewer unnecessary wars, and benefiting from shorter negotiated settlements.

21. We can also think of opportunity costs, not simply the transaction costs of replacing a leader. Within democracies, as partisanship or policy preferences increase, the cost of removing a particular elected official increases. Although it makes little theoretical difference to the model, we could include these costs into the cost of removing the official. The more popular other policies of an elected leader are, the more expensive the leader is to remove, and the freer he or she is to make threats. This leads to the counterintuitive finding that the more popular the leader, the less credible his or her threats.

The general theme of this study is that the risk of losing office enables democratic leaders to more credibly commit to a course of action. Because their domestic political survival rests on living up to their commitments, democrats are more likely to follow through with stated foreign policy goals than are autocratic leaders. This enhanced ability to commit enables democrats to better communicate intentions and consequently deter adversaries. Scholars have proposed that these properties account for the regularities of the democratic peace. For example, Brecher and Wilkenfeld (1997; see also Dixon 1994; Mousseau 1998; Partell and Palmer 1999; Raymond 1994; Schultz 1999b) claim that the enhanced ability of democracies to reveal their intentions encourages peaceful dispute mechanisms between democracies. Fearon (1994) and Eyerman and Hart (1996) argue that this leads to fewer escalatory stages within disputes.

## CONCLUSIONS

While serving as a diplomat for England in the early 1800s, Lord Malmesbury wrote to Lord Camden (1813),

[It] is scarcely necessary to say that no occasion, no provocation, no anxiety to rebut an unjust accusation, no-ideal—however tempting—of promoting the object you have in view—can need, much less justify, a falsehood. Success obtained by one is a precarious and baseless success. Detection would ruin, not only your own reputation for ever, but deeply wound the honour of your Court. (Qtd. in Nicolson [1939] 1964, 59)

In this study, we have distinguished between agents in a way that was not yet available during Lord Malmesbury's time, when agents could be replaced but the Court still stood. Honesty, once tarnished, was hard to restore.

In 1898, negotiations with the Spanish throne over Cuban independence were brought to a standstill by the publication of a private letter from the Spanish minister involved in negotiations, Dupuy de Lome, to the Spanish military leader in Cuba. Although Dupuy's depiction of the American President McKinley as "weak" and a "would-be politician" was easily compensated for with Dupuy's forced resignation, recovering from the portions of the letter that showed that Spain was bluffing in terms of both its political and trade negotiations was less straightforward. Despite the previous U.S. preference for a negotiated settlement, without a change in the throne itself, the U.S. administration was increasingly wary of Spanish claims, and talks faltered (Offner 1992). Similarly, Madison, preceding the War of 1812, awaited news of King George III's failing health. Having been duped once by King George, Madison had hopes that the instatement of the Regent to the throne would allow a chance to negotiate the Ordinances of Council without resorting to war (Stagg 1983).

These cases illustrate how the effectiveness of diplomacy depends on past behavior. They also illustrate the inadequacies of unitary actor models of reputation. Our comparisons of CCR and ACR show that the extent of credibility varies with domestic institutions. If reputations reside with leaders as the examples above suggest, domesti-

cally accountable leaders (i.e., democrats) can more credibly commit themselves because they jeopardize their domestic political tenure should they default. This offers democratic states an additional foreign policy tool unavailable to autocratic leaders who can commit themselves only in smaller sized crises.

## REFERENCES

- Allison, Graham, and Philip Zelikow. 1999. *Essence of decision: Explaining the Cuban missile crisis*. 2d ed. New York: Longman.
- Alt, James, Randall Calvert, and Brian Humes. 1988. Reputation and hegemonic stability: A game theoretic analysis. *American Political Science Review* 82:446-65.
- Austen-Smith, D. 1992. Strategic models of talk in political decisions making. *International Political Science Review* 13 (1): 45-58.
- Axelrod, Robert. 1984. *The evolution of cooperation*. New York: Basic Books.
- Bailey, Thomas A. 1968. *The art of diplomacy: The American experience*. New York: Appleton-Century-Crofts.
- Baldelli, Pia G. Celozzi. 1998. *Power politics, diplomacy, and the avoidance of hostilities between England and the United States in the wake of the Civil War*. Lewiston, NY: Edwin Mellen.
- Berridge, G. R. 1995. *Diplomacy: Theory and practice*. New York: Prentice Hall.
- Brecher, Michael, and Jonathan Wilkenfeld. 1997. *A study of crisis*. Ann Arbor: University of Michigan Press.
- Bremer, Stuart. 1992. Dangerous dyads: Conditions affecting the likelihood of interstate war, 1816-1965. *Journal of Conflict Resolution* 26:309-41.
- Bueno de Mesquita, Bruce, James D. Morrow, Randolph M. Siverson, and Alastair Smith. 1999. An institutional explanation of the democratic peace. *American Political Science Review* 93 (4): 791-808.
- Cambon, Jules. 1931. *The diplomatist*. Translated by C. R. Turner. London: P. Allan.
- Clinton, William J. 1994a. Responding to the Sarajevo marketplace shelling: U.S. leadership and NATO resolve. *Department of State Dispatch*, 21 February.
- . 1994b. NATO's ultimatum regarding Sarajevo. *Department of State Dispatch*, 28 February.
- Cohen says no U.S. troops planned for East Timor. 1999. Associated Press, 9 September.
- Crawford, Vincent, and Joel Sobel. 1982. Strategic information transmission. *Econometrica* 50:1431-51.
- de Callières, François. [1716] 1919. *On the manner of negotiating with princes*. Translated by A. F. Whyte. Reprint, Boston: Houghton Mifflin.
- Dixon, William. 1994. Democracy and the peaceful settlement of international conflict. *American Political Science Review* 88:14-32.
- Eyerman, Joe, and Robert A. Hart, Jr. 1996. An empirical test of the audience cost proposition. *Journal of Conflict Resolution* 40:597-616.
- Farrell, Joseph, and Robert Gibbons. 1989. Cheap talk can matter in bargaining. *Journal of Economic Theory* 48:221-37.
- Fearon, James D. 1994. Domestic political audiences and the escalation of international disputes. *American Political Science Review* 88:577-92.
- . 1995. Rationalist explanation for war. *International Organization* 49:379-414.
- . 1997. Signaling foreign policy interests: Tying hands versus sinking costs. *Journal of Conflict Resolution* 41:68-90.
- Geiss, Imanuel. 1976. *German foreign policy, 1971-1914*. London: Routledge Kegan Paul.
- Goemans, Hein. 1995. The causes of war termination: Domestic politics and war aims. Ph.D. thesis, Department of Political Science, University of Chicago.
- Gowa, Joanne. 1995. Democratic states and international disputes. *International Organization* 49 (3): 511-22.
- Has U.S. wasted effort on Haiti? 1999. Reuters, 17 September.

- Huth, Paul. 1988. *Extended deterrence and the prevention of war*. New Haven, CT: Yale University Press.
- Jervis, Robert. 1984. Deterrence and perception. In *Strategy and nuclear deterrence*, edited by Steven Miller. Princeton, NJ: Princeton University Press.
- Joll, James. 1984. *The origins of the First World War*. New York: Longman.
- Kreps, David, and Robert Wilson. 1982. Reputation and imperfect information. *Journal of Economic Theory* 27:253-79.
- Lake, David A. 1992. Powerful pacifists: Democratic states and war. *American Political Science Review* 86:24-37.
- Levy, Jack. 1988. Domestic politics and war. *Journal of Interdisciplinary History* 18:653-73.
- Maoz, Zeev, and Nazrin Abdolali. 1989. Regime types and international conflict, 1816-1976. *Journal of Conflict Resolution* 33:3-36.
- Maoz, Zeev, and Bruce Russett. 1993. Normative and structural causes of the democratic peace. *American Political Science Review* 87:624-38.
- Martin, L. 1993. Credibility, costs, and institutions: Cooperation on economic sanctions. *World Politics* 45:406-32.
- McGillivray, Fiona, and Alastair Smith. 2000. Trust and cooperation through agent specific punishments. *International Organization* 54 (4): 809-24.
- Mercer, Jonathan. 1996. *Reputation and international politics*. Ithaca, NY: Cornell University Press.
- Milgrom, Paul R., Douglass C. North, and Barry R. Weingast. 1990. The role of institutions in the revival of trade: The law merchant, private judges and the champagne fairs. *Economics and Politics* 2:1-23.
- Morrow, James D. 1989. Capabilities, uncertainty and resolve. *American Journal of Political Science* 33:941-72.
- . 1994. Alliances, credibility, and peacetime costs. *Journal of Conflict Resolution* 38:270-97.
- Mousseau, Michael. 1998. Democracy and compromise in militarized interstate conflicts, 1816-1992. *Journal of Conflict Resolution* 42:210-30.
- Nalebuff, Barry. 1991. Rational deterrence in an imperfect world. *World Politics* 43:313-35.
- Nicolson, Harold (Sir). [1939] 1964. *Diplomacy*. Reprint, Oxford, UK: Oxford University Press.
- Offner, John L. 1992. *An unwanted war: The diplomacy of the United States and Spain over Cuba, 1895-1898*. Chapel Hill: University of North Carolina Press.
- 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. New York: Oxford University Press.
- Oye, Kenneth, ed. 1988. *Cooperation under anarchy*. Princeton, NJ: Princeton University Press.
- Partell, Peter J., and Glenn Palmer. 1999. Audience costs and interstate crises: An empirical assessment of Fearon's model of dispute outcomes. *International Studies Quarterly* 43 (2): 389-406.
- Powell, Robert. 1990. *Nuclear deterrence theory: The search for credibility*. New York: Cambridge University Press.
- Ray, James Lee. 1995. *Democracies in international conflict*. Columbia: University of South Carolina Press.
- Raymond, Gregory A. 1994. Democracies, disputes, and third-party intermediaries. *Journal of Conflict Resolution* 38:24-42.
- Reuters. 1999. 8 September. Cohen says no U.S. troops planned for East Timor.
- Robbins, Keith. 1994. *Politicians, diplomacy, and war in modern British history*. London: Hambledon.
- Rousseau, D. L., Christopher Gelpi, Daniel Reiter, and Paul K. Huth. 1996. Assessing the dyadic nature of the democratic peace. *American Political Science Review* 90:512-33.
- Sartori, Anne. 1998. Deterrence by diplomacy. Ph.D. thesis, Department of Political Science, University of Michigan.
- . Forthcoming. The might of the pen: A reputational theory of communication in international disputes. *International Organization*.
- Schelling, Thomas. 1960. *The strategy of conflict*. Cambridge, MA: Harvard University Press.
- . 1966. *Arms and influence*. New Haven, CT: Yale University Press.
- Schultz, Kenneth. 1998. Domestic opposition and signaling in international crises. *American Political Science Review* 92:829-44.
- . 1999a. Looking for audience costs: A research note. Working paper, Princeton University.

- . 1999b. Do democratic institutions constrain or inform? Contrasting two institutional perspectives on democracy and war. *International Organization* 53 (2): 233-66.
- Selten, Richard. 1978. The chain-store. *Theory and Decision* 9:127-59.
- Smith, Alastair. 1996. Diversionary foreign policy in democratic systems. *International Studies Quarterly* 40:133-53.
- . 1998a. International crises and domestic politics. *American Political Science Review* 92:623-38.
- . 1998b. The effect of foreign policy statements on foreign nations and domestic electorates. In *Strategic politicians, institutions, and foreign policy*, edited by Randolph M. Siverson, 221-54. Ann Arbor: University Michigan Press.
- . 1998c. Fighting battles, winning wars. *Journal of Conflict Resolution* 42 (3): 301-20.
- . 2000. Personalizing crises. In *Essays in public policy*. Stanford, CA: Hoover Institution on War, Revolution and Peace, Stanford University.
- Snyder, Glenn. 1972. Crisis bargaining. In *International crises from behavioral research*, edited by Charles F. Hermann. New York: Free Press.
- Snyder, Glenn, and Paul Diesing. 1977. *Conflict among nations: Bargaining, decision-making, and system structure in international crises*. Princeton, NJ: Princeton University Press.
- Stagg, J.C.A. 1983. *Mr. Madison's war: Politics, diplomacy, and warfare in the early American Republic, 1783-1830*. Princeton, NJ: Princeton University Press.
- Taiwan's high-stakes game. 1999. *The Economist*, 21 August.
- Taylor, A.J.P. [1954] 1980. *The struggle for mastery in Europe 1848-1918*. Oxford, UK: Oxford University Press.
- Taylor, Michael. 1976. *Anarchy and cooperation*. New York: John Wiley.
- Wagner, Harrison. 1989. Uncertainty, rational learning, and bargaining in the Cuban missile crisis. In *Models of strategic choice in politics*, edited by Peter Ordeshook, 177-205. Ann Arbor: University of Michigan Press.
- . 1992. Rationality and misperception in deterrence theory. *Journal of Theoretical Politics* 4 (2): 115-41.
- . 2000. Bargaining and war. *American Journal of Political Science* 44 (3): 469-84.
- Warren, Christopher. 1993. The Governor's Island accord: A victory for diplomacy and democracy in Haiti. *Department of State Dispatch*, 7 July.