

# Compliance, Reputation, and Domestic Politics\*

Nathaniel O. Keohane  
Yale School of Management

December 11, 2003

## Abstract

A prominent explanation for observed compliance with international agreements is that states comply in order to maintain their “reputations” for doing so, allowing them to continue to make agreements in the future. By relying on the assumption that states are unitary actors, however, reputational arguments have ignored the role played by domestic politics. In this paper I analyze a simple formal model of international agreements that highlights the interplay between domestic politics and international affairs. In my model, political parties play a key role; when in power, they decide what new agreements to enter into and whether or not to comply with prior ones. In this context, agreements become unfavorable because of shifting policy preferences: a beneficial agreement to one party is costly to the other. The model thus incorporates changing policy preferences and the possibility for government turnover.

Several conclusions arise from the model. The first involves the tension inherent in preserving a reputation: if a reputation represents an “asset” for future governments, its value depends on what those governments will do with it. A good reputation carries a cost, in the form of added flexibility given to future governments with different policy preferences. Second, the role played by each party’s electoral chances suggests a connection between electoral politics and the “depth of cooperation” that can be reached in international agreements. Third, the model points out a new role for “issue linkage,” rooted in domestic politics: by linking issues, international regimes may help ensure that policy preferences at a domestic level do not influence state compliance decisions at the international level.

## 1 Introduction

International agreements from military alliances to environmental accords embody a central contradiction: the notion that an agreement among sovereign states can be legally binding when no higher power exists to enforce contracts. The crux of the issue is compliance. Why do states comply with costly agreements? Why do other states believe them when they commit to do so?

One prominent explanation, drawing on work in the theoretical economics literature, has centered on the importance of a state’s “reputation.” According to this argument, states comply with agreements because by complying, they maintain their reputation for being “good partners,” reassuring future prospective partners and therefore retaining the option to enter into beneficial agreements in the future. Thus short-term costs are the price of long-term benefits, and even self-interested states can rationally cooperate. However, along with the arguments from economics, the

---

\*I am grateful to Suzanne Cooper, Rachel Deyette, Shigeo Hirano, Anne Joseph, Ruben Lubowski, and Jim Snyder for helpful suggestions and comments. Special thanks are due to Chris Avery and Robert Keohane.

literature has imported a crucial assumption: that states are long-lived unitary actors. In focusing on the relationships sustaining cooperation among states, scholars have largely ignored the ways in which those relationships may be undergirded (or undermined) by domestic politics.

This paper seeks to explore the interplay between domestic politics and international affairs, using a simple formal model of international agreements. In my model, political parties play a key role; when in power, they decide what new agreements to enter into and whether or not to comply with prior ones. The parties themselves act as a “transmission mechanism” linking the interests of governments over time – a crucial step in the workings of reputational arguments. In this context, agreements become unfavorable because of shifting policy preferences: a beneficial agreement to one party is costly to the other. The model thus incorporates changing policy preferences and the possibility for government turnover.

The incorporation of domestic political concerns suggests several implications for international agreements. First, the model highlights the tension inherent in preserving a reputation: to the extent that a reputation represents an “asset” in the hands of future governments, its value depends on what those governments will do with it. The paper thus explores an issue that has gone unnoticed in the literature thus far: the “cost of reputation,” in terms of the added flexibility given to future governments with different policy preferences. Second, the role played by each party’s electoral chances suggests a connection between electoral politics and the “depth of cooperation” that can be reached in international agreements. Third, the model points out a new role for “issue linkage,” rooted in domestic politics: by linking issues, international regimes may help ensure that policy preferences at a domestic level do not influence state compliance decisions at the international level.

The next section reviews various explanations of compliance in the international relations literature, focusing in particular on the reputational arguments and their antecedents in the economics literature. Section 3 introduces the model and presents key results. Section 4 draws out several implications of the model for international agreements, and section 5 concludes.

## 2 The compliance puzzle and the role of reputation

Adopting the premise of states as rational actors brings the question of compliance into sharp relief. If states behave to maximize their own self-interest, why will they willingly comply with costly agreements? In Robert Keohane’s words, “The puzzle of compliance is why governments, seeking to promote their own interests, ever comply with the rules of international regimes when they view these rules as in conflict with... their ‘myopic self-interest.’” (Keohane, 1984, p. 99). As Keohane points out, a strictly self-interested government considering only the immediate consequences of its actions might well be tempted to renege on international agreements that required costly actions, such as a military action in support of an ally, or a removal of import barriers under a trade agreement. Potential partners, anticipating the possibility of such renegeing, may then be more reluctant to sign an agreement. Thus the prospect of noncompliance erodes the ability of states to cooperate, and the *ex post* “compliance puzzle” becomes an *ex ante* commitment problem.

Compliance might be achieved by the threat of enforcement: if wronged parties were able and willing consistently to force violators to abide by their agreements, violations might never occur. This mechanism of retaliation, however, provides a weak ground to stand on. Clearly, no higher body exists in international relations with independent enforcement power – no counterpart to domestic legal courts. Thus any enforcement must come from the wronged party itself, perhaps with the support of its own allies. But enforcement itself is likely to come at considerable short-term

cost; thus the original problem of credible commitment resurfaces. Moreover, even if enforcement is credible it depends crucially on power; while it might support a credible commitment by a weaker power to a stronger power, it may not suffice to resolve the commitment problem between two powers of equal strength (and would almost certainly fail to support a commitment by a stronger power to a weaker power.) As Keohane points out, the result is a “commitment paradox”: a powerful state will find it more difficult credibly to commit itself to comply, since *ex post* a weak state will be unable to enforce such commitment (Keohane, 1997). Enforcement through retaliation, therefore, does not seem to be the sole answer to the compliance puzzle.

Another potential escape from the “compliance puzzle” is to reject the underlying premise of rationality. This is the tack taken by Abram and Antonia Chayes, who reject the characterization of states as rational egoists acting purely out of self-interest, and instead adopt the presumption that states have a “propensity to comply” (Chayes and Chayes 1995). Chayes and Chayes are right to point out the limitations of strict rational-actor analysis, although they are hardly the first to do so. However, their argument suffers from several weaknesses. First of all, to the extent that their presumption is based on empirical observation, it may be merely the result of underlying selection bias (Downs, Rocke, and Barsoom 1996). Second, they defend their presumption on the basis of “efficiency” grounds – essentially making the argument that states economize on decision-making costs by adopting the “rule of thumb” of compliance. Although this might explain routine compliance on mundane matters, it is hard to believe that such decision making costs would truly outweigh the potential advantages from, say, instituting trade barriers to protect politically powerful domestic industries.

Third, Chayes and Chayes argue that states will comply with agreements because such agreements are in their interests. “A treaty is a consensual instrument,” they write. “It is therefore a fair assumption that the parties’ interests were served by entering into the treaty in the first place” (Chayes and Chayes 1995, p. 4). Few would debate that states are unlikely to enter into treaties that run contrary to their interests from the outset. What Chayes and Chayes seem to miss, in making this argument, is that there may be a fundamental distinction between *ex ante* interests and *ex post* incentives. A treaty that looks good for a state “on net” may nonetheless become unattractive when the state is called upon to fulfill a commitment. To this crucial point – the crux of the issue – Chayes and Chayes, almost in passing, simply assert that states “do not negotiate agreements with the idea that they can break them whenever the commitment becomes ‘inconvenient’” (p. 7); but they neglect to provide evidence for this claim.

The strongest argument advanced by Chayes and Chayes is that compliance is a norm among states, which states violate at their peril. Referring to the US-Soviet ABM Treaty, the authors write: “Transgression of such a fundamental engagement would trigger not a limited response but an anxious and hostile reaction across the board, jeopardizing the possibility of cooperative relations between the parties for a long time to come” (p. 9). The reason that this argument is the strongest one, however, is somewhat contrary to the spirit of Chayes’ and Chayes’ argument: it is fundamentally consistent not only with sociological views of rule-following behavior, but also with rational-actor views of self-interested behavior. Norms can be seen as another expression of the importance of reputation, a concept rooted firmly in self-interest, to which I now turn.<sup>1</sup>

---

<sup>1</sup>Indeed, Keohane points out that reputational arguments and arguments based on social norms are fundamentally similar arguments couched in different terms and disciplines; see “Interests, Commitments, and Institutions,” 18. Going a bit further, one might argue that reputation-based arguments are consistent with “social norms” (at an interpretive level) but are also rigorously based on self-interest; compare to the complaint by Chayes and Chayes that

The idea that a state or person can increase the credibility of a threat or promise by staking their reputation on fulfillment of that commitment is not new.<sup>2</sup> The more recent tide of interest in reputation, however, has stemmed from results in the game-theoretic economics literature that provide a formal rational underpinning for reputation. Building on those insights, theorists in international relations have argued that reputation may provide a rational resolution to the “compliance puzzle.”

In *After Hegemony* (*AH*), Keohane framed the fundamental compliance problem as a Prisoner’s Dilemma, in which each party to an agreement would benefit in the short term by reneging. Anticipating such behavior by other states, no state would enter into an agreement, and thus the benefits of cooperation remain out of reach. Drawing on results from game theory, Keohane argued that repeated interaction could overcome this problem: in repeated games, players who are sufficiently patient can sustain cooperation if reneging brings punishment or a reversion to the no-agreement outcome. International institutions, he argued, facilitate cooperation by linking issues and thus creating the repeated interaction that might support cooperation. States find compliance to be in their own self interest, despite short-term costs, because their behavior in one sphere (*e.g.*, violation of one agreement) has repercussions in other spheres. In other words, issue linkage give states an incentive to acquire and maintain reputations for compliance, enhancing their chances of entering into agreements with other states.

This insight has been extended by several other scholars of international relations. Charles Lipson (1991), for example, has argued that the chief distinction between formal and informal international agreements is the degree to which a state’s reputation is at stake; formal agreements acquire more force by implying higher reputational costs if they are broken. “Put simply, treaties are a conventional way of raising the credibility of promises by staking national reputation on adherence” (Lipson 1991, p. 511). A similar point is made by Lisa Martin (1993), who emphasizes the role of “audience costs” in raising the credibility of commitments.<sup>3</sup> Martin argues that states can take actions on both domestic and international levels to increase their own costs of reneging on commitments, thereby enhancing the credibility of those commitments. Having taken such actions, a state’s “reputation within the institution would be damaged by backing down” (p. 418). Like Keohane, Martin ties the issue of reputation to issue linkage: “Once the leading sender has made a public institutional commitment... reversing this policy would tend to decrease the level of benefit derived from other dimensions of the institution” (p. 418).

As Keohane recognized when he discussed reputation in *AH*, the concept has considerable appeal for scholars who take the realist, rational-actor challenge seriously but remain optimistic about international cooperation. Given the acknowledged difficulties of direct enforcement through

---

“the realist argument that national actions are governed entirely by a calculation of interests is essentially a denial of the operation of normative obligation in international affairs” (Chayes and Chayes, *New Sovereignty*, p. 8). If one wants to rebut the realists’ arguments, the stronger approach would appear to be showing that such rebuttal is possible even granting their assumptions, as Keohane did in *After Hegemony*, rather than (with Chayes and Chayes) merely railing against the assumptions.

<sup>2</sup>Reputation has a long history in the bargaining literature, for example. Thomas Schelling discussed reputation at some length in his classic *The Strategy of Conflict* (Cambridge: Harvard University Press, 1960); see especially chapter 2, “An Essay on Bargaining.” See also Philip B. Heymann, “The Problem of Coordination: Bargaining and Rules,” *Harvard Law Review*, vol. 85, no. 5 (March 1973), pp. 822-823.

<sup>3</sup>See also Fearon (1994). Fearon’s model of audience costs, however, does not rely on reputation explicitly, particularly not in the international context, although he does not quite explain how domestic “audience costs” are generated (other than to say that governments can create audience costs by “engaging the national honor,” which sounds at least related to reputation).

retaliation, reputation plays a key role in explaining cooperation among equal powers, even while accepting the realist assumption that states act in their own interests. Adopting the assumption of rational behavior provides a theoretically consistent and rigorous test of theories of cooperation. And while strict rationality is surely not a literal representation of reality, political actors surely act in their (perceived) self-interest at least some of the time; it seems reasonable to suppose that self-interested behavior (as opposed to, say, rote rule-following) is more likely the higher the stakes. Indeed, as discussed above, even Chayes and Chayes attempt to frame their argument in terms of states' interests. An argument for rational, self-interested cooperation seems essential for those who would argue that meaningful international cooperation is possible.

By the same token, if reputation does not work, the arguments advanced in *AH* and elsewhere about self-interested cooperation among egoists become less convincing. Although reputation-based arguments place theories of international cooperation firmly on a rational-actor footing, they are weakened by assumptions imported from economic theory. The economics literature first conceived of reputation in terms of firms contesting markets, and adopted the game-theoretic convention of viewing firms as long-lived, unitary actors. As proponents of reputational theories in international relations themselves acknowledge, the unitary actor assumption is a weak one in models of politics. Abandoning it, however, presents difficulties for reputational arguments.

Consider government turnover. When national leaders are seen as the decision makers, rather than a disembodied "state," reputational arguments appear less solid. Why should a given leader care about preserving a reputation when she will be out of office in a few years? And why should foreign leaders trust past actions to be a guide to present behavior, when governments change? Reputational arguments depend crucially on links between past and future actions: in order for reputations to be meaningful, states must be confident that how another state has behaved in the past provides clues to its future behavior. Indeed, this link is formalized in the notion of "type" used in some reputational models. Such an assumption may be reasonable for a "long-lived" player. But if decision makers change over time, what underlies a state's reputation?

Variable state preferences – for example, due to changing party control over governments – introduce further difficulties. Why should a leader maintain a reputation when the payoff from that reputation will be reaped by future leaders? Reputation is commonly viewed as an "asset": a record of compliance creates a good reputation, which in turn allows a state to enter into agreements with other countries. Thus reputation can be seen as enhancing the flexibility of future leaders to carry out their policy goals. It follows that reputation cannot be an unalloyed good from a partisan perspective; indeed, a government expecting that future governments will have different policy preferences might have an incentive to destroy its country's reputation, in order to hinder the ability of future governments to make policy.

Both critiques highlight the same fundamental issue: existing arguments of reputation rest (implicitly or explicitly) on models of "states" as long-lived, continuous unitary actors.<sup>4</sup> States in the real world, on the other hand, are internally dynamic, their domestic politics replete with partisan squabbles and government turnover. If the mechanism of reputation is to work in the real world, some "transmission mechanism" is needed to tie together the interests of current and future

---

<sup>4</sup>It is important to note here that I am not claiming that scholars who have employed reputational arguments have ignored domestic politics altogether, or are ignorant of the restrictiveness of the unitary-actor assumption. To the contrary, both Keohane and Martin explicitly discuss domestic politics and its links to international relations. My point here is that domestic concerns have not been brought into reputation, so that arguments about reputation still rest on the unitary-actor assumption.

governments: a light source to cast the shadow of the future.

David Kreps (1990) has considered a similar problem in an economics setting. In his simple model, a firm and a customer played a “trust/honor” game in which in each period, a customer decided whether or not to trust a firm, and the firm then decided whether or not to honor that trust. In the one-shot game, the firm would abuse the customer’s trust; in an infinitely repeated game with a long-lived firm, the firm might honor the customer’s trust, in order to earn the trust of future customers. But what if the firm is run by short-lived managers? Kreps points out that in this case, the firm’s reputation can be an asset, with the sale value of this asset providing sufficient incentive for each short-lived manager to honor the customer’s trust. That is, an equilibrium exists (albeit one of many) in which every manager buys the firm from the previous manager, honors the trust of the customer, and then sells the firm to his successor. The value of reputation is thus capitalized in the firm, and each short-lived manager has an incentive to preserve the reputation for the future.

The “sale of the firm” does not carry over immediately into the political setting, but the spirit of Kreps’ model remains. Like Kreps, I seek to uncover a transmission mechanism linking short-lived governments and giving them incentives to act consistent with a long-term reputation. In the next section, I develop a model that explores a particular “transmission mechanism” in politics: political parties.

### **3 A simple model of domestic politics and compliance**

I proceed by constructing a simple model of domestic politics and compliance with international agreements. The aim is to focus on the set of issues surrounding government preferences and government turnover. As is often the case with simple models, I will make some heroic assumptions; the hope is that these capture the essential points and elucidate the basic dynamics at work, without obscuring or neglecting fundamental aspects of the problem. And as I will discuss below, the assumptions themselves are often useful in staking out the range of reputational arguments, specifying the necessary conditions for commitment problems to exist and for reputation to work.

A short digression is worthwhile here, at the outset of my discussion on reputation. As Drew Fudenberg and Jean Tirole (1995) have pointed out, there are two separate conceptions of “reputation” in the formal game-theoretic literature. One approach conceives of reputations within games of imperfect information, in which there is some positive (although perhaps tiny) probability that a given player is a “type” that deviates from the “rational” short-run action; for example, a player who always honors commitments. In such games, player A’s reputation is simply the assessment by other players of the probability that A is of a “cooperative” type; and given a sufficient number of periods remaining in a game, even rational players (who would renege in a single-shot game) may honor their commitments in order to mimic the “crazy” types and thus induce other player to enter into agreements with them. This framework is developed in the classic set of papers by David Kreps, John Milgrom, Paul Roberts, and Robert Wilson (1982).

An alternative approach operates within indefinitely repeated games of complete information, and identifies reputation with a “good record” of past compliance. As Fudenberg and Tirole (1995) note, “in the repeated prisoner’s dilemma the equilibrium in the ‘grim’ strategies ‘cooperate until an opponent defects, and then defect thereafter,’ can be interpreted as describing a situation where each player has a ‘reputation’ for cooperation that vanishes the first time he defects” (p. 367, footnote 1). Kreps (1990), for example, takes this tack in his article on reputation and the firm.

This approach is also roughly consistent with the international relations literature on reputation, which has emphasized the way in which repeated play helps players escape from the mutual-reneging outcome of a one-shot Prisoner's Dilemma.<sup>5</sup>

From a formal standpoint, as Fudenberg and Tirole point out, the latter conception of reputation – framed as it is in infinitely repeated games – has little predictive power, since such games have multiple equilibria (the result known as the Folk Theorem). On the other hand, the predictive power of the former class of models is substantially weakened when multiple long-run players are involved; as Fudenberg and Eric Maskin (1986) have shown, equilibrium in these models depends greatly on the exact assumptions about the form of incomplete information. In the current setting, therefore, any “predictive power” gained from employing imperfect information would be eroded by the need to specify the exact nature of uncertainty and the players’ responses to it. At any rate, because my essential task here is to explore, at a fairly rudimentary level, the implications of domestic politics for the prospects of compliance with international relations, I employ the simpler complete-information conception of “reputation.”

### 3.1 Setup of the model

With these caveats in mind, consider the following infinitely repeated game,<sup>6</sup> which I will call the “international agreements game.” Some number  $N + 1$  countries are potentially engaged in bilateral international agreements with each other, where  $N \geq 1$ . I will refer to country 0 as the “home country,” and countries 1, ...,  $N$  as the “foreign countries.” Throughout this paper, I will (without

---

<sup>5</sup>The international relations literature is a bit murky on this point. Scholars have emphasized the importance of repeated play, and (as a proxy for repeated play) issue linkage, suggesting an affinity for the second of the two interpretations mentioned here. Although Keohane also ties reputation to his discussion of asymmetric information and the “lemons problem” in *After Hegemony*, his treatment of the effects of imperfect information blurs important formal distinctions between models focusing on quality uncertainty (such as the quality of used cars in the “lemons problem”) and models focusing on uncertainty over intentions, or over an actor’s ability credibly to commit to comply. In the more recent “Interests, Commitments, and Institutions,” Keohane bases his discussion more closely on the imperfect-information models of Kreps et al., but such models may be harder to import to international relations than they first appear, as discussed in the text. Lipson and Martin are much less precise about their notions of reputation. On the other hand, a paper by James Alt, Randall Calvert, and Brian Humes is most explicit, exploring a kind of reputation effect in a simple two-period game between a hegemon and allies; the game is a sort of truncated and somewhat simplified version of the market-challenge games explored by Kreps et al. See Alt, Calvert, and Humes (1988). Since their model has such a dramatically different formal representation of “reputation” than the one modeled here, it is not discussed further.

<sup>6</sup>I choose to model infinitely repeated games for two reasons. First, although it may sound paradoxical at first, infinitely repeated games are probably a better representation of reality than finite games; at a formal level, an infinite game can be interpreted as a game which will continue indefinitely (it may end someday, but that date is uncertain), whereas finite games are distinguished by their certain end date. The former situation seems more suitable in the current context. Second, as just discussed, the concept of reputation in infinitely repeated games avoids the need for modeling imperfect information, which would add considerable complexity to the model (without adding much insight to the topics we want to explore).

A further caveat is in order here. Because we consider an infinitely repeated game, we know in advance from the Folk Theorem that multiple equilibria are possible aside from the one explored here. One might defend the equilibrium chosen here on the grounds that it Pareto dominates other equilibria (for example, equilibria in which governments cooperate in every other period, or never cooperate at all). In any case, however, our aim here is less to produce a “predictive” model than to explore how reputational considerations could arise out of strictly self-interested behavior and to suggest links between domestic politics and international relations. Indeed, given the heroic nature of the assumptions required to reduce international relations to a simple model, one might be tempted to doubt the “predictive power” even of models with more restricted sets of equilibria; but we leave that discussion aside here.

loss of generality) focus on the decisions made by the home country.

Suppose there are two types of bilateral agreements that can be signed; call them “type I” and “type II.” The degree of difference between them is left unspecified: these two types of agreements might variously represent trade agreements with respect to two different industries; a trade agreement and a monetary agreement; or a trade agreement and an environmental agreement. However, the two types of agreements are “linked”: a violation of a type I agreement is taken by all countries as equivalent to a violation of a type II agreement, and vice versa.<sup>7</sup> Suppose further that in each country there are two parties,  $X$  and  $Y$ , with different policy preferences: party  $X$  benefits from type I agreements but opposes type II, with  $Y$ ’s preferences diametrically opposed.

In each country, at the start of each period, one of the parties is elected and forms a government. The party controlling each country’s government is assumed to be common knowledge. Let  $p$  be the probability that party  $X$  wins in the home country, and  $q$  be the probability that  $X$  wins in each foreign country.<sup>8</sup> During each period, each government must make a set of simultaneous decisions: for each partner country, it must decide whether or not to sign an agreement of type I; whether or not to sign an agreement of type II; and whether or not to comply with existing international agreements.<sup>9</sup> Each decision is assumed to be common knowledge.

For simplicity, I assume that each agreement, once jointly signed, takes effect during the following period and lasts for just that period.<sup>10</sup> Thus if two governments sign an agreement in period  $t$ , it becomes relevant in period  $t + 1$ . Each government thus inherits up to  $N$  agreements from its predecessor (one per partner country). At the beginning of the next period, the party in power in each country is either reelected or replaced by the opposition, and the process is repeated. Since the parties have different payoffs, in my model the preferences of state governments thus change over time. Even without introducing uncertainty over “types,” therefore, I have captured the notion that the preferences of future governments are not known with certainty. As will be clear below, this uncertainty plays a key role in creating the “commitment problem” explored here.

I model payoffs so that the *benefit* to the home country from an agreement derives from compliance by the foreign country, while the *costs* of the agreement are due solely to home-country compliance. For example, the benefit of a free-trade agreement comes largely from the opening of

---

<sup>7</sup>In section 4.3 below, we allow for “degrees of linking” by allowing some probability that the two agreements are not linked. As we show in that section, a greater degree of linkage makes cooperation “easier” or “more likely,” and thus is preferable for all countries *ex ante* (although it may not be preferable *ex post*.)

<sup>8</sup>Three extensions are worth mentioning here. First, it would be straightforward to allow for different probabilities among the foreign countries; this, however, would add much more algebra than insight. Another possibility, still staying within the framework of “probabilistic” elections, would be to incorporate an “incumbency bonus”; as the model currently stands, a given party has the same probability of election whether it is an incumbent or in opposition. Finally, an attractive but more complicated extension would endogenize electoral outcomes; this would require more detailed model of the underlying interest groups and the electoral system. Since our aim here is to focus on compliance, we have set aside those issues, and model the election process as random.

<sup>9</sup>We assume that decisions are made simultaneously for two reasons. First, if the decisions are to be made sequentially, there is no *a priori* reason to decide which should be made first; but the order affects the possible outcomes. Putting the signing decisions first would not affect the results of the model, since the signing decision is purely partisan, and unaffected by reputation considerations (although that result, of course, is sensitive to assumptions about the international-level equilibrium, as discussed below). On the other hand, if the compliance decision is made first in every period, then the signing decisions later in the period could be made contingent upon the compliance decision. While this is a “reputational” result, it is simpler and perhaps less interesting than the one we investigate in the model that follows.

<sup>10</sup>Extending the life of international agreements would complicate the model considerably, without affecting the basic results. A longer life would simply increase the “commitment payoffs”  $\bar{\pi}_X$  and  $\bar{\pi}_Y$  discussed below.

foreign markets, while the costs are borne by domestic industry.<sup>11</sup> One can think of these payoffs as the benefits and costs to underlying interest groups in the home country. In this simple framework, each party can be thought of as being “allied” with particular interest groups in society; thus each party captures only the benefits and costs associated with that segment of society, rather than the benefits to society as a whole.<sup>12</sup>

In this section, I consider a simple model in which the costs and benefits of agreements are fixed.<sup>13</sup> To be concrete, let  $x$  be the per-period cost to party  $X$  from home-country compliance with a type II agreement, and  $y$  the per-period cost to  $Y$  from compliance with a type I agreement. Let  $b$  represent the benefits of an agreement to the party that prefers it, so that  $X$  gets a payoff of  $b$  when the foreign country complies with a type I agreement, and  $Y$  gets a payoff of  $b$  from foreign compliance with a type II agreement. I also assume that there is a (possibly vanishingly small) cost  $\varepsilon > 0$  associated with renegeing on a favorable agreement; this cost might be thought of as the “cost of bad publicity” or perhaps as the ideological cost associated with not complying with an agreement signed by one’s party.<sup>14</sup>

To illustrate these payoffs, suppose that the home country signs a type I agreement with another country in period  $t$ . Then their payoffs in period  $t + 1$  (looking only at the payoffs from the compliance decision) are shown in Table 1.

Of course, a single party controls the government in each country, and this party makes its decisions based solely on its own payoffs. For example, in the payoff matrix in Table 1, a party  $X$  government in the home country earns a payoff of  $b$  if both countries comply; 0 if it complies but the other country reneges;  $b - \varepsilon$  if it reneges while the other complies; and  $-\varepsilon$  if both countries renege. Importantly, however, both parties receive payoffs regardless of which party is in power. That is, if a type I agreement is in force, and is complied with by the foreign government, party  $X$  gets a payoff of  $b$  even if party  $Y$  controls the government.

I also assume that signing an agreement is costly; this cost could represent the opportunity cost of the time, diplomatic effort, and political capital needed to write an agreement and secure

---

<sup>11</sup>This is clearly somewhat of an oversimplification: in particular, home-country consumers would also benefit from the opening up of home-country markets to foreign goods. The “dichotomous” payoff structure is adopted here to simplify the analysis, but does not fundamentally affect the results: the key characteristic of the payoffs is that the parties’ interests are opposed. Moreover, the approach taken here could be defended by the familiar collective-action argument that from a political standpoint, the payoffs that matter are those captured by organized groups, such as business or organized labor, rather than consumers.

<sup>12</sup>We are clearly sweeping several important aspects of domestic politics under the rhetorical rug. Perhaps the focus on parties is best thought of as a “reduced form” of a more complicated model attending to the relevant interest groups, their organizational strength and influence on elections, etc. For example, the assumption that parties have policy preferences (their payoffs are determined by the international agreements in place) could be derived from a model in which parties seek election through campaign contributions from interest groups, and party differentiation follows from the existence of multiple interest groups. Clearly, the model we develop here implicitly relies on the underlying interest groups having comparable strength in the electoral arena, so that parties emerge on both sides; but this does not seem to be a misrepresentation of real-world party politics.

<sup>13</sup>In section 4, we consider two extensions: first, the possibility that the benefits of an agreement, rather than being constant, vary with the costs of compliance; second, the possibility that the costs of compliance, rather than being the same for all agreements, are drawn from some distribution.

<sup>14</sup>As can be seen in the payoff matrix, this cost is designed merely to eliminate indifference among choices; it does not affect the fundamental commitment problem. The sole effect of  $\varepsilon$  is to eliminate indifference by party  $X$ . In particular,  $\varepsilon$  plays no role in the crucial issue of a party’s choice with unfavorable agreements, such as party  $Y$ ’s choice over whether or not to comply with type I agreements. Our results would be identical if, rather than assuming  $\varepsilon > 0$ , we assumed that parties complied with agreements when they were indifferent.

sufficient domestic support for its enactment or ratification. Let  $c$  denote the cost of signing an agreement. As I will see shortly, some cost is necessary to avoid “cheap talk” outcomes in which governments sign agreements which they know may not be fulfilled. Nonagreement yields a zero payoff for both parties. Finally, I assume that each party applies a discount factor of  $\delta$  to future payoffs, and seeks to maximize the present discounted value of all future payoffs. In the current setting, where the decision makers are implicitly taken to be party leaders, the discount factor can be taken to reflect not only time preference for “getting things done now,” but also the uncertainties inherent in a politician’s future.

In each period (stage game), the strategy of each government consists of one action from each of the following pairs, for each of the  $N$  partner countries: {Sign type I if controlled by party  $X$ , Reject type I if controlled by party  $X$ }; {Sign type I if controlled by party  $Y$ , Reject type I if controlled by party  $Y$ }; {Sign type II if controlled by party  $X$ , Reject type II if controlled by party  $X$ }; {Sign type II if controlled by party  $Y$ , Reject type II if controlled by party  $Y$ }; {Comply with a type I agreement if controlled by party  $X$ , Renege on a type I agreement if controlled by party  $X$ }; {Comply with a type I agreement if controlled by party  $Y$ , Renege on a type I agreement if controlled by party  $Y$ }; {Comply with a type II agreement if controlled by party  $X$ , Renege on a type II agreement if controlled by party  $X$ }; and {Comply with a type II agreement if controlled by party  $Y$ , Renege on a type II agreement if controlled by party  $Y$ }.<sup>15</sup> In a slight abuse of notation, I will denote the stage-game strategies of parties  $X$  and  $Y$  when each is in power by  $s_X$  and  $s_Y$ .

### 3.2 The commitment problem

To illustrate the nature of the “commitment problem” posed by this game, consider a simple version lasting two periods and played by two countries, in which the governments decide in period 1 whether or not to sign agreements, and then decide in period 2 whether or not to comply. As a benchmark, suppose that both parties, at the start of the game (at the beginning of period 1, before the elections), can credibly commit to comply with any treaty in period 2. If such commitments are made, both parties will then comply with both types of agreements in period 2; working backwards, a party  $X$  government in period 1 gets a present-value payoff of  $\delta b$  from signing a type I agreement in period 1. As long as  $\delta b > c$ , then, party  $X$  will sign a type I agreement; similarly if  $\delta b > c$ , party  $Y$  will sign a type II agreement.<sup>16</sup> Given these assumptions, party  $X$ ’s expected payoff *ex ante* is

$$\bar{\pi}_X = pq(\delta b - c) - (1 - p)(1 - q)\delta x \quad (1)$$

This is party  $X$ ’s “commitment payoff”: the expected payoff it receives in the two-period game when commitment is possible. With probability  $pq$ , party  $X$  is elected in both countries in period 1, both governments sign a type I agreement (at a signing cost of  $c$ ), and compliance yields  $b$  in period 2 (for a discounted payoff of  $\delta b$ ). With probability  $(1 - p)(1 - q)$ , both governments are controlled by party  $Y$  and sign a type II agreement, yielding a discounted payoff of  $-\delta x$  for party

<sup>15</sup>Stage-game strategies are spelled out more formally in the Appendix. Note that there are eight pairs of actions, corresponding to the three decisions, because a strategy must specify a course of action in each possible situation (and therefore must be contingent upon the type of agreement and the party in the other country).

<sup>16</sup>Although the assumption that  $\delta > c$  ensures that both parties will prefer to sign favorable agreements, it does not guarantee that society is thereby made better off; if, for example,  $\delta > c > \delta(1 - x)$ , overall social welfare would be higher under nonagreement on type I. This is a simple case of a “political externality”: party  $X$  ignores the costs on party  $Y$ , and as a result “too many” agreements are signed. This interpretation of partisan politics might present an interesting subject for further exploration, but we set it aside for now to concentrate on the model at hand.

$X$ . As long as  $x < \frac{pq(\delta b - c)}{\delta(1-p)(1-q)}$ , this net expected payoff  $\bar{\pi}_X > 0$ . Similarly, if  $y < \frac{(1-p)(1-q)(\delta b - c)}{\delta pq}$ , party  $Y$ 's commitment payoff  $\bar{\pi}_Y$  is positive.

Unhappily, however, external commitment mechanisms are unavailable in a world of sovereign states with no higher government or court to enforce compliance. If commitment is not feasible, what happens to international agreements in my simple model? Consider the decisions facing governments in the second period, when a type I agreement has been signed in period 1, but commitment is not possible. A party  $Y$  government will always Renege on type I (unfavorable) agreements. A party  $X$  government, meanwhile, will always Comply with type I agreements, since once signed those agreements are costless to party  $X$  (note that the “reneging cost”  $\varepsilon$  eliminates indifference). Thus there are four possible (dominant strategy) equilibria in the second period, depending on the parties in power: If two party  $Y$  governments face each other, they will both Renege; if two party  $X$  governments face each other, they will both Comply; if the governments are of different parties, the  $Y$  government will Renege, and the  $X$  government will Comply. In the absence of commitment, a type I agreement yields party  $X$  in country 0 a payoff of  $b$  if party  $X$  is in power, and nothing if party  $Y$  is in power.

In the first period, the discounted expected benefit to party  $X$  from signing a type I agreement is  $\delta qb$ . If the signing cost  $c$  is greater than  $\delta qb$ , therefore, party  $X$  will not be willing to sign even a favorable type I agreement. Similarly, if  $c > \delta(1-q)b$ , party  $Y$  will not sign a type II agreement. If both conditions hold at once, so that  $c > \delta qb$  and  $c > \delta(1-q)b$ , then no agreements are signed in period 1, and both parties get a payoff of zero.

This is the essence of the commitment problem: neither party signs favorable agreements, even though both parties would be better off *ex ante* if they could commit themselves to comply. Note that the structure of the commitment problem depends on the cost of signing an agreement, relative to the benefits and the discount rate. When agreements are very costly, or the future is discounted heavily, agreements would not be worthwhile in any case – even if commitment were possible. On the other hand, if agreements are very cheap, or the benefits are very high, governments can sign them even if they expect that their political opponents will renege in the future. When costs are high enough to make “conditional fulfillment” unattractive, but low enough that universal fulfillment would yield net benefits, both parties would be better off if they could commit to compliance.

All three cases may occur in the real world. On some issues, international cooperation might offer benefits in theory; but the complexities involved in writing a comprehensive accord or securing political support might make an actual agreement prohibitively costly.<sup>17</sup> On other issues, “throw-away” or “symbolic” agreements may be commonplace, producing net gains to their proponents even if they are ultimately unfulfilled. To the extent that I am interested in the “commitment problem” or the “puzzle of compliance,” however, only the last case is relevant. In the remainder of the paper, therefore, I shall assume that a commitment problem exists: that is, that  $\delta b > c > \delta qb$  and  $\delta b > c > \delta(1-q)b$ , and that no external commitment mechanism is available to bind states to comply with agreements.

Note that this commitment problem results not from a Prisoner’s Dilemma, but from partisan differences. The obstacle to compliance in my model is not that the same parties who signed an agreement face an *ex post* temptation to renege (a temptation they could have foreseen from

---

<sup>17</sup>There may be parallels here to the “incomplete contracting” literature in economics: writing “complete” international agreements that provide for all possible contingencies may prove unfeasible. In such cases, “fulfillment” would be hard to define, and therefore hard to verify, and a possible result would be the breakdown of agreement, even though gains might be possible “in theory.”

the outset). Although that framework is bound to apply in many cases, it has been extensively studied in the literature. I focus here on a different mechanism by which commitments become unattractive: changes in government preferences driven by shifts in political power. The model here draws a sharper distinction between *ex ante* and *ex post* optimality. In the Prisoner’s Dilemma, both parties would prefer to commit themselves to comply, even *ex post*; in the current model, however, a party *Y* prefers mutual renegeing to mutual cooperation on a type I agreement (considered in isolation). Nonetheless, the party would, *ex ante*, prefer a world of credible commitment. In a sense, the issue is one of preferences rather than of incentives. Moreover, the framework studied here seems more compatible with the criticisms of Chayes and Chayes, for example: the issue is no longer “how can I sign an agreement now and renege later?” but “how can I sign an agreement now, knowing that my successors may disagree with its goals and thus renege?”

### 3.3 A reputational equilibrium

The commitment problem in the previous section is a product not only of relative costs and benefits, but also of the short time horizon. Suppose, then, that the “international agreements game” is played an infinite number of times; or, what is the same thing from a formal point of view, an indefinite number of times, so that the end is never certain. The “non-cooperation” outcome remains an equilibrium: if both parties refuse to comply with unfavorable agreements, and the costs are as assumed above, then one possible equilibrium is for no-one to sign any agreements, earning a payoff of zero every time. However, another equilibrium is possible. Indeed, from the Folk Theorem we know that an infinite number of equilibria are possible; here I focus on the one which is Pareto-optimal for all countries, so that (if the two parties were to sit down and agree on an equilibrium) this is what they would choose.

I propose the following “reputational equilibrium.”<sup>18</sup> Choose a time horizon  $T \geq T^*$ , where  $T^*$  is defined precisely in equation (6) below; as will become clear, it is chosen to ensure sufficient incentive for compliance. In period  $t$ , the government in each country complies with any existing international agreements as long as all governments in all countries have complied with all international agreements for at least the last  $T + 1$  periods:  $t - T - 1, t - T, \dots, t - 1$ . (If  $t < T$  then each country complies with existing international agreements as long as none has ever renege.) Moreover, each government signs the type of treaty it prefers (type I if party *X* is in power, type II for party *Y*) and rejects the type of treaty it dislikes. I call this the “normal phase” of play.

If any country reneges on an agreement, play enters a “punishment phase” starting in the next period. As punishment, the punishing countries refuse to sign or comply with any agreements with the offender. This punishment lasts  $T$  periods, after which play reverts to the “normal phase” described above. If any country violates the punishment phase (*e.g.*, a punishing country signs an agreement with the offender during the punishment phase), then that country becomes the “object of punishment” and a new punishment phase begins.

I thus interpret a violation of an agreement to involve “loss of a good reputation”: that is, a country’s reputation is identified as its record of past compliance with international agreements.<sup>19</sup>

---

<sup>18</sup>This equilibrium is stated rigorously in the appendix, where a full proof that it is a subgame perfect equilibrium can also be found.

<sup>19</sup>In this model, as in unitary-actor models of reputation in international relations, reputation is identified with a country. One might ask: since the parties are making the decisions, why not identify reputation with each political party – so that party *X* might have one reputation, while party *Y* has another? If reputations adhered to parties, however, rather than to countries, the incentive to comply would be diminished. Party *X*, for example, is tempted

I also allow the punishment horizon to be finite, which can be interpreted as allowing “damaged” reputations to be repaired with the passage of time (although the model is also perfectly consistent with infinitely long punishments, involving so-called “grim” punishment strategies.) Note also that in the proposed equilibrium punishment follows any instance of renegeing: this is a *full compliance* equilibrium.

I now argue that the strategies described above comprise a subgame perfect equilibrium in the international agreements game.<sup>20</sup> Consider the choices facing party  $X$  in period  $t$  when it is in power in country 0, and when all countries have complied with agreements in every previous period. (The conclusions generalize readily to the case when  $Y$  is in power.) The signing decisions are straightforward: party  $X$  benefits from type I agreements, and prefers no agreement to type II agreements.

The compliance decision is more complicated. I will begin by considering the “hardest case” for compliance: namely, when a party  $X$  government faces type II agreements and  $Y$ -controlled foreign governments in every other country. Let  $s_X^* = \{\text{Sign I with all countries; Reject II with all countries; Comply with I with all countries; Comply with II with all countries}\}$ , and define  $s'_X = \{\text{Sign I with all countries; Reject II with all countries; Comply with I with all countries; Renege on II with all countries}\}$ , so that  $s_X^*$  represents the equilibrium (compliant) strategy and  $s'_X$  represents a “renegeing” strategy.

By complying, party  $X$  receives an immediate payoff of  $-x$  for each of the  $N$  existing agreements. But compliance also brings a benefit: by maintaining a history of compliance – that is, a good “reputation” – the government ensures a stream of payoffs from future agreements. Since both parties comply with all agreements in equilibrium, the expected payoff to  $X$  from an agreement with each country in any given period is exactly its commitment payoff. This flow of payoffs starts next period and continues on forever; the present value of the benefit from compliance is therefore  $\frac{\delta}{1-\delta}$ . The payoff from playing  $s_X^*$  in this case is thus

$$g(s_X^*) = N \left( -x + \frac{\delta}{1-\delta} \bar{\pi}_X \right). \quad (2)$$

---

to renege on unfavorable type II agreements; a reputation for breaking such agreements would only undermine the rival party’s ability to sign those agreements, playing into party  $X$ ’s hands. One might construct an equilibrium in which a violation by one of the parties is punished by a refusal by other countries to sign treaties preferred by that party; but such an equilibrium would be one of “retaliation” and retribution, rather than reputation. Moreover, in such a case punishment becomes rather difficult, as a punished country may face negative payoffs during punishment (by complying with unfavorable treaties) with only a distant reward for compliance. And a party-based reputational equilibrium in which renegeing by one party resulted in no treaties being signed for the duration of the punishment period would be identical to the country-based reputational equilibrium considered here.

More importantly, a model of reputations based on countries and not parties coincides more closely with the common conception of “reputation” in international affairs. When policymakers invoke reputation in support of a given course of action, they do so in the name of a national reputation.

More generally still, the spirit of a “transmission mechanism” is to tie the actions of current governments to future ones, and therefore provide some basis for a reputation at the country level. While our use of parties as the transmission mechanism still sets aside much of domestic politics, the parties here can be understood as proxies for the interest groups they represent; that is, one might interpret our model of domestic politics as a “reduced form” of a more complicated model in which interest groups attempt to influence elections. In this interpretation, reputation might be seen as the degree to which other countries expect interest groups to pressure governments to comply.

<sup>20</sup>A “subgame perfect equilibrium” is one in which the strategies to be played in every “subgame” – e.g., the game starting from the punishment phase – comprise a Nash equilibrium. Importantly, this holds even for those subgames which are “off the equilibrium path,” i.e., which will never actually be reached during play. Subgame perfection thus rules out “noncredible threats,” such as punishments that would not be carried out *ex post*.

If the government reneges in period  $t$  (playing  $s'_X$ ), its immediate payoff is zero, rather than  $-x$ . However, renegeing triggers the punishment phase. Thus renegeing yields

$$g(s'_X) = N \left( 0 + \frac{\delta^{T+1}}{1-\delta} \bar{\pi}_X \right). \quad (3)$$

Punishment implies that for the next  $T$  periods, no agreements can be signed. Starting in  $T + 1$  periods, country 0 will be able to sign international agreements again, earning (in expectation) its commitment payoff in each period. The present value of that future stream of payoffs is the second term within the brackets.

For compliance to be a rational strategy in equilibrium requires that  $g(s_X^*) \geq g(s'_X)$  – that is, that

$$x \leq \bar{\pi}_X \left( \frac{\delta(1-\delta^T)}{1-\delta} \right). \quad (4)$$

Note that  $N$  (the number of countries) drops out of condition (4): because it appears in both  $g(s_X^*)$  and  $g(s'_X)$ , it does not affect the compliance decision. More generally, therefore, the number of players has no effect on the compliance decision.

The right-hand side of inequality (4) is the gain to party  $X$  from complying in period  $t$ : the discounted expected value of the stream of payoffs from agreements over  $T$  periods, which will be forfeited by renegeing. Loosely, therefore, one might label this term the *value of a reputation*. Similarly, the left-hand side of (4) is simply the up-front *cost of compliance*.

Clearly, for a government to comply with “unfavorable” treaties, the value of a reputation must outweigh the cost of compliance. Substituting equation (1) for  $\bar{\pi}_X$  into (4) and solving for  $x$  yields

$$x \leq \bar{x}(T) \equiv \frac{pq(\delta b - c)}{\frac{1-\delta}{\delta(1-\delta^T)} + \delta(1-p)(1-q)}. \quad (5)$$

The term on the right-hand side of (5), defined as  $\bar{x}(T)$ , is the compliance cost ceiling for party  $X$  as a function of the punishment horizon  $T$  and the other parameters.<sup>21</sup> This ceiling is the maximum short-term compliance cost that a party- $X$  government will bear to preserve its reputation, given the time horizon  $T$ . A similar cost ceiling can also be defined, identical to (5) except that  $p$  and  $q$  replace  $1-p$  and  $1-q$ , and vice versa.

While (5) gives an upper bound on  $x$  as a function of  $T$ , I can use it to derive the lower bound on  $T$  for a given  $x$ . Rearranging (5) and taking logarithms yields an equation for  $T^*$  in terms of  $x$ ,  $b$ ,  $\delta$ ,  $p$ , and  $q$ :

$$T^* = \frac{\log(1-\xi)}{\log \delta}. \quad (6)$$

where  $\xi \equiv \frac{1-\delta}{\delta} \left[ \frac{x}{\bar{\pi}_X} \right]$ , so that  $\partial \xi / \partial x > 0$ . The punishment length required to ensure compliance need not be long. For example, letting  $p = q = 0.5$ ,  $\delta = 0.95$ ,  $b = 1$ ,  $x = 0.5$ , and  $c = 0.1$ , a punishment horizon of  $T^* = 5$  periods would suffice to ensure compliance.

Equation (6) can be interpreted as conveying the “difficulty” of maintaining compliance: as the right-hand side term increases, the severity of punishment needed to ensure compliance increases.

---

<sup>21</sup>Note that the right-hand side of (5) is not equal to the right-hand side of (4). The latter – the “value of a reputation” – includes  $x$  (since it includes future unfavorable treaties, as well as favorable ones.)

As shown in the appendix, the right-hand side is increasing in  $x$  and decreasing in  $b, \delta, p,$  and  $q$ . That is, a longer punishment horizon is required the higher the cost of compliance  $x$ , the smaller the payoff from agreements  $b$ , the smaller the discount rate  $\delta$ , and the smaller the probabilities of election  $p$  and  $q$ . These results are intuitive, since a longer punishment is a harsher one; as the cost of compliance increases relative to the payoff from an agreement, the incentive to renege grows stronger; as the discount rate increases, the “shadow of the future” grows longer and a given punishment weighs more heavily (thus the horizon can be shortened); as the probability of  $X$ -election increases, the future benefits of compliance are greater, because party  $X$  is more likely to be in power (and thus to sign favorable agreements).

Inequality (5) and equation (6) (they are equivalent) impose implicit requirements on  $x$  and  $y$ : if the costs of compliance are too high, a government may prefer to renege on unfavorable agreements. Note from (5) that the compliance cost ceiling  $\bar{x}(T)$  grows as the punishment horizon recedes into the future. A longer punishment is a more severe punishment, making party  $X$  more willing to bear compliance costs to avoid it. At some point, however, the cost of abiding by an unfavorable agreement grows too large, so that party  $X$  will not comply even in the face of infinitely long punishment. This absolute ceiling on the cost of compliance can be derived by letting the punishment horizon  $T$  in (5) go to infinity:

$$\bar{x}_\infty \equiv \frac{pq(\delta b - c)}{\frac{1-\delta}{\delta} + \delta(1-p)(1-q)}. \quad (7)$$

The right-hand side of (7) imposes an upper bound on the cost of unfavorable agreements compatible with full compliance by party  $X$ . Again, a similar ceiling  $\bar{y}_\infty$  can be defined for compliance by party  $Y$ .<sup>22</sup>

The comparative statics results on  $\bar{x}_\infty$  are intuitive. Taking partial derivatives yields

$$\frac{\partial \bar{x}_\infty}{\partial \delta} > 0, \quad \frac{\partial \bar{x}_\infty}{\partial b} > 0, \quad \frac{\partial \bar{x}_\infty}{\partial c} < 0, \quad \frac{\partial \bar{x}_\infty}{\partial p} > 0, \quad \text{and} \quad \frac{\partial \bar{x}_\infty}{\partial q} > 0.$$

The compliance cost ceiling rises with the discount factor  $\delta$ : more weight on the future means a greater value attached to reputation, and thus a greater willingness to bear the cost of complying with agreements. Likewise,  $\bar{x}_\infty$  increases as the benefit of agreements grows, and decreases as they become more costly. Finally, as  $p$  and  $q$  increase, party  $X$  is more likely to be in power in the future, and therefore more likely to enjoy favorable agreements; as a result, it is willing to sacrifice more today to preserve its reputation.

A necessary and sufficient condition for some suitable finite punishment horizon  $T^*$  to exist<sup>23</sup> is therefore that the costs of compliance are not too high for either party; that is,

$$x < \bar{x}_\infty \quad \text{and} \quad y < \bar{y}_\infty. \quad (8)$$

If condition (8) holds, therefore, inequality (4) will be satisfied for  $T \geq T^*$ , and a government of party  $X$  will comply with its international agreements – even when facing an “unfavorable” agreement, and a foreign government controlled by the other party.

<sup>22</sup>Note that when  $p = q = 1/2$ , the bounds on both  $x$  and  $y$  will be identical; as  $p$  or  $q$  shrinks, the bound on  $x$  shrinks accordingly, while the bound on  $y$  grows. The roles of  $\delta, p,$  and  $q$  are discussed in more detail in section 4.

<sup>23</sup>If either condition in (8) holds with equality, the required  $T^*$  will be infinite, corresponding to the “ultimate punishment” in which an offender is permanently isolated from international agreements with the other country.

Moreover, if (8) is satisfied, then the conditions for compliance by party  $X$  are satisfied in every possible case. For example, consider a party  $X$  government facing only  $m < N$  type II treaties; the costs of compliance are now reduced ( $-mx$  rather than  $-Nx$ ) while the costs of renegeing (being punished) remain the same. Similarly, if party  $X$  controls governments in at least one foreign countries, the costs of renegeing will be higher, since renegeing this period will eliminate the payoff tomorrow from any favorable treaties signed this period.

To show that the proposed equilibrium is subgame perfect, I need to show that the “punishment phase” is “credible”: that is, that no player wishes unilaterally to deviate from the punishment routine. I show this rigorously in the appendix, but the intuition is straightforward: when the other players are following the equilibrium, one player’s refusal to punish an offender (say, by complying with an existing agreement, or by signing a new agreement) yields it nothing except punishment. Moreover, because only one offender is punished at a time, the punishing countries continue to earn payoffs from their bilateral agreements among themselves.

## 4 Implications of domestic politics

### 4.1 The costs of reputation

The model explored here uncovers a cost of reputation in the domestic setting: when policy preferences differ among potential governments – for example, when future governments might be controlled by a different political party – reputation has a downside. By complying with an agreement, and maintaining the country’s “reputation,” a government preserves the ability of its successors to make international agreements of their own. In effect, passing on a reputation to future governments is giving the opposition party an “asset” to use in the future.

This *cost* of reputation appears to have been overlooked in the literature. Interestingly, the general theme of intertemporal relationships has received a fair amount of attention, even by the same theorists who focus on reputation. Keohane, for example, emphasizes the value of reputation, and then almost immediately goes on to discuss international institutions as a means of “protecting against changes in preferences” (1984, p. 116). These two arguments are at odds. While international agreements may constrain the behavior of future governments with different policy preferences, preserving a reputation for compliance will do just the opposite, *enhancing* the ability of future governments to sign agreements they prefer.

When domestic politics are considered, the value of reputation depends on the tradeoff between the expected benefits of future “favorable” agreements, and the expected costs of future “unfavorable” ones. Expanding the right-hand side of (4), the value of a reputation is

$$[pq(\delta b - c) - (1 - p)(1 - q)\delta x] (1 - \delta^T) \left( \frac{\delta}{1 - \delta} \right). \quad (9)$$

The terms in square brackets make up the commitment payoff,  $\bar{\pi}_X$ ; together, they can be thought of as the “immediate” or per-period value of reputation. The connection is intuitive, since reputation is valuable only as a means of commitment. At the start of any given period, before the realization of elections, the expected benefit of a good reputation to party  $X$  is  $pq(\delta b - c)$ , or the discounted payoff from signing a favorable agreement. The expected *cost* of a reputation, on the other hand, is  $(1 - p)(1 - q)\delta x$ . Their difference is the expected net benefit – the short-term value of a reputation. The second and third terms in (9) then scale up this short-term value into its value over  $T$  periods.

From (9), it is clear that the net value of a reputation to party  $X$  increases in  $p$  and  $q$  – that is, the more likely  $X$  is to return to power in the future (and sign favorable agreements).

The model thus points out one crucial way domestic politics may matter for compliance: because reputation only matters if it brings future benefits, dimmer electoral chances translate into lower reputational incentives. In a way, the likelihood of future election augments the “shadow of the future” in determining the strength of reputational incentives.<sup>24</sup> Thus parties with little future electoral chances – for example parties elected on a short-term wave of popular resentment with the “establishment” – should be less likely to uphold prior international agreements. Indeed, as the probability of reelection becomes smaller, the game becomes more like the “one-shot” compliance game studied above. Of course, such weak parties may come to power only rarely, so that noncompliance would be the exception, rather than the rule. In the extreme case in which only one party ever wins elections, agreements would always be complied with (although only one type of agreement will ever be signed). At the other extreme, in which two parties are evenly matched, and both have solid electoral chances, we might also expect compliance to be high, and agreements would be more varied. There might, however, be some “middle ground” in which compliance was less likely: in which one party was elected infrequently enough that it saw little value in reputation, but still came to power frequently enough to disrupt patterns of compliance. This seems particularly relevant for parliamentary governments with multiple parties; an extension of the model to more than two parties might yield fruitful insights on this point.

## 4.2 Party politics and international agreements

Party politics and electoral power also have implications for the international agreements that may be signed by each party. Consider again the definition of  $\bar{x}_\infty$  in equation (7) above. Like the value of reputation, this compliance cost “ceiling” depends crucially on the probabilities of election: as the electoral chances of party  $X$  increase (at home or abroad), so does the maximum short-term compliance cost that party is willing to bear to preserve a reputation. On the other hand, as party  $Y$  is more likely to hold power, the agreements signed in the future are more likely to be *unfavorable* to party  $X$ .

One implication, at least within the confines of this simple model, is that more successful political parties should be willing to bear more costly international agreements. Suppose that compliance costs are variable and depend on type of agreement being signed; a far-reaching trade accord, for example, may exact higher costs of compliance (with correspondingly higher benefits) than a simple agreement. (As mentioned at the outset of the previous section, nothing in the model rules out the possibility that compliance costs vary over agreements of a given type – as long as they remain below  $\bar{x}$ .)

Condition (7) then suggests that “deeper” cooperation can be achieved on agreements that the weaker party prefers: that is, although a party with a lower electoral probability may come to power infrequently, it will be able (in my model) to sign agreements with relatively high compliance costs. The reason is simply that the depth of cooperation is determined by compliance costs, which are borne here by the other party; thus the stronger the other party, the deeper the cooperation that can be achieved. Weakness in one arena yields strength in another.<sup>25</sup>

---

<sup>24</sup>Indeed, from a formal standpoint, the probabilities of election could be folded into the discount factor; although in this case, each party would have its own discount factor, reflecting its probability of reelection.

<sup>25</sup>There are interesting parallels here to other “paradoxical” results along similar lines: Schelling (1960) highlights

### 4.3 Issue linkage at the domestic level

A third implication of the model is the importance of *issue linkage* at the domestic level. While issue linkage has been much discussed in the literature, most analysts have focused on two roles linkage plays in the international sphere. First, issue linkage may increase the scope for side payments among countries.<sup>26</sup> Second, as discussed above, issue linkage provides a kind of proxy for repeated interaction, and so increases the force of reputation (Keohane 1984).

The model explored here suggests another, complementary role for issue linkage: providing political leaders who oppose existing agreements an incentive to comply with them nonetheless. That is, if different leaders have different preferences over agreements, as in my model, they will have an incentive to renege on agreements they dislike – an incentive that does not exist in the “unitary actor” model of stable preferences. In the absence of linkage, therefore, governments opposed to a given agreement might simply renege on it – knowing that the reputation costs, if any, might actually work in their favor, by inhibiting future agreements of the same type. In a sense, without linkage, reneging could provide a “double dividend”: avoiding the costs of compliance, plus undermining the possibility for future unfavorable agreements.

In this model, an absence of linkage would imply that violation of a type II agreement would initiate “type II” punishment, while type I agreements could continue on unaffected. Such a scenario would provide a powerful incentive for party  $X$  governments to renege on unfavorable agreements, complying only with those agreements they preferred. But party  $Y$  would follow the same course, and compliance would break down. If parties selectively renege on unfavorable agreements, the original commitment problem recurs. Seen in this light, issue linkage provides the “cement” holding compliance together at the domestic level.

The model can be modified to explore the issue of linkage in greater depth. Rather than supposing, as above, that linkage is “automatic,” assume instead that whether or not these agreements are linked in a particular case is revealed after each incident of reneging, and let  $\lambda$  represent the *ex ante* probability that a violation of a type  $i$  agreement will be “linked” to type  $j$  agreements. The payoff from complying is unaffected; but the payoff from reneging is now greater, because with probability  $1 - \lambda$  a state which reneges on type  $i$  will still have the prospect of type  $j$  agreements during the punishment period. The payoff from reneging is now

$$g(s'_X) = N \left( 0 + \frac{\delta}{1 - \delta} (1 - \lambda) \bar{\pi}_X + \frac{\delta^{T+1}}{1 - \delta} \lambda \bar{\pi}_X \right). \quad (10)$$

Punishment for noncompliance is now represented by the last term above. As the probability of linkage goes to zero, the cost of being punished vanishes as well. Equation (4), the condition on the compliance cost  $x$  that ensures compliance by party  $X$ , now becomes

$$x \leq \lambda \bar{\pi}_X \left( \frac{\delta (1 - \delta^T)}{1 - \delta} \right) \quad (4')$$

where the only difference from equation (4) is the inclusion of the parameter  $\lambda$ . Again, the right-hand side can be interpreted as the value of a reputation. By making punishment more onerous, a the possibility that a bargaining party’s weakness – a driver’s inability to veer off a collision course, for example – becomes a strength in bargaining situations. And as already discussed, Keohane (1997) has pointed out a “commitment paradox”: weak states are better able to make credible commitments when commitment is enforced by the strong.

<sup>26</sup>See, for example, Sebenius (1983); Stein (1980); Keohane (1984), pp. 91-2; and Martin (1995).

higher value of  $\lambda$  increases the value of a good reputation.

Thus a higher value of  $\lambda$  increases the maximum compliance cost compatible with the “reputational equilibrium.” The more likely is linkage, the wider is the scope for compliance. The connection between linkage and reputational incentives is especially stark in this model, since (as discussed above) it offsets an underlying incentive for party  $X$  to renege on agreements it dislikes. Loosely speaking, condition (4′) thus shows that if the degree of linkage is endogenous (*i.e.*, a matter of choice), governments should rationally prefer a higher degree of linkage; in away, linking agreements creates a “commitment device” of sorts to overcome the obstacles inherent in policy disagreements at the domestic level.<sup>27</sup>

## 5 Conclusion

In a world of sovereign states, reputation may play a key role in motivating compliance, and making commitments credible. Most reputational arguments in the literature, however, have relied on the assumption, imported from the economics literature, that states are long-lived unitary actors. In this paper, I have taken a step towards incorporating the messy reality of domestic politics into stories of reputation. In the process, I have explored the implications of domestic politics – in particular, government turnover and partisan contests – for the abilities of state governments to make and keep commitments in the international arena. Parties (and by extension the interest groups that support them) can be seen as a key missing link: a “transmission mechanism” which connects the interests of future governments to the interests of current ones. Without such a transmission mechanism, reputational arguments fall apart: short-lived political leaders have little incentive, on their own, to develop a reputation for the future.

Introducing domestic politics, even at a simple level, suggests three insights for international agreements and the puzzle of compliance. First, in a world with partisan differences, reputation may involve costs for present governments, as well as benefits, since it represents an investment in the future – which may be controlled by other parties. The mechanism of reputation, then, depends, at least in part, on the probabilities that a given party will return to power. The shadow of the future is cast partly by the ballot box.

Second, the significance of electoral strength in determining the benefits of a reputation suggests, somewhat paradoxically, that *weak* parties may be able to sign “stronger” agreements. If the “depth” of cooperation is related to the costs of compliance, and if those costs fall primarily on other parties, then the strength of a party’s *opponents* determines the depth of cooperation it can achieve. The stronger the opposition party – the more secure its political future – the higher a compliance cost it will willingly bear to preserve a reputation.

Third, the need to bridge the gulf between partisan positions suggests a new, domestically-oriented role for “issue linkage.” When parties disagree on policy, each may be tempted to renege

---

<sup>27</sup>One difficulty with less-than-certain linkage is that it complicates the punishment phase somewhat. Suppose a party  $X$  government reneges on a type II agreement, but signs a type I agreement. If linkage does not occur, a party  $Y$  government entering office would have little incentive to comply with the existing type I agreement: renegeing only delays the renewal of “normal” relations by a few periods, off in the future (by restarting the “punishment clock.”) Moreover, renegeing on the type I agreement plays into the preferences of party  $Y$  anyway; in fact, it benefits from type-I punishment. Meanwhile, compliance imposes upfront costs. Thus opposition parties are unlikely to comply with unfavorable agreements early in the punishment phase. But then payoffs in the punishment period depend on the various possible permutations of party power, making them much more difficult to calculate.

on agreements which are in their opponent's interests. Linking behavior in one sphere to consequences in another sphere can counter such partisan temptations, providing incentives for each party to comply with the other's agreements – so that they can continue to capture future gains for themselves.

A number of interesting extensions of the model have been noted in passing. In particular, the assumption that electoral probabilities are exogenous is the most glaring oversimplification, and relaxing it could yield further insights. Much depends on how compliance affects elections; if voters reward compliance (perhaps because they recognize the need to maintain a reputation), then the incentives for compliance will be strengthened. A second extension would be to allow for a range of different agreements of different “depths” (and therefore costs) within each category, and make the choice of “depth” exogenous.

Reputation will not be a panacea, and should not be interpreted as solely responsible for supporting compliance. Indeed, the model explored here has underscored several necessary assumptions for reputation to be effective. If reputation is not the only means to ensure compliance, it nonetheless retains central importance – partly because it bears on other compliance pathways. Enforcement of compliance, for example, depends in turn on the ability of states credibly to commit to exacting punishment, as pointed out above; thus reputation can play a supporting role even in international arenas dominated by the pathway of “reciprocity.”

Even more interesting is the connection between reputation and norm-following, alluded to earlier. Keohane (1997) has pointed out that reputational arguments and arguments based on social norms are closely related arguments, couched in different terms and disciplines. He writes: “From an observational standpoint... it may be difficult to disentangle these two motivations.... Indeed, valuable reputations are often reputations for behaving in normatively conventional ways.... [S]eeking to maintain a reputation and behaving appropriately may be so closely connected that they are not even distinct in the minds of practitioners” (p. 18).

This is a powerful point, but one can carry it even further. Not only are reputational and normative arguments “observationally equivalent”: they can be seen as expressions of an *identical* underlying argument. Game theory and sociological approaches, in some sense, are two paths to the same vista. Of course, the posited behavioral models – the carefully calculating utility-maximizer versus the conventional rule-follower – are different, even contradictory; but then again, neither is taken to be a literal, universally applicable representation of behavior. In fundamental ways they complement each other. Reputational arguments can be interpreted as providing a rigorous rational justification for social norms, making normative arguments logically *consistent* with the maximization of self-interest. And the concept of norms, in turn, can be interpreted as guiding us through the minefield of multiple equilibria: where formal logic fails (in specifying the equilibrium to be played, or the prior beliefs to be assumed), arguments based on observation can provide signposts and benchmarks to explain why reputations develop in a certain way: why, for example, backing down from a confrontation might be taken as revealing “weakness” in one setting, but might create a subsequent expectation of “resolve” in another.<sup>28</sup>

Linked with normative interpretations, theories of reputation will continue to play a central role in arguments of self-interested cooperation. And as conventional military alliances recede in importance in international politics, replaced by monetary union, economic “globalization,” and international environmental accords, exploring the prospects for cooperation among sovereign states

---

<sup>28</sup>On this point, see Jonathan Mercer's (1995) criticism of formal theories of reputation.

remains a fundamental task. Connecting the decisions made by states in the international arena to the dynamics of domestic politics can help us understand, and perhaps improve, those prospects.

## References

- [1] Alt, James E., Randall L. Calvert, and Brian D. Humes. 1988. "Reputation and Hegemonic Stability: A Game-Theoretic Analysis." *American Political Science Review* 82(2): 445-466.
- [2] Chayes, Abram, and Antonia Handler Chayes. 1995. *The New Sovereignty*. Cambridge: Harvard University Press.
- [3] Downs, George W., David M. Rocke, and Peter N. Barsoom. 1996. "Is the Good News About Compliance Good News About Cooperation?" *International Organization* 50(3): 379-406.
- [4] Fearon, James D. 1994. "Domestic Political Audiences and the Escalation of International Disputes" *American Political Science Review* 88(3): 577-592.
- [5] Fudenberg, Drew, and Eric Maskin. 1986. "The Folk Theorem in Repeated Games With Discounting or With Incomplete Information." *Econometrica* 54: 533-554.
- [6] Fudenberg, Drew, and Jean Tirole. 1995. *Game Theory*. Cambridge: MIT Press.
- [7] Heymann, Philip B. 1973. "The Problem of Coordination: Bargaining and Rules." *Harvard Law Review* 85(5).
- [8] Keohane, Robert O. 1984. *AH*. Princeton: Princeton University Press.
- [9] Keohane, Robert O. 1997. "Interests, Commitments, and Institutions in U. S. Foreign Policy." Manuscript, Duke University.
- [10] Kreps, David. 1990. "Corporate Culture and Economic Theory." In James E. Alt and Kenneth A. Shepsle, eds., *Perspectives on Positive Political Economy*, pp. 90-143. Cambridge, UK: Cambridge University Press.
- [11] Kreps, David, and Robert Wilson. 1982. "Reputation and Imperfect Information." *Journal of Economic Theory* 27: 253-279.  
Kreps, David, Paul Milgrom, John Roberts, and Robert Wilson. 1982. "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma." *Journal of Economic Theory* 27: 245-252.
- [12] Lipson, Charles. 1991. "Why Are Some International Agreements Informal?" *International Organization* 45(4): 495-538.
- [13] Martin, Lisa L. 1993. "Credibility, Costs, and Institutions: Cooperation on Economic Sanctions." *World Politics* 45(3): 406-432.
- [14] Martin, Lisa L. 1995. "Heterogeneity, Linkage, and Commons Problems." In Robert O. Keohane and Elinor Ostrom, eds., *Local Commons and Global Independence*, pp. 71-91. London: Sage.

- [15] Mercer, Jonathan. 1995. *Reputation and International Politics*. Ithaca: Cornell University Press.
- [16] Milgrom, Paul, and John Roberts. 1982. "Predation, Reputation, and Entry Deterrence." *Journal of Economic Theory* 27: 280-312.
- [17] Schelling, Thomas. 1960. *The Strategy of Conflict*. Cambridge: Harvard University Press.
- [18] Sebenius, James K. 1983. "Negotiation Arithmetic: Adding and Subtracting Issues and Parties." *International Organization* 37: 281-316.
- [19] Stein, Arthur A.. 1980. "The Politics of Linkage." *World Politics* 33(1): 62-81.

## Appendix

In this appendix, I state the proposed equilibrium formally, and show that it is indeed a subgame perfect equilibrium. For the sake of notation, I let  $N = 2$ , but the proof generalizes readily.

The following notation is used:

$x$	cost to party $X$ of home-country compliance with type II agreement
$y$	cost to party $Y$ of home-country compliance with type I agreement
$c_I$	cost of signing a type I agreement
$c_{II}$	cost of signing a type II agreement
$p$	probability that party $X$ wins election in country 0
$q$	probability that party $X$ wins election in countries 1 and 2
$\delta$	discount factor.

As in the text, define:

$$\begin{aligned}\bar{x} &\equiv \frac{pq(\delta b - c)}{\frac{1-\delta}{\delta} + \delta(1-p)(1-q)}; \text{ and} \\ \bar{y} &\equiv \frac{(1-p)(1-q)(\delta b - c)}{\frac{1-\delta}{\delta} + \delta pq}.\end{aligned}\tag{11}$$

I make the following assumptions on the parameters:

- (A1)  $x < \bar{x}$ ;
- (A2)  $y < \bar{y}$ ; and
- (A3)  $\delta b > c > \delta b q$  and  $\delta b > c > \delta b(1 - q)$ .

Assumptions (A1) and (A2) are the only assumptions actually required to support the equilibrium discussed here; they are needed to ensure that some punishment exists that will deter renegeing. (If they are not met, the costs of compliance are so high that no party will comply with unfavorable agreements, whatever the length of punishment). The bounds shrink as the discount factor decreases or as the relevant party becomes less successful at the polls; if, for example,  $X$  is unlikely to win election either at home or abroad, so that both  $p$  and  $q$  are small, then its cost of compliance will be relatively constrained (although  $Y$  will be able to bear a correspondingly higher cost of compliance.)

As explained in the text, (A3) ensures that a commitment problem exists; if it does not hold, then the reputational equilibrium explored here still stands, but would lose its force in explaining signing. If the right-hand side inequalities did not hold, so that  $c$  was smaller than  $\delta b q$  and  $\delta b(1 - q)$  respectively, parties would sign agreements even in the absence of reputation effects (although they would not comply with unfavorable agreements). If  $c$  was greater than  $\delta b$ , on the other hand, agreements would not be worthwhile even with full commitment.

As before, without loss of generality I consider the decision by the government in the ‘‘home country,’’ country 0. In a slight abuse of notation, let  $s_X$  denote a stage-game strategy by a party  $X$  government, and let  $s_X^i$  denote the set of actions dictated by  $s_X$  towards country  $i$ , so that  $s_X = \{s_X^1, s_X^2\}$ . Then define the following sets of actions towards each country:

$$\begin{aligned}s_X^{i*} &= \{\text{Sign type I with country } i; \text{ Reject II with country } i; \text{ Comply with type I with country } i; \text{ Comply with II with } i\} \\ s_Y^{i*} &= \{\text{Reject I with } i; \text{ Sign II with } i; \text{ Comply with I with } i; \text{ Comply with II with } i\} \\ s_X^i &= s_Y^i = \{\text{Reject I with } i; \text{ Reject II with } i; \text{ Renege on I with } i; \text{ Renege on II with } i\}\end{aligned}$$

Finally, for party  $m$ ,  $m = X, Y$ , define the stage-game strategies  $s_m^* = \{s_m^{1*}, s_m^{2*}\}$  and  $s_X^0 = \{s_m^1, s_m^2\}$ .

As in the text, define the commitment payoffs for  $X$  and  $Y$  to be:

$$\begin{aligned}\bar{\pi}_X &= pq(\delta b - c) - (1 - p)(1 - q)\delta x \\ \bar{\pi}_Y &= (1 - p)(1 - q)(\delta b - c) - pq\delta y.\end{aligned}$$

Note that both commitment payoffs are positive by (A1) and (A2). Next, define  $T_X$  such that

$$\bar{\pi}_X \left(1 - \delta^{T_X}\right) \left(\frac{\delta}{1 - \delta}\right) = x. \quad (12)$$

Rearranging equation (12) and taking logs yields:

$$\begin{aligned}\delta^{T_X} &= 1 - \frac{(1 - \delta)}{\delta^2} \frac{x}{\bar{\pi}_X} \equiv 1 - \xi \\ \Rightarrow T_X &= \frac{\log(1 - \xi)}{\log \delta}.\end{aligned} \quad (13)$$

The second line, which defines  $T_X$  as a function of  $\delta, p, q, b$ , and  $x$ , is given as equation (6) in the text.

**Remark 1** *Given assumption (A1), a finite  $T_X > 0$  satisfying equations (12) and (13) exists.*

**Proof.** Assumption (A1) implies that  $\xi < 1$ . For  $\delta < 1$ , therefore,  $\log(1 - \xi)$  exists, and there is a finite  $T_X$  satisfying (13) and therefore (12). (Note that although  $\log(1 - \xi)$  is negative,  $\log \delta$  is also negative, so the right-hand side of (13) is positive, as required.) For  $\delta = 1$ , however, both the numerator and denominator of the right-hand side of (13) go to  $\log 1 = 0$ , so that (13) cannot be evaluated directly. Applying L'Hôpital's Rule to (13),

$$\lim_{\delta \rightarrow 1} T_X = \frac{\left(\frac{1}{1 - \xi}\right) \left(-\frac{\partial \xi}{\partial \delta}\right)}{\frac{1}{\delta}} \bigg|_{\delta=1} = \frac{x}{2[pq - (1 - p)(1 - q)x]},$$

so that  $T_X$  satisfying (13), and therefore satisfying (12), exists. **q.e.d.** ■

**Remark 2** *Given (A1) and  $T_X$  as defined in equation (13), the following statements hold:*

1.  $\frac{\partial T_X}{\partial p} < 0$ ,  $\frac{\partial T_X}{\partial q} < 0$ ;
2.  $\frac{\partial T_X}{\partial x} > 0$ ; and
3.  $\frac{\partial T_X}{\partial \delta} < 0$ .

**Proof.** By definition

$$\xi \equiv \frac{(1 - \delta)}{\delta} \frac{x}{\bar{\pi}_X} = \frac{(1 - \delta)}{\delta} \frac{x}{pq(\delta b - c) - \delta(1 - p)(1 - q)x}.$$

Taking derivatives of  $\xi$  yields

$$\frac{\partial \xi}{\partial p} < 0, \frac{\partial \xi}{\partial q} < 0, \frac{\partial \xi}{\partial x} < 0, \frac{\partial^2 \xi}{\partial \delta^2} > 0.$$

Finally, note that  $\frac{\partial T_X}{\partial \xi} = \frac{-(\log \delta) \left( \frac{1}{1-\xi} \right) - \frac{1}{\delta} \log(1-\xi)}{(\log \delta)^2} > 0$ , where I have used the facts that both  $\delta$  and  $1 - \xi$  are less than one. Parts (i) and (ii) follow immediately (using the Chain Rule).

The proof of (iii) is somewhat more involved:

$$\begin{aligned} \frac{\partial T_X}{\partial \delta} &= \frac{(\log \delta) \left( \frac{1}{1-\xi} \right) \left( -\frac{\partial \xi}{\partial \delta} \right) - \left( \frac{1}{\delta} \right) \log(1-\xi)}{(\log \delta)^2} \\ \text{sign} \left[ \frac{\partial T_X}{\partial \delta} \right] &= \text{sign} \left[ (\log \delta) \left( \frac{1}{1-\xi} \right) \left( -\frac{\partial \xi}{\partial \delta} \right) - \left( \frac{1}{\delta} \right) \log(1-\xi) \right] \\ &= \text{sign} \left[ (\delta \log \delta) \left( -\frac{\partial \xi}{\partial \delta} \right) - (1-\xi) \log(1-\xi) \right]. \end{aligned}$$

Define  $h(\delta) = (\delta \log \delta) \left( -\frac{\partial \xi}{\partial \delta} \right) - (1-\xi) \log(1-\xi)$ . Then  $h(1) = 0$  and

$$\begin{aligned} \frac{\partial h}{\partial \delta} &= (1 + \log \delta) \left( -\frac{\partial \xi}{\partial \delta} \right) + (\delta \log \delta) \left( -\frac{\partial^2 \xi}{\partial \delta^2} \right) - (1 + \log(1-\xi)) \left( -\frac{\partial \xi}{\partial \delta} \right) \\ &= \log \left( \frac{\delta}{1-\xi} \right) \left( -\frac{\partial \xi}{\partial \delta} \right) + (\delta \log \delta) \left( -\frac{\partial^2 \xi}{\partial \delta^2} \right) > 0. \end{aligned}$$

In the last inequality above, I have used the fact that  $T_X$  is defined so that  $\delta^{T_X} = 1 - \xi$ , which implies that  $\delta > 1 - \xi$  and thus that  $\log(\delta/1 - \xi) > 1$ . Since  $h(1) = 0$  and  $\frac{\partial h}{\partial \delta} > 0$ ,  $h(\delta) < 0$  for  $\delta < 1$ , and part (iii) of the Remark follows. **q.e.d.** ■

Along similar lines, define

$$T_Y \equiv \frac{\log \left[ 1 - \frac{1-\delta}{\delta^2} \frac{y}{\bar{\pi}_Y} \right]}{\log \delta}, \quad (14)$$

so that  $T_Y$  satisfies the analogous condition to inequality (12) for party  $Y$ :

$$\bar{\pi}_Y \left( 1 - \delta^{T_Y} \right) \left( \frac{\delta}{1-\delta} \right) \geq y, \quad (15)$$

which captures the tradeoff for party  $Y$  between the value of a reputation and the costs of compliance. Along the lines of Remarks 1 and 2 above, it is straightforward to show, given (A2), that such a  $T_Y$  exists and that it is decreasing in  $p$  and  $q$ , increasing in  $y$ , and decreasing in  $\delta$ .

$T_X$  and  $T_Y$  represent the necessary punishment time horizons to ensure compliance by governments controlled by party  $X$  and party  $Y$  respectively. Pick the larger of the two and call it  $T^*$ . Note that which time horizon is larger depends on the electoral probabilities and on the relative sizes of  $x$  and  $y$ . When  $p = q = 1/2$ , so that both parties are equally likely to win in every country,  $T_X > T_Y$  if and only if  $x > y$ , that is, if and only if the cost of complying with unfavorable treaties is higher for party  $X$  than for party  $Y$ . An increase in  $p$  or  $q$  tends to “favor” party  $X$ , making  $T_X$  shorter relative to  $T_Y$ .

I am now almost ready to state the proposed equilibrium formally and demonstrate that it is subgame perfect. First, choose a time horizon  $T \geq T^*$ . With this time horizon  $T$  in hand, the proposed equilibrium strategy for a party  $m$ -controlled government in country 0 is: “Play  $s_m^*$  if no country has reneged on any agreement in periods  $t - T - 1, \dots, t - 1$ , or (if  $t < T$ ) if no country has ever reneged on any agreement. Otherwise, let  $i$  denote the country that has most recently reneged on an agreement, and let  $t'$  denote the period in which they reneged. If  $i = 1$  or  $2$ , play  $\{s_m^i, s_m^{j*}\}$  every period up to and including period  $t' + T$ . If  $i = 0$ , play  $s_m^0$  up to and including period  $t' + T$ . Finally, if any country  $k$  reneges during the punishment phase, or fails to enforce the punishment, terminate the punishment phase for the previous offender and restart punishment on country  $k$ .” The strategies for the governments of countries 1 and 2 are identical, except that in the punishment phase country 1, for example, will play  $\{s_m^i, s_m^{j*}\}$  when  $i = 0$  or  $2$  and will play  $\{s_m^0, s_m^2\}$  when  $i = 1$ .

**Claim 3** *These strategies form a subgame perfect equilibrium of the “international agreements game.”*

**Proof.** To show that these strategies comprise a subgame perfect equilibrium, I need to show that these strategies form a Nash equilibrium in every possible subgame. Without loss of generality, I assume that  $T_X > T_Y$ , so that  $T^* = T_X$ ; extending the proof for the case of  $T_X < T_Y$  is straightforward.

First, consider the “normal phase” when no player has recently reneged. In this case, party  $m$  in country 0 plays  $s_m^*$ . Clearly, the signing behavior is a best response to the other countries’ actions: party  $X$ , for example, benefits from type I agreements and suffers from type II agreements, so it rationally signs the former and rejects the latter.

To prove that compliance is a best response, I first examine the decision facing a government controlled by party  $X$ . Although there are still eighteen possible cases to consider (*e.g.*, party  $X$  facing two  $X$  governments and two type I agreements, party  $X$  facing one  $Y$  and one  $X$  government and one of each type of agreement, etc.) I need only look at the “most difficult” case: when a government controlled by party  $X$  faces two foreign governments controlled by party  $Y$ , and two existing type II agreements. This is the “most difficult” case because the presence of  $Y$  governments in countries 1 and 2 lowers the cost of reneging by eliminating any payoff next period from agreements signed today (which would be forfeited by noncompliance), while the existence of two unfavorable agreements raises the short-term costs of compliance. The discussion in the text demonstrates that in this case, the payoff from compliance is higher than the payoff from reneging if the following condition holds:

$$\bar{\pi}_X (1 - \delta^T) \left( \frac{\delta}{1 - \delta} \right) \geq x.$$

$T^*$  satisfies this condition by construction, and (as I showed in Remark 1) exists if assumption A1 is met. It follows that this condition holds for  $T^* \geq T$ . The corresponding “hardest case” condition for a party  $Y$  government is

$$\bar{\pi}_Y (1 - \delta^T) \left( \frac{\delta}{1 - \delta} \right) \geq y.$$

Since by assumption  $T^* \geq T = T_X > T_Y$ , this condition is also satisfied, and the party  $Y$  government also complies with unfavorable agreements (in its case, type I agreements). Moreover, if the

parties willingly comply in the “hardest case” scenarios, they will also comply in every other possible scenario, in which they have less to lose from compliance (in the case of preexisting “favorable” agreements) or more to lose from reneging (in the case of friendly foreign governments). Thus the “normal phase” strategies are Nash equilibrium strategies.

Next, consider the strategies of the “punishers” during the punishment phase. In the proposed equilibrium, these countries play normal strategies with respect to each other, but “isolate” the offender by rejecting all agreements and reneging on any existing agreements. It is straightforward to see that playing normal strategies between themselves is rational; indeed, the decision whether or not to comply with agreements with the other punishing country is identical to the compliance decision in the “normal” phase. Given that the punished country is rejecting and reneging on all agreements during the punishment phase, it is also rational for the punishing country to reject any new agreements with the offender. Furthermore, reneging on any existing treaty with the punished country is also rational: complying with the offending country yields country 0 zero today (since, as I assumed at the outset, payoffs from compliance depend on the other country’s actions), but results in  $T$  periods of punishment for country 0, which lowers its payoffs relative to the equilibrium.

Finally, consider the strategy of the “offender” during the punishment phase. Given that the other countries are “isolating” it, it gains nothing from signing agreements (which the other countries are rejecting) or from complying with agreements (which the other countries are violating). Deviating from the equilibrium strategies, however, restarts the “punishment clock,” and thus represents a reduction in payoffs relative to the equilibrium. Thus the offender’s strategy during the punishment phase is a best response to the other countries’ strategies.

In both the “normal” and “punishment” phases, therefore, the proposed strategies are best responses to each other. Therefore, the proposed strategies form a subgame perfect equilibrium as claimed. **q.e.d.** ■

Table 1: Payoffs from compliance with a Type I agreement

		<i>Foreign Country</i>	
		Comply	Reneg
<i>Home Country</i>	Comply	$b, b$	$0, b - \epsilon$
	Reneg	$-y, -y$	$-y, 0$
	$X$	$b - \epsilon, 0$	$-\epsilon, -\epsilon$
	$Y$	$0, -y$	$0, 0$

*Note:* Each cell gives the payoffs to parties  $X$  and  $Y$  in each country. The payoff to the party in the home country is listed first.