Reputations and Change in International Relations

Ekrem T. Başer*

September 1, 2023

Forthcoming in *International Studies Quarterly*

Abstract

Reputations for resolve are critical in international relations for deterring adversaries and reassuring partners. However, a state's resolve is unobservable and can change unbeknownst to its audience. How does the possibility of unobserved change impact reputation dynamics? I provide a theory of long-run reputations with changing resolve via a formal model covering conflict and cooperation domains. In the model, the possibility that current reputations are based on outdated information makes the audience extend the benefit of the doubt to states with poor reputations. This leads to states building or spending their reputations depending on their current reputations. Importantly, when damaged reputations can be rebuilt, states with better reputations face stronger temptations to spend them. Thus, reputations constrain states with poor reputations the most. Further, because demonstrations of resolve improve reputations, which, in turn, reduce incentives for future demonstrations of resolve, there is a cyclical rhythm to conflict and cooperation. A major implication is that a state's behavior changes with its reputation even if its resolve is unchanged and the stakes are identical. Reputational enforcement works, but the price is occasional breaches of trust. These results also settle a few long-standing controversies in the IR-reputation literature.

*Syracuse University, Political Science, etbaser@syr.edu. I would like to thank Muhammet Baş, Çağlayan Başer, Rob Carroll, Stephen Chaudoin, Nuole Chen, Xinyuan Dai, Brian Gaines, Chris Grady, Alice Iannantuoni, Brenton Kenkel, Korhan Koçak, Jim Kuklinski, Bob Pahre, Charla Waeiss, Matt Winters, seminar participants at University of Illinois, Washington University St. Louis, NYU Abu Dhabi, EITM Institute, MPSA, SPSA, and Formal Models of International Relations Conference for helpful comments and discussions. I also would like to thank the Institute for Humane Studies for supporting this research through the Humane Studies Fellowship.

1 Introduction

A state's reputation refers to the audience's beliefs about its persistent characteristics informed by past behavior (Dafoe, Renshon, and Huth 2014, 374). I focus on "reputations for resolve," defining resolve as "steadfastness of purpose, maintaining a policy despite contrary inclinations or temptations to back down (Kertzer 2016, 8)." This definition highlights the common logic operating across conflictual and cooperative interactions, and addresses calls for greater integration of these domains regarding resolve and reputation mechanisms (Kertzer 2016; Jervis, Yarhi-Milo, and Casler 2021). Resolve is central from refusing to acquiesce to threats despite the risk of conflict (McManus 2017), to fulfilling debts notwithstanding the temptation to default (Tomz 2007), to protecting foreign investors' property rights regardless of the rents from expropriation (Johns 2021).

At the core of reputations lies a repeated information problem. Audiences (adversaries, creditors, allies) are uncertain about a state's resolve. They try to diminish their uncertainty by learning from the state's past behavior and forming beliefs about its resolve, which constitute the state's reputation. A good reputation entails significant benefits by deterring adversaries and reassuring partners (Crescenzi 2018; Harvey and Mitton 2017; Weisiger and Yarhi-Milo 2015). Seeking these benefits, otherwise irresolute states want to appear resolute to their audiences. Thereby reputations constrain state behavior. In the absence of strong contract-enforcing institutions, that is, under anarchy, reputations are considered a decentralized "enforcement" mechanism due to this constraining power (Keohane 1984; Simmons 2010).

I argue that two features of resolve pose challenges to reputations. First, resolve is not fixed but changes over time. Whether a state is resolute or irresolute depends on inputs both situational (fighting costs, need for credit, threat environment) and dispositional (leader personalities, culture) (Kertzer 2016). Resolve thus changes following moving parts in governments, societies, and environments, which is consequential for reputations. Recent research finds that changes in many sources of resolve, situational and dispositional, affects reputation dynamics. Examples include leaders' cultural backgrounds (Dafoe and Caughey 2016) and personality (Yarhi-Milo 2018), threat environment (Tingley and Walter 2011; Sechser 2018), and broader contextual roots of resolve (Crescenzi 2018; Harvey and Mitton 2017). While this literature unpacks when reputations

matter, it also highlights the challenge of changing resolve for reputations: if resolve changes, uncertainty surrounding it keeps replenished.

The second challenging feature of resolve is that resolve is an unobservable which only imperfectly correlates with its observable inputs (Kertzer 2016). An observed change, like leader turnover, does not guarantee a change in resolve. If observed changes perfectly revealed a state's resolve, the uncertainty that drives reputations would not exist. Conversely, the absence of observed changes does not guarantee unchanged resolve. The fact that resolve can change without the audience's knowledge poses a challenge because reputations rely on the relevance of past information in the present. Reputations require a state's resolve to be *sticky*.

I argue that a theory of reputations in IR should consider resolve as both sticky and changing, acknowledging the *possibility* that audience beliefs may be based on outdated information, even in the absence of *observed* or *realized* changes. Reputation is a long-run, repeated information mechanism where audiences face persistent and shifting uncertainty about resolve. This necessitates a dynamic theory that explains the consequences of shifting uncertainty about resolve for state behavior. How do reputation dynamics unfold under shifting uncertainty? What incentives do states face in building and spending their reputations if audiences suspect that today might differ from yesterday? How does a state's current reputation affect its behavior?

Reputation research has seen a resurgence over the past decade, yet, our understanding of long-run reputations, which require grappling with shifting uncertainty, remains limited. Recent research has established the importance of reputations in IR (Crescenzi 2018; Harvey and Mitton 2017; Weisiger and Yarhi-Milo 2015), charted scope conditions for when reputations matter (Clare and Danilovic 2010; Dafoe and Caughey 2016; Sechser 2018; Yarhi-Milo 2018), improved our understanding on the interaction of leader and state reputations (Renshon, Dafoe, and Huth 2018; Lupton 2020; Yarhi-Milo 2018; Wu and Wolford 2018), and on the microfoundations of reputational concerns (Brutger and Kertzer 2018; Kertzer, Renshon, and Yarhi-Milo 2021).[1] We nevertheless lack a dynamic theory of long-run reputations under change because existing theories in IR tend to ignore that resolve changes (Nalebuff 1991; Schelling 1966; Simmons 2000; Walter 2006) or that resolve is sticky (Sartori 2002; Wolford 2007).

If a state's resolve is fixed, it finds reputation-building worthwhile and never falters, or it

1. See Jervis, Yarhi-Milo, and Casler (2021) for a survey.

never bothers. There is no reputation spending or rebuilding; the behavior is static (Kreps and Wilson 1982). If a state's resolve is not sticky, observing a leadership change or a different context, audiences believe lessons from the past are no longer relevant. If reputations emerge, they are short-lived, relevant only until the next observed change. Then, one cannot explain long-run reputations. The static picture still prevails, only to be punctuated by periods of observed change.

The few works which consider both changing and sticky resolve, in turn, either focus on short-run dynamics, ignoring the long-run (Alt, Calvert, and Humes 1988; Wu, Licht, and Wolford 2021), or on audience reactions at the expense of state behavior (Lupton 2020; Tomz 2007). They overlook the core challenge of change for a dynamic reputation theory: the implications for how a state's reputation conditions its behavior.[2] Therefore, our understanding of long-run reputation dynamics in IR remains limited.

This paper presents a dynamic theory of long-run reputations in IR using a formal model that incorporates the sticky yet changeable nature of a state's resolve. The model encompasses interactions in both cooperation (sovereign borrowing, alliances) and conflict (deterrence, sanctions) domains. States build reputations through costly actions, such as resisting threats or honoring debts, while reputational spending occurs when states yield to short-term temptations like acquiescing to threats or defaulting on debts. Audience members possess uncertainty about state resolve and acknowledge the potential for unobserved changes. This shifting uncertainty poses challenges for reputation-building as the audience remains somewhat skeptical of positive reputations. However, it also allows for reputation rebuilding, as the audience extends some benefit of the doubt to states with poor reputations.

The model produces important insights. First, I provide scope conditions regarding the stability in resolve for long-run reputations to emerge. Influential works argued that changing resolve can render reputations irrelevant (Mercer 1996; Snyder and Diesing 1977; Press 2005). While recent scholarship established the relevance of reputations, we know little about the extent of change that can sustain reputations. I demonstrate that reputations arise unless resolve changes excessively fast, rendering past information irrelevant. Whether change is excessive depends on the importance of interacting with the state for its audience relative to their status quo payoffs. Very good or poor status quo payoffs require greater stability in resolve to sustain reputations.

2. An exception is Gent et al. (2015), which I discuss below.

Second, I discuss how states' current reputations determine their future behavior. When reputations are poor, states who would otherwise yield to short-run temptations—defaulting, acquiescing, not following through on threats—resist them in expectation of future reputational returns. When reputations are better, they are more likely to yield to short-run temptations and spend their reputations. Better reputations leave less room for improvement, implying diminishing returns to reputation building. In contrast, better reputations generate greater temptation to spend them because taking advantage of others' beliefs is more profitable when they are favorable. Further, due to shifting uncertainty, the audience is suspicious of good reputations and extends the benefit of the doubt to those with poor reputations. Then, forward-looking states know that damaged reputations can be rebuilt and better rewards will be forthcoming. When resolve is sticky yet changing, states with better reputations face greater incentives to spend their reputations today and start rebuilding tomorrow.

Therefore, there is a cyclical rhythm to conflict and cooperation. A state's past behavior affects its reputation, which affects current behavior *even when stakes are identical, and its resolve remains unchanged*. Cooperation scholars commonly focus on how cooperation breaks down due to changes to compliance costs (Koremenos 2005; Rosendorff and Milner 2001; Tomz 2007). I show that noncompliance can directly result from reputation dynamics, even when compliance costs remain identical. Conflict scholars commonly concentrate on changes in the costs of war, capabilities, and issue stakes to explain variation in state behavior (Fearon 1995; Powell 2006). I show that demonstrations of resolve in conflictual interactions can reduce incentives for future resolute behavior and vice versa, even with otherwise identical stakes.

This logic is reflected in Walt Rostow's advice to Kennedy in a 1961 memorandum entitled "The Shape of Battle." According to Rostow, if the U.S. proved its mettle by prevailing in Vietnam, this would enable the pursuit of conciliatory policies toward the Soviets and China (Gibbons 2014, 25). Whereas a key reason why the U.S. chose to stand firm in Vietnam in the first place was the need to demonstrate after the Bay of Pigs fiasco that it was not a "paper tiger" (23–24).

Third, I settle the controversy regarding the relationship between current reputations and current behavior among IR scholars. An influential argument known as the "domino theory paradox" contends that while reputational deterrence theories expect audiences to believe an actor with a blemished reputation to back down, policymakers can try extra hard to contradict this expecta-

tion (Jervis 1997). Several others criticized reputational theories as paradox-ridden via analogous arguments (Mercer 2013; Press 2005). Others disagree, claiming these paradoxes themselves to be illogical (Sartori 2002).

I resolve this controversy by showing that such "paradoxes" are a logical consequence of reputations under shifting uncertainty, something existing reputational theories cannot address. Additionally, I discuss the sources of confusion stemming from disequilibrium reasoning by reputation critics. While irresolute states with better reputations are likely to spend their reputations, audiences remain uncertain about states' future behavior. In fact, for the audience, better reputations are associated with a weakly *increasing* expected probability of reputation-building, even though better reputations *decrease* states' incentives to build reputations.

Fourth, I discuss how increased uncertainty on state resolve following *observed* changes like leader turnover affect behavior. Some studies find that increased uncertainty leads to a greater willingness to demonstrate resolve (Lupton 2020; Wolford 2007; Wu, Licht, and Wolford 2021), others find the opposite (Thyne 2012). I argue that, depending on prior reputations, observed changes can increase or decrease incentives to build reputations. The effect of observed changes on behavior is non-monotonic in prior reputations.

Finally, I whether reputations can act as a decentralized enforcement mechanism. I show that, even when reputational enforcement functions well, it is still an imperfect substitute for contract enforcement. Reputational enforcement works, but the price is occasional breaches of trust.

This paper is related to several others modeling reputations under type replacements (Alt, Calvert, and Humes 1988; Mailath and Samuelson 2001; Gent et al. 2015; Phelan 2006). Alt, Calvert, and Humes (1988) model a hegemon with changing coercion costs responding to an ally's challenge. Their two-period model restricts reputational concerns to the first period, precluding an analysis of long-run reputations where endogenous reputations condition future behavior, which is the focus here. In Mailath and Samuelson (2001) and Gent et al. (2015), agents (firms and NGOs, respectively) fear being seen as inept, thus continually build reputations because tarnishing reputations is catastrophic. Mailath and Samuelson (2001) prevent rebuilding by assuming that change is always bad, and Gent et al. (2015) by assuming that a no-cooperation outcome ends the NGO. Whereas here, a tarnished reputation is costly but not game-ending, and states can rebuild reputations because resolve can decrease but also increase.

The closest work to this paper is Phelan (2006), which looks at government taxation with type replacements. The game in Phelan (2006) is similar to my cooperation environment, and I follow his equilibrium characterization. Beyond adapting Phelan (2006)'s model to IR, I add to his results in three respects. One, while in Phelan (2006) reputation-building is costless without the audience's trust, here reputation-building is always costly. Interestingly, unlike in Phelan (2006), the equilibrium here requires forward-looking states recovering the classic result in Kreps and Wilson (1982). Two, Phelan (2006) considers none of the implications I discuss here (Propositions 2-5). Three, neither Phelan (2006) nor the others above analyze together contexts where reputations are pursued to deter adversaries (conflict), and to reassure partners (cooperation), which I do.[3] I show that a single mechanism operates across both domains.

2 Model

A state L (long-lived) interacts with a different audience member in each period for infinitely many periods. Generically referred to as R (short-lived), each audience member enters the game for a single period and is replaced by a new R in the next period. The entire audience observes interactions. L's discount factor is $\delta \in (0,1)$. Following much of the literature, the main actor facing reputational concerns vis-a-vis an audience of third parties is a state (Crescenzi 2018; Schelling 1966; Sartori 2002; Tomz 2007; Weisiger and Yarhi-Milo 2015). Focusing on state reputations is useful for examining long-run dynamics. It can be justified by path dependence due to bureaucratic structures, formal and informal social institutions, among others. Whether states build reputations to deter adversaries or reassure partners, the audience could be potential challengers, allies, international lenders, foreign investors, or rebel groups.

2.1 Stage Game

In a given period, L can be a behavioral resolute type or a strategic irresolute type, which is its private information. First, R decides whether to enter into an interaction with L or stay out, $a_R \in \{E, O\}$. Then, if L is an irresolute type, it chooses between a costly strong action or a cheap weak

3. In cooperation, costly signals are Pareto improving. In conflict, costly signals benefit the sender and hurts the audience. Relabeling one to get the other is impossible.

action $a_L \in \{S, W\}$. If L is resolute, it plays the strong action by assumption. Since the resolute L is a behavioral type, the action sets are defined only for the irresolute L. I assume the following about the stage-game payoffs of R and the irresolute L, $u_i(a_R, a_L)$ for $i \in \{R, L\}$.

Assumption 1 (*Strong action is costly*)**.** $u_L(a_R, W) - u_L(a_R, S) > 0$ for all a_R

Assumption 2 (*Independent status quo payoff*)**.** $u_R(O, a_L) = d$ for all a_L

If R stays out, $a_R = O$, it receives its status quo payoff d, independent of L's actions.

Assumption 3. Payoffs are described by either the cooperation or the conflict environments:

A. *Cooperation:*

i. $u_L(E, a_L) - u_L(O, a_L) > u_L(a_R, W) - u_L(a_R, S)$ for all a_L, a_R

ii. $u_L(E, W) - u_L(E, S) \geq u_L(O, W) - u_L(O, S)$

iii. $u_R(E, S) > d > u_R(E, W)$

B. *Conflict:*

i. $u_L(O, a_L) - u_L(E, a_L) > u_L(a_R, W) - u_L(a_R, S)$ for all a_L, a_R

ii. $u_L(O, W) - u_L(O, S) \geq u_L(E, W) - u_L(E, S)$

iii. $u_R(E, W) > d > u_R(E, S)$

In cooperation, (i) indicates that the irresolute L's gain from having R enter is always greater than the cost of the strong action. Otherwise irresolute L would never build its reputation in the repeated game, as L's gains from a good reputation (having R enter) would not be worth the trouble (the cost of strong action). Point (ii) indicates that R entering (weakly) increases the irresolute L's temptation of choosing W instead of S; and (iii) indicates that R prefers to enter if L plays S but not otherwise.

In conflict, (i) indicates that the irresolute L's gain from having R stay out is always greater than the cost of the strong action; (ii) indicates that R staying out (weakly) increases the irresolute L's temptation of choosing W; and (iii) indicates that R prefers to enter if L plays W but not otherwise.

2.1.1 Discussion on the Stage Game

Figure 1 displays examples of substantive contexts captured by each environment.[4] In both panels, R's information sets include scenarios where L is resolute and irresolute.[5]

For cooperation, an example is sovereign borrowing (Figure 1a). R enters by lending money and stays out by not lending money. If R lends money, a costly strong action by L settles its debts, and a weak action defaults. If R stays out, a strong action demonstrates fiscal discipline, and a weak action is populist spending. The irresolute L is tempted to take the weak action: defaulting is more tempting than honoring debts, and populist spending is more tempting than pursuing fiscal discipline.

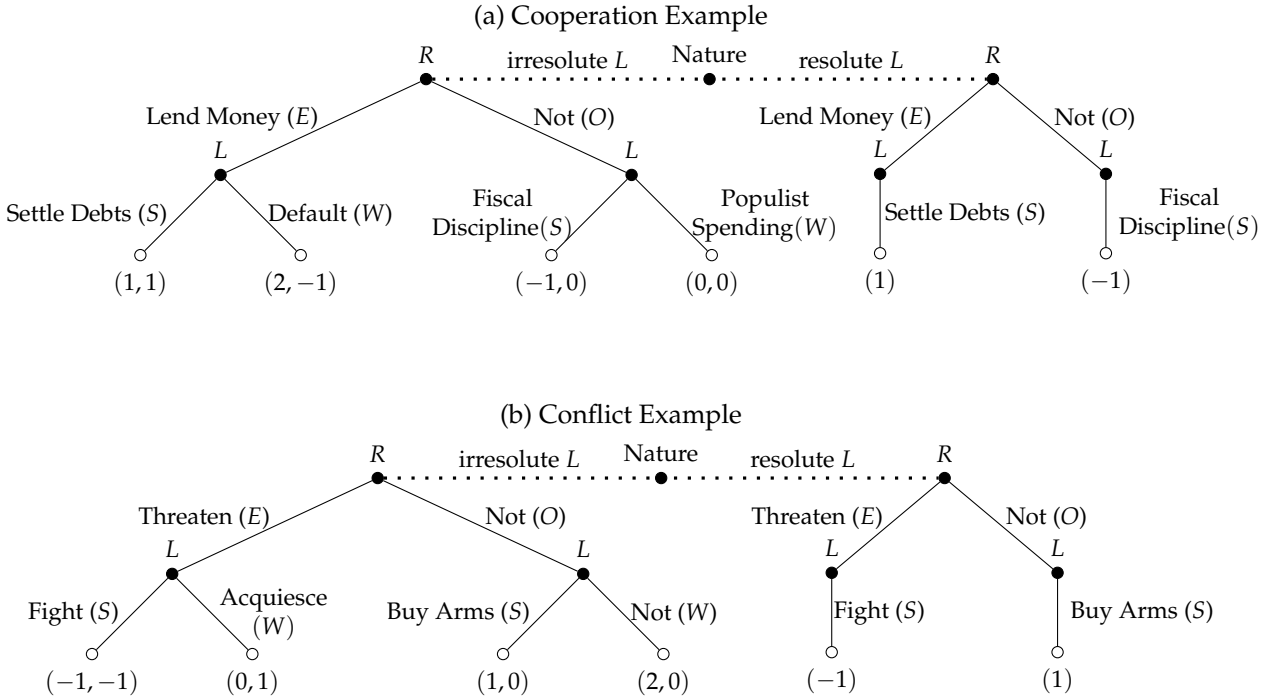

Figure 1: Panel (a) is a cooperation example capturing sovereign borrowing dynamics. Panel (b) is a conflict example capturing deterrence dynamics. R's payoffs are on the left, and irresolute L's payoffs are on the right. When L is resolute, the reported payoffs pertain to R.

If R does not lend money, it receives its status quo payoff. If R lends money, it prefers L to settle its debts than default. R prefers the outcome where L settles its debt to the status quo, but prefers the status quo to the outcome where L defaults.

4. The stage game is fixed, pertaining to either conflict or cooperation environments.

5. Abusing game tree representation, I capture the resolute L scenario by L having one choice (S). Strictly speaking, the resolute L is not making a choice, as it is a behavioral type.

For conflict, an example is deterrence (Figure 1b). R enters by challenging L, threatening military action in case of non-compliance, and R stays out by not issuing threats. If R threatens, a costly strong action by L is fighting or escalating, and a weak action is acquiescing. If R stays out, a strong action is investing in arms, and a weak action is not investing in arms. The strong action is costlier for the irresolute L: fighting and investing in arms is costlier.

If R does not threaten, it receives its status quo payoff. If R threatens L, it prefers L to acquiesce than fight. R prefers the outcome where L acquiesces to the status quo, but prefers the status quo to the outcome where L fights.

When R knows that L is irresolute, the cooperation stage game has a unique subgame-perfect Nash equilibrium (O, W): R does not lend money, and the irresolute L engages in populist spending. Herein lies L's problem: if it could commit to honoring debts, R would lend, and the irresolute L would be better off. Unfortunately, R knows that the irresolute L would default after borrowing and refuses to lend. R prefers lending and have L honor its debt to not lending, but this cannot happen in equilibrium due to irresolute L's temptation to default. The irresolute L's problem hurts both sides, generating a Pareto-inferior outcome.

In conflict, the irresolute L in faces a similar problem without reputational incentives; when R knows that L is irresolute. In equilibrium, R threatens L, and L acquiesces (E, W). If L could commit to fighting after threatened, R would stay out, making the irresolute L better off. However, R knows that the irresolute L would acquiesce, and threatens. Unlike cooperation, R benefits from L's irresoluteness in conflict because otherwise R would have to stay out to avoid fighting and be worse off.

Cooperation environment captures settings where mutually beneficial cooperation is hindered by temptations to exploit others' trust, as in the sovereign debt, foreign direct investment, or alliance literature (Crescenzi 2018; Johns 2021; Tomz 2007; Leeds 2003). The temptation to default after borrowing money prevents borrowing in the first place (Tomz 2007). The temptation to expropriate foreign investments once they are in—the "obsolescing bargain"—scares away investors (Johns 2021). The temptation to refuse helping allies precludes alliance formation or renders alliances ineffective (Leeds 2003).

Conflict environment captures settings where the problem is deterring challengers due to temptations to acquiesce, such as generalized deterrence, where challenges involve threats of vi-

olence and challengers are typically other states (Schelling 1966; Nalebuff 1991). It can also speak to civil war and repression dynamics where challengers are not states but rebels (Walter 2006) or anti-government groups (Licht and Allen 2018). Deterring sanctions is similar, where demands are accompanied not by violence but economic punishment or withdrawal of cooperation (Cilizoglu and Bapat 2020).

Finally, I allow L to take strong or weak actions without explicit challenges or gestures of cooperation by R. Therefore, L can signal resolve in the repeated game when R stays out. This allows ignoring what "no information" means, providing significant tractability gains. Further, as recent research suggests, this better captures dynamics of signaling resolve (Goldfien, Joseph, and McManus 2023). A state can signal resolve to challengers by fighting against threats, but also via military investments in the absence of threats. It can signal resolve to foreign investors by protecting their investments once they are in place but also by strengthening judicial institutions in the absence of investments.

When Russia invaded Ukraine in 2022, observers assessed NATO members' resolve as reliable allies both via their actions during the invasion (supplying arms to Ukraine) and also via their prior performance in fulfilling alliance duties in the absence of an explicit challenge (whether military spending fulfills promises). Similarly, when assessing states' resolve in honoring debts, international lenders not only consider prior debt repayments but also states' economic policies indicative of good debtor behavior, such as increasing taxes and cutting public spending (Tomz 2007).

Even in the classic chain-store setup, assuming that challengers learn nothing about the monopolist without entry is unrealistic. The limit-pricing literature suggests that monopolists signal their ability to deter entrants via the prices they set even without any realized competition (Milgrom and Roberts 1982). Nevertheless, I discuss in the appendix what equilibrium dynamics might look like if L cannot signal resolve when R stays out.

2.2 Repeated Game, Information, Strategies

The stage game is played between L and a new R in each period, for infinitely many periods. R does not know L's resolve but has beliefs about it. These beliefs constitute L's reputation:

Definition 1 (*Reputation*)**.** The probability R assigns to L being resolute at the beginning of time t, μ_t, is L's *reputation*.

I model resolve as sticky yet subject to change without the audience's knowledge by allowing L's resolve to change at the beginning of each period following a Markov process. R knows that L's type might change but does not observe whether the change occurs. If L is resolute in period t, it becomes irresolute in $t+1$ with $1-\lambda$ probability, and remains resolute with λ probability. If L is irresolute in period t, it becomes resolute in $t+1$ with ϵ probability, and remains irresolute with $1-\epsilon$ probability. I set L's initial reputation at $\mu_0 = \epsilon$. Transition probabilities are common knowledge. The following assumption ensures that resolve is sticky:

Assumption 4. *(Resolve is sticky)*

- $\epsilon < \frac{d - u_R(E,W)}{u_R(E,S) - u_R(E,W)}$
- $\lambda > \frac{d - u_R(E,W)}{u_R(E,S) - u_R(E,W)}$

Assumption 4 ensures sufficient stability over types to sustain reputations. If the first inequality is violated, irresolute types become resolute so frequently that even if the audience knows that L is irresolute today, they will think that L is likely resolute tomorrow. The audience extends too much benefit of the doubt to states with damaged reputations to make reputation-building worthwhile. If the second inequality is violated, resolute types become irresolute so frequently that even if the audience knows that L is resolute today, they will think that L is likely irresolute tomorrow. Then the audience is too suspicious of good reputations to make reputation-building worthwhile.

I consider Markovian strategies with L's reputation as the state variable. Beside tractability gains, there are substantive reasons for this restriction. It enables a sharp focus on reputations: how L's reputation, formed via R's learning based on past observations, conditions everyone's behavior. This restriction eliminates strategies based on norms of reciprocity and retaliation, which are not of interest here. Further, scholars questioned whether the significant cognitive sophistication signaling games demand was realistic for real-world policymakers (Downs and Jones 2002; Jervis 1976; Mercer 1996). The low cognitive load Markovian strategies entail helps alleviate these concerns.

A Markovian strategy takes L's reputation μ at the beginning of that period and returns a probability of S for the irresolute L and E for R; denoted σ_L and σ_R respectively. The resolute L always plays S. Each R chooses σ_R to maximize its expected payoff for the period it plays the game:

$$\sigma_R \underbrace{\Big(\mu u_R(E,S) + (1-\mu)(\sigma_L u_R(E,S)) + (1-\sigma_L)u_R(E,W))\Big)}_{\text{expected payoff from entering}} + (1-\sigma_R) \underbrace{d}_{\text{status quo payoff}} \tag{1}$$

Having observed L's action, R updates its belief about L's resolve, μ_t, according to Bayes' Rule. R adjusts this posterior according to the type transition probabilities to arrive at L's reputation at the beginning of the next period, μ_{t+1}. This process is captured by $\Phi(\mu|a_L)$. If R observes W, it infers that L is irresolute since the resolute L never plays W. Then, L's reputation at the beginning of the next period is the probability that yesterday's irresolute L becomes resolute today: $\Phi(\mu|W) = \epsilon$. If R observes S, then L's reputation in the next period is:

$$\Phi(\mu|S) = \underbrace{\lambda\left(\frac{\mu}{\mu+(1-\mu)\sigma_L}\right)}_{L \text{ was resolute yesterday, stays resolute today}} + \underbrace{\epsilon\left(1-\frac{\mu}{\mu+(1-\mu)\sigma_L}\right)}_{L \text{ was irresolute yesterday, becomes resolute today}} \tag{2}$$

$\Phi(\mu|S)$ is decreasing in σ_L: the more R expects S from the irresolute L, the less R learns about L's resolve. If R knows that L is resolute at the end of t, L's reputation at $t+1$ will be $\mu_{t+1} = \lambda$. Hence λ represents the upper bound and ϵ the lower bound for L's reputation at the beginning of any period.

L's choices influence the future via its reputation, therefore, the irresolute L maximizes its flow payoffs. Let $V(\mu)$ denote the continuation value of the game for the irresolute L when its reputation is μ.

$$V(\mu) = \sigma_L \underbrace{\Big(\sigma_R u_L(E,S) + (1-\sigma_R)u_L(O,S) + \delta V(\Phi(\mu|S))\Big)}_{\text{continuation payoff after playing S, and starting the next period with reputation } \Phi(\mu|S)} + (1-\sigma_L) \underbrace{\Big(\sigma_R u_L(E,W) + (1-\sigma_R)u_L(O,W) + \delta V(\epsilon)\Big)}_{\text{continuation payoff after playing W, and starting the next period with reputation } \epsilon} \tag{3}$$

The irresolute L chooses σ_L to maximize expression 3. The pair (σ_L, σ_R) is a Markov Perfect Equilibrium (MPE), which I henceforth refer to as equilibrium.

3 Results

3.1 No uncertainty, no reputations

Suppose resolve is fixed, and R knows that L is irresolute ($\mu_0 = \epsilon = 0$). L has no incentive to play S because it will not improve its reputation—the audience has nothing to learn. Strong acts do not matter for L's future outcomes. Given L's behavior, R stays out in cooperation and enters in conflict. Without reputations, there is no cooperation (no (E, S) outcome), and while there is no costly conflict—L always acquiesces—L bares the brunt of anarchy by facing constant predation. This scenario reiterates that uncertainty about resolve is necessary for reputational incentives to emerge, without which, the equilibrium is tragic for both players in cooperation, and for only L in conflict.

3.2 Without changing resolve, everything is static

Suppose resolve is fixed, and R is uncertain about L's resolve ($\mu_0 > 0$, $\epsilon = 0$, $\lambda = 1$). If the irresolute L plays W, it reveals itself irresolute, uncertainty disappears, and the equilibrium reverts to the no-uncertainty case above. Recognizing this grim prospect, a forward-looking irresolute L decides playing S is worthwhile notwithstanding the costs and never falters. R enters in cooperation and stays out in conflict. Thereby, introduction of reputational incentives via uncertainty solves the irresolute L's problem.

The amount of cooperation here is the same as if L commits ex-ante to playing S via an enforceable contract. Reputational enforcement achieves first-best outcomes in cooperation. In conflict, there is no conflict, like the no-uncertainty case. However, L is better off and R worse off, because L playing S due to reputational concerns deters challengers.[6] Since reputations solve irresolute L's problem, making it better-off in both environments, overlooking changing resolve leads to overemphasizing the costs of reputational damage.

6. I do not present a proof, as this is the classic result from Kreps and Wilson (1982).

Here, L's reputation never changes because the irresolute L perfectly mimics the resolute type, and the audience learns nothing about L's resolve. Reputations without change are not dynamic; they are static with no reputation spending or rebuilding. While the common fixed-resolve assumption is useful to connect reputational incentives to certain outcomes (war, cooperation), it is ill-suited to explain dynamics of reputations.

3.3 Reputations with changing resolve

Consider the main scenario where the audience is uncertain about L's resolve, and acknowledges that it can change without their knowledge ($\mu_0 = \epsilon, \epsilon > 0, \lambda < 1$).

By expression 1, when L's reputation μ is high enough, R's expected payoff from entering exceeds its status quo payoff even if R expects the irresolute L to play W with certainty. Let μ^* be the threshold reputation level which makes R indifferent when $\sigma_L = 0$:

$$\mu^* = \frac{d - u_R(E, W)}{u_R(E, S) - u_R(E, W)} \tag{4}$$

This expression gives the most R can secure by staying out as a proportion of the most it can secure by entering: μ^* measures the importance of interacting with L.

When L's reputation is above this threshold $\mu > \mu^*$, R enters in cooperation and stays out in conflict. Suppose for $\mu > \mu^*$ the irresolute L plays W, $\hat{\sigma}_L(\mu) = 0$. I discuss below why this is true in equilibrium. If L's reputation is lower ($\mu \leq \mu^*$), the irresolute L can compensate by taking strong actions with greater probability to induce R to enter in cooperation and stay out in conflict. Specifically, the irresolute L will choose σ_L to make R indifferent between its expected returns to entering and staying out in expression 1:

$$\hat{\sigma}_L(\mu) = \frac{\mu^* - \mu}{1 - \mu} \tag{5}$$

$\hat{\sigma}_L(\mu)$ is strictly decreasing in μ. When $\mu \leq \mu^*$, the irresolute L is less willing to play S the better its reputation. This specifies the irresolute L's equilibrium strategy $\hat{\sigma}_L(\mu)$.

The audience observes L's action and updates its reputation by taking into account irresolute

L's strategy $\hat{\sigma}_L(\mu)$. If they observe S, the learning process in expression 2 becomes:

$$\hat{\Phi}(\mu|S) = \left(\frac{\lambda - \epsilon}{\mu^*}\right)\mu + \epsilon \tag{6}$$

If they observe W, then L's reputation is $\hat{\Phi}(\mu|S) = \epsilon$.

R's equilibrium strategy when L's reputation is $\mu > \mu^*$ is to enter in cooperation ($\hat{\sigma}_R = 1$) and stay out in conflict ($\hat{\sigma}_R = 0$) with certainty. When L's reputation is lower $\mu \leq \mu^*$, R chooses its probability of entry to make the irresolute L indifferent between strong and weak actions. The irresolute L chooses S only if the returns on its reputational investment are worthwhile. The better L's reputation, R must enter in cooperation (or stay out in conflict) with higher probability to keep the irresolute L indifferent. Therefore, σ_R is strictly increasing in L's reputation in cooperation and strictly decreasing in conflict. This increased willingness to do what L wants as L's reputation increases provides the necessary rewards for the irresolute L to pay reputation-building costs.

Why is it optimal for the irresolute L to choose W when $\mu > \mu^*$? For the irresolute L to choose S, the rewards should be worthwhile. However, there is an upper limit to these rewards. Once L's reputation is $\mu > \mu^*$, R is sufficiently confident of L's resolve that it is willing to accept the residual risk of being wrong. Beyond this point, further reputational improvements bring L no additional rewards. Not only that, but L's temptation to spend its reputation is highest when $\mu > \mu^*$.

If L's type was fixed, damaging its reputation would mark it irresolute forever. Equilibrium would revert to the no-uncertainty case, and this grim prospect would suppress the irresolute L's temptations. Here, such a threat is not credible because L's resolve could improve. Hence $\Phi(\mu|W) = \epsilon > 0$ is the lower bound for L's reputation at the beginning of any period. R has reason to think that its beliefs might be due to outdated circumstances. Since interacting with L is potentially profitable, R is willing to give it a try, if slowly. A sufficiently forward-looking irresolute L ($\delta > \bar{\delta}$) strictly prefers to take advantage of its reputation when $\mu > \mu^*$ instead of paying reputation-building costs indefinitely. It knows that its reputation will improve and better rewards will be forthcoming.

The upper limit to reputation-building rewards is also the reason why L is less willing to play S the better its reputation when $\mu \leq \mu^*$. The upper limit to reputation-building rewards implies diminishing returns to further reputation building at all reputation levels. The reason is

straightforward: the irresolute L needs further rewards to build its reputation but R's ability to provide them is diminishing. If at reputation level μ, R is willing to enter in cooperation with $\sigma_R(\mu)$ probability, the size of the future rewards are capped at $1 - \sigma_R(\mu)$. The more the audience members are willing to do what L wants, the less room for improvement, implying diminishing returns to reputation building. To summarize:

Proposition 1. *Maintain Assumptions 1-4 and $\delta > \bar{\delta}$. In equilibrium, if $\mu \leq \mu^*$, the irresolute L plays S with strictly lower probability the better its reputation $\left(\hat{\sigma}_L(\mu) = \frac{\mu^* - \mu}{1 - \mu}\right)$. If $\mu > \mu^*$ the irresolute L plays W with certainty ($\hat{\sigma}_L(\mu) = 0$). The resolute L always plays S by assumption.*

If $\mu \leq \mu^*$, R enters in cooperation and stays out in conflict with strictly greater probability the better L's reputation. If $\mu > \mu^*$, R enters in cooperation and stays out in conflict with certainty.

The proof, including the expressions for σ_R and $\bar{\delta}$, is in the appendix. I leave as conjecture that this is the unique MPE of the game and include further comments in the appendix about the uniqueness of the equilibrium. Next, I discuss several major implications of this equilibrium.

4 Discussion

4.1 Long-run reputations emerge with changing resolve, but not if change is too fast

Several influential works argued that unobserved change in resolve might render reputations irrelevant (Mercer 1996; Snyder and Diesing 1977; Press 2005). According to Snyder and Diesing (1977, 186–187):

> With the passage of time and shifts in the climate of public opinion, a nation's behavior in a past crisis becomes an increasingly unreliable indicator of future behavior... Since crises tend to be diverse in structure, background, emotional content, and so on, predictions of a state's behavior from one case to another would seem unreliable.

To Mercer (1996, 38–39), failing to address unobserved change means not addressing long-run reputations:

> Game theorists have... yet to address general reputations, for two reasons. First, if we define resolve as a function of things that vary according to the situation, then resolve cannot have cross-situational validity... Second, formal work has not addressed general reputations because those models assume... the situation is always similar... By defining a reputation as having validity only in the same situation, formal approaches so truncate the concept that it misses most of what we are interested in.

Although recent research provided significant evidence for the relevance of reputations (e.g., Crescenzi 2018; Weisiger and Yarhi-Milo 2015), the problem of changing resolve has not been addressed by our theories. Formal models in IR with changing resolve do not tackle this problem because they do not consider resolve to be sticky (Wolford 2007; Sartori 2002). If today's resolve is independent of yesterday's, this assumes long-run reputations away because audiences cannot learn across situations. In Wolford (2007), reputations are short-run, resetting with each leadership change. In Sartori (2002), resolve is independent across situations, and reputations for resolve do not emerge. Instead, actors get a reputation for honesty or dishonesty that transcends time.[7]

Wu, Licht, and Wolford (2021) and Gent et al. (2015) are exceptions. In Wu, Licht, and Wolford (2021) new leaders inherit part of their resolve from predecessors. They analyze how the inherited resolve affects the risk of conflict by *assuming* that long-run reputations exist rather than problematizing their emergence. In Gent et al. (2015), an NGO's competence changes with its leadership but it is sticky. They assume that a no-cooperation outcome is a game-ending event, in so doing, prevent change to be reflected in reputation dynamics and retain the static picture.

In my model, a state's resolve is changing yet sticky. The audience acknowledges that resolve can change without their knowledge. This shifting uncertainty makes the audience suspicious of good reputations and willing to extend the benefit of the doubt to those with poor reputations. Proposition 1 demonstrates that long-run reputations emerge and constrain state behavior even under such shifting uncertainty, suggesting that reputation critics maybe too pessimistic.

That said, reputation critics are not wrong because change can too dramatic for the audience's learning to keep up, rendering long-run reputations irrelevant. Assumption 4 captures the precondition regarding the pace of change to sustain reputations:

$$\epsilon < \frac{d - u_R(E, W)}{u_R(E, S) - u_R(E, W)} = \mu^* < \lambda$$

Volatility in states' resolve can unravel reputations via two channels. If λ is small, the audience thinks yesterday's resolute states are unlikely to stay resolute. This extreme suspicion means

7. Reputations in Sartori (2002) are different. Here reputation is a learning mechanism based on private information. Reputations in Sartori (2002) capture a retaliatory enforcement mechanism. The model in Sartori (2002) demonstrates that states can make commitments via truth-telling norms with retaliatory enforcement, even when reputations—as conceptualized here—cannot emerge because resolve is not sticky (129, 136–137). See Tomz (2007) for the difference between retaliatory and reputational enforcement.

good reputations decay too quickly to make reputation-building worthwhile. If ϵ is large, the audience thinks yesterday's irresolute states will likely become resolute. Such excessive benefit-of-the-doubt makes taking advantage of this free goodwill preferable to costly reputation-building.

Whether L's resolve is sufficiently sticky depends on μ^*; the importance of interacting with L relative to R's status quo payoff d. A very high or low d destroys reputational incentives. A high d makes R sensitive to the risk of facing an irresolute L in cooperation (resolute L in conflict). This unravels reputations via excessive suspicion. A low d makes R insensitive to the risk of facing an irresolute L in cooperation (resolute L in conflict), unraveling reputations via the excessive benefit-of-the-doubt channel. To summarize:

Proposition 2. *Fix a process of change* $(\epsilon > 0, \lambda > 0)$. *There exists a lower bound* $\left(\underline{d} = \frac{\epsilon(u_R(E,S)-u_R(E,W))}{u_R(E,W)}\right)$ *and an upper bound* $\left(\bar{d} = \frac{\lambda(u_R(E,S)-u_R(E,W))}{u_R(E,W)}\right)$ *such that the equilibrium in Proposition 1 is sustained if* $d \in (\underline{d}, \bar{d})$.

This result follows directly from Proposition 1 and Assumption 4.

4.2 A state's behavior changes with its reputation even if its resolve is unchanged

A critical feature of the equilibrium is that a state's current reputation determines its behavior *even when its resolve is unchanged and stakes are identical*. A resolute state maintains its policies notwithstanding the contrary temptations, hence plays the strong action regardless of reputational concerns. An irresolute state's strategy depends on reputational concerns. By Proposition 1, the irresolute L builds its reputation with $\hat{\sigma}_L(\mu) = \frac{\mu^*-\mu}{1-\mu}$ probability when its reputation is $\mu \leq \mu^*$ and spends it $\hat{\sigma}_L(\mu) = 0$ when $\mu > \mu^*$. Figure 2 shows how an irresolute state's reputation affects its probability of reputation-building.

The better an irresolute state's reputation, the less likely it is to build its reputation: each strong action increases the probability of weak actions, which, in turn, increases the probability of strong actions. Therefore, the evolution of an irresolute state's reputation appears cyclical. The cyclical pattern is more apparent in the examples in Figure 1 where Assumption 2 holds with equality. There is a one-step transition between the period immediately after L spends its reputation, and the period at which L spends its reputation again.

Suppose $\delta > 0.5$. Figure 3 shows that two phases "doubt" and "trust" endogenously emerge

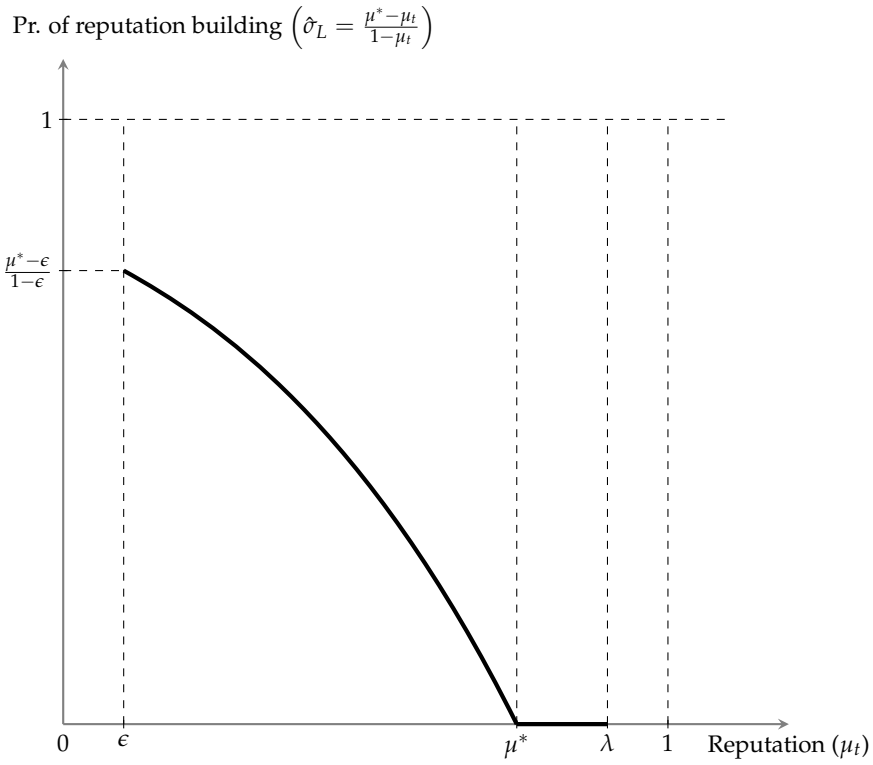

Figure 2: The probability of an irresolute state building its reputation depends on its current reputation.

in equilibrium of the repeated games with stage games in Figure 1. In the doubt phase, R believes L is likely irresolute and enters (stays out) with low probability in cooperation (conflict), whereas the irresolute L is willing to build its reputation. Once L builds its reputation, the game transitions to the trust phase, where R believes L is likely resolute and enters (stay out) with certainty in cooperation (conflict), whereas the irresolute L exploits this trust and spends its reputation. Equilibrium play alternates between these two phases.

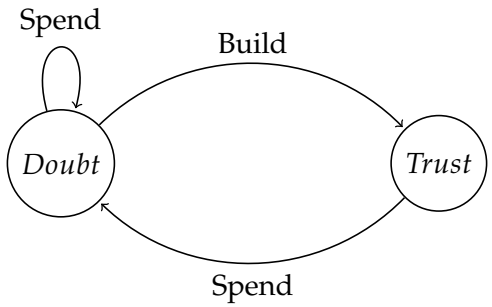

Figure 3: Cyclical equilibrium dynamics based on stage game examples in Figure 1. Transitions depend on audience observing a strong or weak action.

Therefore, a state's current reputation affects its current behavior *even when no change is realized, observed, or otherwise*. Cooperation scholars commonly focus on compliance costs to international agreements, how the variation in these costs explains differential compliance, and how changes to compliance costs break down cooperation (Koremenos 2005; Rosendorff and Milner 2001; Tomz 2007). I show that past cooperation can cause current cooperation to break down even without realized domestic or international shocks forcing states' hands to terminate a cooperation regime. This could explain cases where existing theories underpredict noncooperative behavior, such as in the sovereign debt literature regarding the default frequency (Aguiar and Gopinath 2006).

Tomz and Wright (2007) find a puzzlingly weak relationship between default and declines in economic output. Countries had suspended payments when the domestic economy was favorable and maintained debt service when it was unfavorable. I argue that, holding economic output constant, states' default behavior should vary by reputations, and thus their payment history. This variation could be why the relationship between debt payments and economic conditions is weaker than expected.

Conflict scholars similarly focus on changes in states' costs of conflict, capabilities, or issue stakes to explain variation in behavior (Fearon 1995; Powell 2006). I show that demonstrations of strength can increase incentives for reconciliation, and demonstrations of weakness can increase incentives for aggressive acts, even without changes to costs, capabilities, or issue stakes. In other words, conflict and cooperation have an inherent cyclical rhythm.

For instance, the reputation loss following the Bay of Pigs fiasco was a crucial reason why the U.S. stood firm in Vietnam (Gibbons 2014). The Cuban Study Group, assembled by President Kennedy immediately after the Bay of Pigs, criticized the failed operation and underlined that there was a need for "a changed attitude on the part of the government and of the people" to meet the global challenge of communism (23). According to James C. Thomson, who was Under Secretary Bowles' assistant and later a member of the NSC staff on Vietnam, the Kennedy administration had an uneasy sense of a worldwide challenge after the Bay of Pigs fiasco, and this "created an atmosphere in which President Kennedy undoubtedly felt under special pressure to show his nation's mettle in Vietnam (Thomson 1968)." The U.S. decided that the stage to prove its mettle would be Vietnam, not Laos, because, unlike Laotians, Vietnamese were thought to be "real fighters" and provide the necessary challenge (Gibbons 2014, 24–25).

The Kennedy administration behaved more aggressively after the U.S.' reputation was damaged, but also believed improving the U.S.' reputation would allow moderating their position. Consistent with my results that better reputations enable contrary actions, the U.S. administration believed prevailing in Vietnam would, in turn, enable the pursuit of more conciliatory policies towards the Soviet Union and China. Deputy National Security Advisor Walt Rostow advised Kennedy in a 1961 memo entitled the "Shape of Battle" that "a moderation of Communist policy" would be possible if the U.S. could prevail in Vietnam and Berlin. Then, the U.S. could "provide a golden bridge of retreat from their present aggressive positions for both Moscow and Peking.[8]"

Since current reputations affect behavior even in otherwise *identical* conditions, examining interactions in isolation without considering long-run dynamics can be highly misleading.

4.3 Increased uncertainty around observed changes, such as leader turnover

So far, periods in the model were interpreted as the passage of time, remaining agnostic about the mechanisms through which resolve changes. Instead, suppose each t captures a period of continuity, e.g., an administration. At the beginning of each t, L has new leadership. The audience observes this leadership change but not whether L's resolve changes: ϵ and λ determine whether L's resolve changes with leader turnover. If (i) L's resolve is sticky over transitions (Assumption 4), and (ii) each administration cares about welfare under future administrations $\left(\delta > \bar{\delta}\right)$, Proposition 1 applies. The relationship between the new administration's behavior and its inherited reputation is the same as in Figure 2. If L under the new administration is irresolute, the better the inherited reputation, the more likely the new administration is to spend this reputation.

That said, leadership change is typically associated with increased uncertainty about resolve (Lupton 2020; Wolford 2007; Wu, Licht, and Wolford 2021). To capture how this increased uncertainty affects behavior compared to normal times, return to the original interpretation of time and suppose a leadership change occurs at the beginning of a given t. Type transition probabilities at t are different, ϵ' and λ', about which I assume the following:

Assumption 5. *(Transition period)*

8. Papers of John F. Kennedy. Presidential Papers. President's Office Files. Staff Memoranda. Rostow, Walt W., 1961: June-December.https://www.jfklibrary.org/asset-viewer/archives/JFKPOF/065/JFKPOF-065-001?image_identifier=JFKPOF-065-001-p0008

(i) $\epsilon' > \epsilon$ and $\lambda' < \lambda$

(ii) $\frac{\epsilon}{1-\lambda+\epsilon} = \frac{\epsilon'}{1-\lambda'+\epsilon'}$

(iii) $\mu^* > \frac{\epsilon'}{1-\lambda'+\epsilon'}$

This assumption makes L's resolve under the new leadership more uncertain. Point (i) indicates that L's resolve is more likely to change during the transition. Point (ii) indicates that the transition and normal-time steady states of the Markov process governing L's resolve are identical. These keep the focus on increased uncertainty and prevents beliefs from also being skewed toward L being resolute or irresolute.[9] Point (iii) prevents L's reputation from improving beyond μ^* purely via the drift in beliefs.

Let μ be L's reputation at the end of $t-1$: after L's action but before adjusting for transitions. Let $\mu_t(\mu)$ be L's reputation at the beginning of t without leadership change, and let $\mu'_t(\mu)$ be L's reputation at the beginning of t with leadership change.

Increased uncertainty has two effects on beliefs. One, the audience is more willing to extend the benefit of the doubt if L's reputation was poor because a previously irresolute L is more likely to be resolute under the new leadership. Two, the audience is more suspicious if L's reputation was better because a previously resolute L is more likely to be irresolute under the new leadership. This implies another threshold μ^{**}, such that if L's reputation μ was lower than this threshold $\mu^{**} > \mu$ the benefit-of-the-doubt channel dominates. Then the new administration starts with a reputation $\mu'_t(\mu) > \mu_t(\mu)$. If $\mu > \mu^{**}$, the suspicion channel dominates, and the new administration starts with a reputation $\mu_t(\mu) > \mu'_t(\mu)$.

I will hold L's type constant as irresolute across scenarios to isolate the effect of increased uncertainty. This case is the most interesting since the resolute L always plays S. The following proposition describes how L's inherited reputation μ affects whether the new administration is more willing to build its reputation than the scenario with no leadership change.

Proposition 3. *Assume L remains irresolute from $t-1$ to t across scenarios. Define $\Delta(\mu) = \hat{\sigma}_L(\mu'_t) - \hat{\sigma}_L(\mu_t)$ as the change in the probability of reputation building ($a_L = S$) by L under the new administration*

9. Points (i) and (ii) jointly are akin to a second-order-stochastic-dominance relationship.

compared to the no-leadership-change scenario. We have:

$$\Delta(\mu) = \begin{cases} \frac{\mu^* - \epsilon'}{1-\epsilon'} - \frac{\mu^* - \epsilon}{1-\epsilon} & \text{if } \epsilon > \mu \\ \frac{\mu^* - \mu_t'}{1-\mu_t'} - \frac{\mu^* - \mu_t}{1-\mu_t} & \text{if } \frac{\mu^* - \epsilon}{\lambda - \epsilon} \geq \mu \geq \epsilon \\ \frac{\mu^* - \mu_t'}{1-\mu_t'} & \text{if } \frac{\mu^* - \epsilon'}{\lambda' - \epsilon'} \geq \mu > \frac{\mu^* - \epsilon}{\lambda - \epsilon} \\ 0 & \text{if } \mu > \frac{\mu^* - \epsilon'}{\lambda' - \epsilon'} \end{cases} \tag{7}$$

$\Delta(\mu)$ is strictly negative if $\mu^{**} > \mu$, strictly positive if $\frac{\mu^* - \epsilon'}{\lambda' - \epsilon'} \geq \mu > \mu^{**}$, and zero otherwise.

The proof is in the appendix. Figure 4 displays the relationship between $\Delta(\mu)$ and μ. Whether and how increased uncertainty surrounding leader turnover affects that state's behavior depends on its prior reputation μ. If the prior reputation is poor $(\mu^{**} > \mu)$, leader turnover decreases the probability that the state builds its reputation. If the inherited reputation is good $\left(\frac{\mu^* - \epsilon'}{\lambda' - \epsilon'} \geq \mu > \mu^{**}\right)$, the opposite is true. Leader turnover increases the probability that the state builds its reputation.

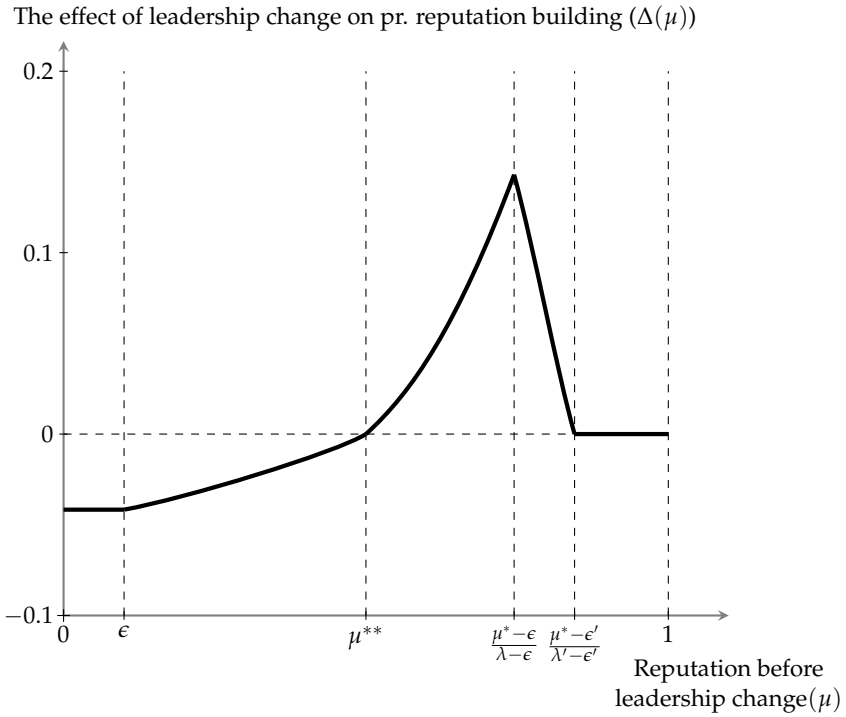

Figure 4: Y-axis is the effect of leadership change on the probability that L builds its reputation in the period of transition. X-axis is L's reputation right before the leadership change. ($\mu^* = 0.7, \mu^{**} = 0.5, \epsilon' = 0.2, \lambda' = 0.8, \epsilon = 0.1, \lambda = 0.9$)

The effect of leader turnover on behavior is non-monotonic in inherited reputations. In Figure

4, when $\frac{\mu^*-\epsilon}{\lambda-\epsilon} \geq \mu$, better prior reputations increase the effect of leader turnover. However, when $\frac{\mu^*-\epsilon'}{\lambda'-\epsilon'} \geq \mu > \frac{\mu^*-\epsilon}{\lambda-\epsilon}$, better prior reputations decrease the effect of leader turnover. Here good reputations decay faster during transition due to increased uncertainty. In the no-leadership-change scenario, a lower μ is sufficient for the state's reputation to reach the threshold μ^*. Once the state's reputation exceeds μ^*, that state spends its reputation with certainty. When prior reputation is $\frac{\mu^*-\epsilon'}{\lambda'-\epsilon'} \geq \mu > \frac{\mu^*-\epsilon}{\lambda-\epsilon}$, L's reputation at the beginning of t is lower than μ^* after leader turnover, but higher than μ^* if leadership change did not occur. This means the difference is $\Delta(\mu) = \frac{\mu^*-\mu_t'(\mu)}{1-\mu_t'(\mu)}$, which is decreasing in prior reputation.

In IR, the predominant reputational argument about leadership turnover is that increased uncertainty surrounding new leaders' resolve encourages them to prove their mettle (Lupton 2020; Wolford 2007; Wu, Licht, and Wolford 2021). Wolford (2007) labels this effect the "turnover trap." The closest paper to the above analysis is Wu, Licht, and Wolford (2021), where new leaders partly inherit their resolve from predecessors. Like here, the audience perceives continuity across leaders while recognizing the possibility of change. They find that new leaders have increased incentives to signal resolve when they inherit less of their resolve, that is, when the audience faces greater uncertainty.[10] This is consistent with Proposition 3 when L's prior reputation is high $\mu > \mu^{**}$. When $\mu > \mu^{**}$, the suspicion channel dominates the drift in audience beliefs, and increased uncertainty surrounding resolve boosts reputation-building incentives ($\Delta(\mu) > 0$).

In contrast, I argue that the effect of leader turnover is the opposite when inherited reputation is low $\mu^{**} \geq \mu$. When $\mu^{**} \geq \mu$, the benefit-of-the-doubt channel dominates the drift in audience beliefs and increased uncertainty *decreases* reputation-building incentives ($\Delta(\mu) < 0$): a *reverse* turnover trap. This does not contradict Wu, Licht, and Wolford (2021) because, being interested in the conditions under which turnover trap occurs, they do not focus on what happens when turnover trap is absent.[11] There, turnover trap is absent when the previous leader was known to have low resolve, which corresponds to the low-prior-reputation scenario here. I argue that, when inherited reputation is poor, turnover trap is not just absent but reverses direction. I also show that the effect of increased uncertainty on new leaders' reputation-building incentives is

10. They also highlight the importance of a long time horizon for reputations to matter, similar to $\delta > \bar{\delta}$ in Proposition 1.

11. They analyze only scenarios when $Pr(war_1|x^*) - Pr(war_2|w^*) > 0$ and $Pr(war_1|x^*) - Pr(war_2|y^*) > 0$. By their Proposition 2 (appendix), this happens when k, the inherited war cost, is small—when the previous leader had high resolve.

non-monotonic in their inherited reputation.

Several studies use leader turnover to proxy increased uncertainty and arrive at contradictory empirical findings. Thyne (2012) finds that civil wars involving states with stable leadership are more likely to end with peace settlements because those states can commit to peace agreements more easily. Uzonyi and Wells (2015) find that stable leadership has the opposite effect because reduced uncertainty around their resolve makes commitment problems more acute. Similarly, Rider (2013) argues that new leaders are more likely to engage in costly arms races to demonstrate resolve. I argue that these findings may not be contradictory, and that the inherited reputation is a scope condition through which these seemingly conflicting results can be rationalized.

4.4 Reputation critics, changing resolve, and disequilibrium reasoning

In a famous argument entitled the "domino theory paradox," Jervis (1997, 267) criticizes domino theory—a general deterrence theory based on reputations for resolve which formed the basis of the U.S.' Cold War foreign policy—as follows:

> The domino theory holds that even small defeats produce positive feedback because the state's adversaries and allies will infer that it is weak and prone to retreat in other conflicts. But statesmen who believe the theory and who suffer limited defeats may act especially boldly to try to show that the theory is incorrect, or at least does not apply to them. In seeking to prevent the operation of the anticipated dynamics, statesmen then disconfirm the theory.

An analogous argument, the "never again theory," is advanced in Press (2005) to highlight contradictions in reputational deterrence theories: while reputational deterrence theories expect the audience to believe that an actor with a blemished reputation is weak or distrustful, we often find that states are extra motivated to prove their mettle after defeats. Mercer (2013, 224) makes a parallel argument but focuses on states with better reputations:

> If I know that I have a reputation for resolve based on my past behavior, then I am more likely to bluff in the future (because others are unlikely to believe I am bluffing). But because others know this to be true, they are more likely to think I am bluffing - creating the paradox that a reputation for resolution means others think one is more likely to bluff and a reputation for irresolution means others think one is less likely to bluff.

Sartori (2002, 135) disagrees, claiming that the type of behavior reputation critics describe "is probably not a logical outcome of international interactions." If states received greater credibility

after defeats or bluffs, this would disincentivize reputation building due to its associated costs. Further, empirical findings contradict that audiences tend to believe states are more resolute after yielding to adversaries or defaulting on their debts (Crescenzi 2018; Tomz 2007; Weisiger and Yarhi-Milo 2015).

My results resolve this controversy in two ways. First and most importantly, reputation critics argue about how a state's current reputation conditions its behavior, which my model addresses by taking resolve as changing *and* sticky. One cannot explain that relationship by assuming resolve as either fixed or non-sticky. With non-sticky resolve, reputations do not form or are short-lived; with fixed resolve, state behavior is static in current reputations. Existing reputation models are not designed for this purpose.

In contrast, I show that when resolve is changing *and* sticky, the patterns reputation critics identify naturally emerge in reputation dynamics—they are not paradoxes. Irresolute states with poor reputations are more willing to signal resolve and those with better reputations are less willing to signal resolve (Figure 2). Nevertheless, consistent with empirical findings, audiences are increasingly convinced that a state is resolute the more they observe strong actions and not otherwise.

Second, even with changing and sticky resolve, disequilibrium reasoning leads to confusion for two reasons, which my equilibrium analysis rectifies. One, reputational concerns do not affect some states' behavior; these are the resolute types. This does not mean such states do not benefit from better reputations. Instead, it means that resolute states are able to maintain policies notwithstanding the contrary short-run temptations due to other reasons, intrinsic or extrinsic, such that reputations have no additional causal effect on their actions. Other states' behavior does change with reputational incentives; these are the irresolute types. Reputations have causal power in determining these states' behavior. The relationship between current reputations and behavior in Figure 2 pertains to those "irresolute types."

The second source of confusion due to disequilibrium reasoning is conflating the state's and its audience's perspectives. While irresolute states behave as reputation critics describe, the audience's uncertainty about states' resolve and future behavior is always present. To see this more clearly, suppose L's current reputation is μ_t. Given the irresolute L's strategy $\hat{\sigma}_L(\mu)$ in Proposition

1, the expected probability of observing strong actions across both types of L when $\mu^* \geq \mu_t$ is:

$$\underbrace{\mu_t \times 1}_{\text{Pr that } L \text{ is resolute and plays } S} + \underbrace{(1-\mu_t) \times \frac{\mu^* - \mu_t}{1-\mu_t}}_{\text{Pr that } L \text{ is irresolute and plays } S} = \underbrace{\mu^*}_{\text{Constant in } \mu_t}$$

When $\mu_t > \mu^*$, this probability becomes:

$$\underbrace{\mu_t \times 1}_{\text{Pr that } L \text{ is resolute and plays } S} + \underbrace{(1-\mu_t) \times 0}_{\text{Pr that } L \text{ is irresolute and plays } S} = \underbrace{\mu_t}_{\text{Increasing in } \mu_t}$$

Proposition 4 summarizes this discussion and Figure 5 depicts the relationship between L's current reputation and the expected probability of observing strong actions.

Proposition 4. *The expected probability of observing strong actions by L (across types) is weakly increasing in L's reputation. It is constant when $\mu^* \geq \mu$ and strictly increasing when $\mu > \mu^*$.*

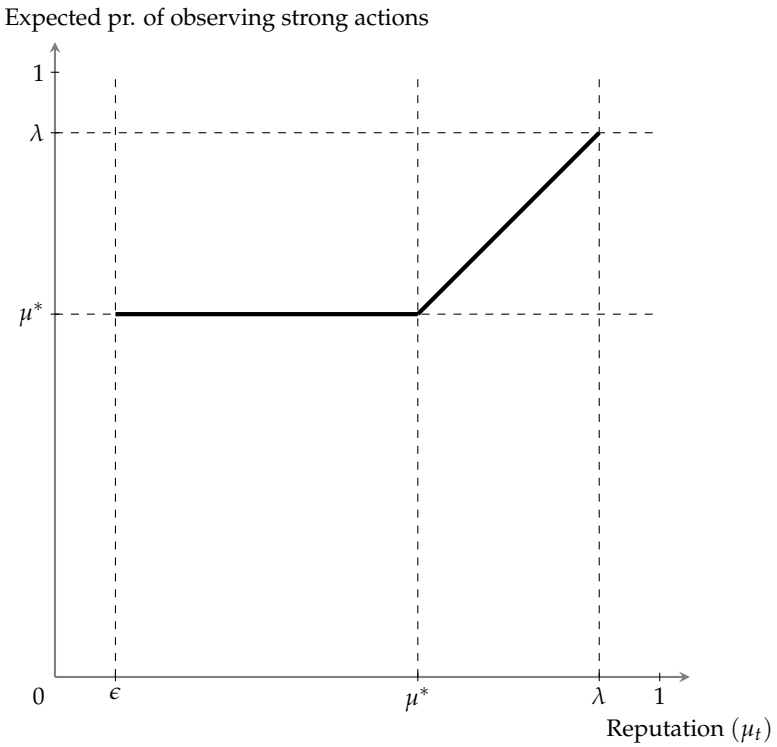

Figure 5: L's reputation and the expected probability of observing strong actions across both types of L.

The resolute L always plays S. When L's reputation is below the threshold, the irresolute

L, wanting to engage in this costly behavior no more than necessary, adjusts its probability of playing S just enough to induce R to enter in cooperation and stay out in conflict. The better L's reputation, the lower that probability needs to be. Specifically, the irresolute L adjusts the probability of S proportionally to its reputation. The result is that R faces a constant expected probability of reputation-building when $\mu^* \geq \mu$. When L's reputation is above the threshold, the irresolute L plays W and spends its reputation. Then the expected probability of observing S equals L's reputation μ. Here, the better L's reputation, the higher the expected probability of observing a strong action by L.

To conclude, while irresolute states, those which act on their reputational concerns, are *less* likely to take strong actions the better their reputations, the expected probability of observing a strong action by L is the opposite. By Proposition 4, a better reputation by L is associated with a (weakly) greater probability of observing a strong action.

4.5 Reputational enforcement under anarchy

In IR, reputations are often portrayed as a decentralized "enforcement" mechanism in the absence of strong contract-enforcing institutions, that is, under anarchy (Keohane 1984; Simmons 2010; Tomz 2007). In order to compare how reputational enforcement in the current model compares with a scenario where states commit via some perfectly enforceable contract, consider the probability of observing the (E, S) outcome in a period. This captures the probability of observing cooperation (R lends money, L pays it back) in the cooperation environment and conflict (R issues threats, L fights) in the conflict environment. An enforcement regime seeks to maximize cooperation (more (E, S) in cooperation) and minimize conflict (less (E, S) in conflict) in equilibrium. In the contract-enforcement scenario, suppose the irresolute L commits to playing S ex-ante with an enforceable contract.

Given equilibrium strategies $\hat{\sigma}_R$ and $\hat{\sigma}_L$ in Proposition 1, the probability of observing (E, S) at any period is interior. Whereas in the contract-enforcement scenario, there is always cooperation (since L always pays its debts) and no conflict (since L never backs down). In the reputational enforcement equilibrium, states spend their reputations endogenously when their current reputations improve; therefore, reputational enforcement results in *less* cooperation and *more* conflict

compared to the ideal of contract enforcement.

Proposition 5. *There is less cooperation and more conflict under reputational enforcement compared to contractual enforcement.*

Reputational enforcement is worse than the ideal of contract enforcement. IR scholars already recognize that a reputation system may fail to function well. The reasons include psychological mechanisms causing information processing problems (e.g. Jervis 1976; Mercer 1996) or multiple audiences with opposing priorities or segmented reputations on conflicting dimensions (Downs and Jones 2002). I show that, even when reputational enforcement functions well, it is still an imperfect substitute for contract enforcement. These results should reduce our confidence about the level of international cooperation supportable by reputations alone. Reputational enforcement works in cooperation, but the price is occasional breaches of trust. We should also reduce our confidence about the lengths states will go to in conflictual interactions in the name of reputation.

5 Conclusion

I analyze a dynamic model of reputations covering conflictual and cooperative interactions, where states' resolve is sticky and changing over time. The audience knows that states' resolve can change without their knowledge. I show that if audiences grapple with the possibility of unobserved change, states spend and rebuild their reputations based on their current reputations, even if *no change is realized*. Introducing the possibility of unobserved change enables explaining how states' current reputations condition their behavior.

I conclude by highlighting three areas for future research. First, this model generates a rational recency bias in audiences. The possibility of unobserved change in resolve reduces today's value of yesterday's data. One can imagine other channels that reduce today's value of old data via domestic institutions, leader psychology, culture, environmental change, and physical memory constraints. Future research could look at other sources of recency bias that can generate behavioral patterns as here. Second, this framework could be helpful in examining how various sources of global (in)stability relate to patterns of conflict and cooperation by determining the degree to which state resolve fluctuates and how that fluctuation, in turn, affects the salience of reputations.

Third, I defined resolute states as those maintaining policies notwithstanding contrary temptations. Therefore, the constraining power of enforcement mechanisms other than reputations can determine whether and when a state is resolute or irresolute. Given that the current framework examines how changes in state resolve affect reputation, it could be helpful to examine whether and when reputational enforcement can substitute or complement other forms of enforcement.

References

Aguiar, Mark, and Gita Gopinath. 2006. "Defaultable debt, interest rates and the current account." *Journal of International Economics* 69 (1): 64–83.

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.

Brutger, Ryan, and Joshua D. Kertzer. 2018. "A Dispositional Theory of Reputation Costs." *International Organization* 72 (3): 693–724.

Cilizoglu, Menevis, and Navin Bapat. 2020. "Economic coercion and the problem of sanctions-proofing." *Conflict Management and Peace Science* 37 (4): 385–408.

Clare, Joe, and Vesna Danilovic. 2010. "Multiple Audiences and Reputation Building in International Conflicts." *Journal of Conflict Resolution* 54 (6): 860–882.

Crescenzi, Mark J. C. 2018. *Of Friends and Foes: Reputation and Learning in International Politics.* Oxford: Oxford University Press.

Dafoe, Allan, and Devin Caughey. 2016. "Honor and War." *World Politics* 68 (02): 341–381.

Dafoe, Allan, Jonathan Renshon, and Paul Huth. 2014. "Reputation and Status as Motives for War." *Annual Review of Political Science* 17 (1): 371–393.

Downs, George W., and Michael A. Jones. 2002. "Reputation, Compliance, and International Law." *Journal of Legal Studies* 31 (1): 95–114.

Fearon, James D. 1995. "Rationalist Explanations for War." *International Organization* 49 (3): 379–414.

Gent, Stephen E., Mark J. C. Crescenzi, Elizabeth Menninga, and Lindsay Reid. 2015. "The reputation trap of NGO accountability." *International Theory* 7 (3): 426–463.

Gibbons, William C. 2014. *The U.S. Government and the Vietnam War: Executive and Legislative Roles and Relationships, Part II.* Princeton: Princeton University Press.

Goldfien, Michael A., Michael F. Joseph, and Roseanne W. McManus. 2023. "The Domestic Sources of International Reputation." *American Political Science Review* 117 (2): 609–628.

Harvey, Frank P., and John Mitton. 2017. *Fighting for Credibility: US Reputation and International Politics.* Toronto: University of Toronto Press.

Jervis, Robert. 1976. *Perception and Misperception in International Politics.* Princeton: Princeton University Press.

Jervis, Robert. 1997. *System Effects: Complexity in Political and Social Life.* Princeton: Princeton University Press.

Jervis, Robert, Keren Yarhi-Milo, and Don Casler. 2021. "Redefining the Debate Over Reputation and Credibility in International Security: Promises and Limits of New Scholarship." *World Politics* 73 (1): 167–203.

Johns, Leslie. 2021. "Formal Models of International Political Economy." In *The Oxford Handbook of International Political Economy*, 1–20. Oxford: Oxford University Press.

Keohane, Robert O. 1984. *After Hegemony: Cooperation and Discord in the World Political Economy.* Princeton: Princeton University Press.

Kertzer, Joshua D. 2016. *Resolve in international politics.* Princeton: Princeton University Press.

Kertzer, Joshua D., Jonathan Renshon, and Keren Yarhi-Milo. 2021. "How Do Observers Assess Resolve?" *British Journal of Political Science* 51 (1): 308–330.

Koremenos, Barbara. 2005. "Contracting around International Uncertainty." *American Political Science Review* 99 (4): 549–565.

Kreps, David M, and Robert Wilson. 1982. "Reputation and imperfect information." *Journal of Economic Theory* 27 (2): 253–279.

Leeds, Brett A. 2003. "Alliance Reliability in Times of War: Explaining State Decisions to Violate Treaties." *International Organization* 57 (4): 801–827.

Licht, Amanda A, and Susan Hannah Allen. 2018. "Repressing for Reputation: Leadership Transitions, Uncertainty, and the Repression of Domestic Populations." *Journal of Peace Research* 55 (5): 582–595.

Lupton, Danielle L. 2020. *Reputation for Resolve.* Ithaca: Cornell University Press.

Mailath, George J., and Larry Samuelson. 2001. "Who wants a good reputation?" *Review of Economic Studies* 68 (2): 415–441.

McManus, Roseanne W. 2017. *Statements of Resolve: Achieving Coercive Credibility in International Conflict.* Cambridge: Cambridge University Press.

Mercer, Jonathan. 1996. *Reputation and International Politics.* Ithaca: Cornell University Press.

———. 2013. "Emotion and Strategy in the Korean War." *International Organization* 67 (2): 221–252.

Milgrom, Paul, and John Roberts. 1982. "Limit Pricing and Entry under Incomplete Information: An Equilibrium Analysis." *Econometrica* 50 (2): 443.

Nalebuff, Barry. 1991. "Rational Deterrence in an Imperfect World." *World Politics* 43 (3): 313–335.

Phelan, Christopher. 2006. "Public trust and government betrayal." *Journal of Economic Theory* 130 (1): 27–43.

Powell, Robert. 2006. "War as a Commitment Problem." *International Organization* 60 (1): 169–203.

Press, Daryl. 2005. *Calculating Credibility: How Leaders Assess Military Threats.* Ithaca: Cornell University Press.

Renshon, Jonathan, Allan Dafoe, and Paul Huth. 2018. "Leader Influence and Reputation Formation in World Politics." *American Journal of Political Science* 62 (2): 325–339.

Rider, Toby J. 2013. "Uncertainty, salient stakes, and the causes of conventional arms races." *International Studies Quarterly* 57 (3): 580–591.

Rosendorff, B. Peter, and Helen V. Milner. 2001. "The Optimal Design of International Trade Institutions: Uncertainty and Escape." *International Organization* 55 (4): 829–857.

Sartori, Anne E. 2002. "The Might of the Pen: A Reputational Theory of Communication in International Disputes." *International Organization* 56 (1): 121–149.

Schelling, Thomas C. 1966. *Arms and Influence.* New Haven: Yale University Press.

Sechser, Todd S. 2018. "Reputations and Signaling in Coercive Bargaining." *Journal of Conflict Resolution* 62 (2): 318–345.

Simmons, Beth A. 2000. "International Law and State Behavior: Commitment and Compliance in International Monetary Affairs." *American Political Science Review* 94 (04): 819–835.

———. 2010. "Treaty Compliance and Violation." *Annual Review of Political Science* 13 (1): 273–296.

Snyder, Glenn H., and Paul Diesing. 1977. *Conflict Among Nations: Bargaining, Decision Making, and System Structure in International Crises.* Princeton: Princeton University Press.

Thomson, James C. 1968. *How Could Vietnam Happen? An Autopsy.* https://www.theatlantic.com/magazine/archive/1968/04/how-could-vietnam-happen-an-autopsy/306462/.

Thyne, Clayton L. 2012. "Information, Commitment, and Intra-War Bargaining: The Effect of Governmental Constraints on Civil War Duration." *International Studies Quarterly* 56 (2): 307–321.

Tingley, Dustin H., and Barbara F. Walter. 2011. "The Effect of Repeated Play on Reputation Building: An Experimental Approach." *International Organization* 65 (02): 343–365.

Tomz, Michael. 2007. *Reputation and International Cooperation: Sovereign Debt across Three Centuries.* Princeton: Princeton University Press.

Tomz, Michael, and Mark L. J. Wright. 2007. "Do Countries Default in 'Bad times'?" *Journal of the European Economic Association* 5 (2-3): 352–360.

Uzonyi, Gary, and Matthew Wells. 2015. "Domestic institutions, leader tenure and the duration of civil war:" *Conflict Management and Peace Science* 33 (3): 294–310.

Walter, Barbara F. 2006. "Building reputation: Why governments fight some separatists but not others." *American Journal of Political Science* 50 (2): 313–330.

Weisiger, Alex, and Keren Yarhi-Milo. 2015. "Revisiting Reputation: How Past Actions Matter in International Politics." *International Organization* 69 (02): 473–495.

Wiseman, Thomas. 2008. "Reputation and impermanent types." *Games and Economic Behavior* 62 (1): 190–210.

Wolford, Scott. 2007. "The turnover trap: New leaders, reputation, and international conflict." *American Journal of Political Science* 51 (4): 772–788.

Wu, Cathy X., Amanda A Licht, and Scott Wolford. 2021. "Same as the Old Boss? Domestic Politics and the Turnover Trap." *International Studies Quarterly* 65 (1): 173–183.

Wu, Cathy X., and Scott Wolford. 2018. "Leaders, States, and Reputations." *Journal of Conflict Resolution* 62 (10): 2087–2117.

Yarhi-Milo, Keren. 2018. *Who Fights for Reputation.* Princeton: Princeton University Press.

6 Supplementary Appendix

6.1 Proof of Proposition 1

Proof. The proof follows Phelan (2006)'s analysis, with appropriate modifications. Throughout we are maintaining Assumptions 1-4. Define $Y = \frac{u_L(E,W)-u_L(E,S)}{u_L(E,W)-u_L(O,W)}$, and $Z = \frac{u_L(O,W)-u_L(O,S)}{u_L(E,W)-u_L(O,W)}$. Assume that L discounts future payoffs by $\delta > \bar{\delta}$, where $\bar{\delta} = max\{Y - Z, |Z|\}$.

The reputation cutoff μ^* and the equilibrium strategy of the irresolute L are specified in the main text. The reputation cutoff is $\mu^* = \frac{d-u_R(E,W)}{u_R(E,S)-u_R(E,W)}$. The irresolute L's equilibrium strategy is $\hat{\sigma}_L(\mu) = \frac{\mu^*-\mu}{1-\mu}$ when $\mu^* \geq \mu$ and $\hat{\sigma}_L(\mu) = 0$ when $\mu^* > \mu$.

To complete the proof of Proposition 1, I will (i) show that given $\hat{\sigma}_L(\mu)$, L can push its reputation above μ^* in finite periods; (ii) derive the strategy of R, $\hat{\sigma}_R$, and show that when L's reputation is $\mu^* \geq \mu$, $\hat{\sigma}_R$ is increasing in L's reputation in cooperation and decreasing in L's reputation in conflict; (iii) show that when L's reputation is $\mu > \mu^*$, it is optimal for the irresolute L to spend its reputation with certainty.

Focus on (i). Given $\hat{\sigma}_L$, and R observes S, L's reputation evolves as follows in equilibrium (Equation 6 in the text):

$$\begin{aligned}\hat{\Phi}(\mu|S) &= \lambda\Big(\frac{\mu}{\mu+(1-\mu)\hat{\sigma}_L}\Big) + \epsilon\Big(1 - \frac{\mu}{\mu+(1-\mu)\hat{\sigma}_L}\Big) \\ &= \Big(\frac{\lambda-\epsilon}{\mu^*}\Big)\mu + \epsilon\end{aligned}$$

Suppose $\frac{\lambda-\epsilon}{\mu^*} \geq 1$, then the expression above is linear in μ with a slope weakly greater than 1. Therefore, μ can exceed μ^* in finite steps. Suppose $\frac{\lambda-\epsilon}{\mu^*} < 1$. The fixed point of the process $\hat{\Phi}(\mu|S)$ is:

$$\begin{aligned}\Big(\frac{\lambda-\epsilon}{\mu^*}\Big)\mu + \epsilon &= \mu \\ &= \frac{\epsilon\mu^*}{\mu^*-\lambda+\epsilon}\end{aligned}$$

Note that by Assumption 4 $\lambda > \mu^*$. This implies

$$\frac{\epsilon\mu^*}{\mu^*-\lambda+\epsilon} > \frac{\epsilon\mu^*}{\lambda-\lambda+\epsilon} = \mu^*$$

The fixed point of $\hat{\Phi}(\mu|S)$ is strictly greater than μ^*. Thus, L's reputation can exceed μ^* in finite steps.

Next focus on (ii), the strategy of R. Starting from $\mu = \mu_0 = \epsilon$, let N be the minimum number of steps required for μ to exceed μ^*. L's reputation will generically not equal μ^* at the N^{th} step, therefore I will maintain that this will strictly exceed μ^*.

Let μ^k represent L's reputation after k consecutive realizations of S, where $\mu^0 = \hat{\Phi}(\mu|W) = \epsilon$, $\mu^1 = \hat{\Phi}(\mu^0|S)$, $\mu^2 = \hat{\Phi}(\mu^1|S)$, and so on. I will specify R' strategy on this grid of beliefs $\mu \in \{\mu^0, \mu^1, ...\}$, which include all levels of reputation L can achieve starting from the prior $\mu_0 = \epsilon$. Recall that the irresolute L's strategy $\hat{\sigma}_L$ was constructed to induce indifference on R when $\mu \leq \mu^*$, and was set at $\hat{\sigma}_L = 0$ for $\mu > \mu^*$. Similarly, the strategy of R will induce indifference on the irresolute L between S and W when $\mu \leq \mu^*$, and when $\mu > \mu^*$ its strategy will be set at $\hat{\sigma}_R = 1$ in cooperation and $\hat{\sigma}_R = 0$ in conflict. Let $\hat{\sigma}_R^k = \hat{\sigma}_R(\mu = \mu^k)$. Let $V^k = V(\mu = \mu^k)$ be the continuation value of the game for the irresolute L starting from the reputation μ^k. Then for $k \in \{0, 1, ..., N-1\}$, the strategy of R will satisfy the following:

$$V^k = (1-\delta)\Big(\hat{\sigma}_R^k u_L(E,S) + (1-\hat{\sigma}_R^k)u_L(O,S)\Big) + \delta V^{k+1} \tag{8}$$

$$V^k = (1-\delta)\Big(\hat{\sigma}_R^k u_L(E,W) + (1-\hat{\sigma}_R^k)u_L(O,W)\Big) + \delta V^0 \tag{9}$$

Since the irresolute L will play W with certainty once $\mu > \mu^*$, for $k \geq N$ we have different expressions for V^k depending on the environment:

$$V^{k\geq N} = (1-\delta)u_L(E,W) + \delta V^0 \qquad \text{(Cooperation)}$$

$$V^{k\geq N} = (1-\delta)u_L(O,W) + \delta V^0 \qquad \text{(Conflict)}$$

There are $(N+1) \times V + N \times \hat{\sigma}_R^k = 2N+1$ unknowns and $2N+1$ equations laid out in the above linear system for both environments, meaning it's full rank with a unique solution. Setting $k = 0$ in equation 9 we get:

$$V^0 = \hat{\sigma}_R^0 \left(u_L(E,W) - u_L(O,W)\right) + u_L(0,W) \tag{10}$$

We solve equation 8 with 9:

$$V^{k+1} = \frac{1-\delta}{\delta}\left(\hat{\sigma}_R^k(u_L(E,W) - u_L(E,S)) + (1-\hat{\sigma}_R^k)(u_L(O,W) - u_L(O,S))\right) + V^0 \tag{11}$$

Set $k = k+1$ in equation 9, solve with equation 11, substituting the V^0 in equation 10:

$$\hat{\sigma}_R^{k+1} = \sigma_R^k\left(\frac{Y-Z}{\delta}\right) + \frac{Z}{\delta} + \hat{\sigma}_R^0 \tag{12}$$

where $Y = \frac{u_L(E,W)-u_L(E,S)}{u_L(E,W)-u_L(O,W)}$, and $Z = \frac{u_L(O,W)-u_L(O,S)}{u_L(E,W)-u_L(O,W)}$. Setting $k = 0$ here and iterating, we get the following for $\hat{\sigma}_R^k$:

$$\hat{\sigma}_R^k = \hat{\sigma}_R^0 \sum_{i=0}^{k}\left(\frac{Y-Z}{\delta}\right)^i + \frac{Z}{\delta}\sum_{i=0}^{k-1}\left(\frac{Y-Z}{\delta}\right)^i \tag{13}$$

Set $k = N-1$ in equation 10 and solve for $\hat{\sigma}_R^{N-1}$. Since $\hat{\sigma}_R^k = 1$ in cooperation and $\hat{\sigma}_R^k = 0$ in conflict for $k \geq N$ by hypothesis, we get two different expressions by environment:

$$\hat{\sigma}_R^{N-1} = (1-\hat{\sigma}_R^0)\frac{\delta}{Y-Z} - \frac{Z}{Y-Z} \qquad \text{(Cooperation)}$$

$$\hat{\sigma}_R^{N-1} = (-\hat{\sigma}_R^{N-1})\frac{\delta}{Y-Z} - \frac{Z}{Y-Z} \qquad \text{(Conflict)}$$

Set $k = N-1$ in equation 13 and solve for σ_R^0 using above expressions:

$$\hat{\sigma}_R^0 = \frac{1}{\sum_{i=0}^{N}\left(\frac{Y-Z}{\delta}\right)^i} - \frac{Z}{\delta}\left(\frac{\sum_{i=0}^{N-1}\left(\frac{Y-Z}{\delta}\right)^i}{\sum_{i=0}^{N}\left(\frac{Y-Z}{\delta}\right)^i}\right) \qquad \text{(Cooperation)}$$

$$\hat{\sigma}_R^0 = -\frac{Z}{\delta}\left(\frac{\sum_{i=0}^{N-1}\left(\frac{Y-Z}{\delta}\right)^i}{\sum_{i=0}^{N}\left(\frac{Y-Z}{\delta}\right)^i}\right) \qquad \text{(Conflict)}$$

Substituting the above for $\hat{\sigma}_R^0$ in equation 13 provides the solution for $\hat{\sigma}_R^k \in \{\hat{\sigma}_R^0, \hat{\sigma}_R^1, ..., \hat{\sigma}_R^{N-1}\}$:

$$\hat{\sigma}_R^k = \left(\frac{\Sigma_{i=0}^{k} (\frac{Y-Z}{\delta})^i}{\Sigma_{i=0}^{N} (\frac{Y-Z}{\delta})^i} \right) - \frac{Z}{\delta} \left(\frac{\Sigma_{i=k}^{N-1} (\frac{Y-Z}{\delta})^i}{\Sigma_{i=0}^{N} (\frac{Y-Z}{\delta})^i} \right) \quad \text{(Cooperation)}$$

$$\hat{\sigma}_R^k = -\frac{Z}{\delta} \left(\frac{\Sigma_{i=k}^{N-1} (\frac{Y-Z}{\delta})^i}{\Sigma_{i=0}^{N} (\frac{Y-Z}{\delta})^i} \right) \quad \text{(Conflict)}$$

Given that $\delta > \bar{\delta} = max\{Y - Z, |Z|\}$ by assumption, we have $0 < \hat{\sigma}_R^k < 1$ for $k \leq N - 1$. This concludes the specification of equilibrium strategies. Further note that Z is strictly positive in cooperation and strictly negative in conflict by Assumption 3. Then it immediately follows from the expressions above that when L's reputation is $\mu^* \geq \mu$, $\hat{\sigma}_R^k$ is increasing in μ in cooperation and decreasing in μ in conflict, as needed.

Finally, focus on (iii) it is optimal for the irresolute L to spend its reputation with certainty when $\mu > \mu^*$. The proof is by contradiction. For $k \geq N$, we want to show:

$$V^k > (1-\delta) u_L(E, S) + \delta V^{k+1} \quad \text{(Coopeartion)}$$

$$V^k > (1-\delta) u_L(O, S) + \delta V^{k+1} \quad \text{(Conflict)}$$

Suppose $k \geq N$, yet L weakly prefers S.

$$(1-\delta) u_L(E, S) + \delta V^{k+1} \geq (1-\delta) u_L(E, W) + \delta V^0 \quad \text{(Cooperation)}$$

$$V^{k+1} \geq \frac{1-\delta}{\delta} \Big(u_L(E, W) - u_L(E, S) \Big) + V^0$$

$$(1-\delta) u_L(O, S) + \delta V^{k+1} \geq (1-\delta) u_L(O, W) + \delta V^0 \quad \text{(Conflict)}$$

$$V^{k+1} \geq \frac{1-\delta}{\delta} \Big(u_L(O, W) - u_L(O, S) \Big) + V^0$$

Set $k = N - 1$ in equation 8 and solve for V^N using 9:

$$V^N = \frac{1-\delta}{\delta} \hat{\sigma}_R^{N-1} \Big(u_L(E, W) - u_L(E, S) \Big) + \frac{1-\delta}{\delta} (1 - \hat{\sigma}_R^{N-1}) \Big(u_L(O, W) - u_L(O, S) \Big) + V^0$$

Since for $k < N$ we have $0 < \hat{\sigma}_R^k < 1$, and by Assumption 3, the above expression implies

$V^N < V^{N+1}$. However, since $\hat{\sigma}_R^{k \geq N} = 1$ in cooperation and $\hat{\sigma}_R^{k \geq N} = 0$ in conflict, for $k \geq N$ it should be that $V^k = V^{k+1}$, because the continuation game starting at any prior $\mu > \mu^*$ should be identical — a contradiction. Therefore it must be that when $\mu > \mu^*$, L strictly prefers playing W as required.

This completes the proof of Proposition 1. □

6.2 Proof of Proposition 3

Proof. Let μ be L's reputation at the end of period $t-1$, that is after the realization of L's action but before the type transitions. Let $\mu_t(\mu)$ be L's reputation at the beginning of period t given no leadership change, and let $\mu'_t(\mu)$ be L's reputation at the beginning of period t given leadership change. we have $\mu'_t > \mu_t$, if:

$$\mu^{**} = \frac{\epsilon' - \epsilon}{(\lambda - \lambda') + (\epsilon' - \epsilon)} > \mu$$

Note that this expression is equal to the steady state of both Markov chains (during normal times and the period of leadership change) if and only if the steady state of both Markov chains are equal each other:

$$\frac{\epsilon' - \epsilon}{(\lambda - \lambda') + (\epsilon' - \epsilon)} = \frac{\epsilon}{1 - \lambda + \epsilon}$$

$$(\epsilon' - \epsilon)(1 - \lambda) = \epsilon(\lambda - \lambda')$$

$$\epsilon'(1 - \lambda) = \epsilon(1 - \lambda')$$

$$\epsilon'(1 - \lambda) + \epsilon\epsilon' = \epsilon(1 - \lambda') + \epsilon\epsilon'$$

$$\frac{\epsilon}{1 - \lambda + \epsilon} = \frac{\epsilon'}{1 - \lambda' + \epsilon'}$$

Which is satisfied by Assumption 5ii. This, in turn, implies $\mu^* > \mu^{**}$ by Assumption 5iii.

If the following is true for μ, then $\mu_t > \mu^*$

$$(\lambda - \epsilon)\mu + \epsilon > \mu^*$$

$$\mu > \frac{\mu^* - \epsilon}{\lambda - \epsilon}$$

Similarly, if the following is true for μ, then $\mu'_t > \mu^*$

$$\mu > \frac{\mu^* - \epsilon'}{\lambda' - \epsilon'}$$

The second condition is greater than the first:

$$\frac{\mu^* - \epsilon'}{\lambda' - \epsilon'} > \frac{\mu^* - \epsilon}{\lambda - \epsilon}$$

$$\mu^*((\lambda - \epsilon) - (\lambda' - \epsilon')) > \epsilon'(\lambda - \epsilon) - \epsilon(\lambda' - \epsilon')$$

$$\mu^* > \frac{\epsilon'\lambda - \epsilon\lambda'}{(\lambda - \lambda') + (\epsilon' - \epsilon)}$$

This inequality always holds because the RHR is strictly less than μ^{**} given that $\epsilon' - \epsilon > \epsilon'\lambda - \epsilon\lambda'$ and $\mu^* > \mu^{**}$. Further, by Assumption 5iii μ^* is strictly smaller than both $\frac{\mu^* - \epsilon}{\lambda - \epsilon}$ and $\frac{\mu^* - \epsilon'}{\lambda' - \epsilon'}$.

Finally, if the following is true for μ, then $\mu_t > \epsilon'$

$$(\lambda - \epsilon)\mu + \epsilon > \epsilon'$$

$$\mu > \frac{\epsilon' - \epsilon}{\lambda - \epsilon}$$

Now let $\Delta(\mu) = \hat{\sigma}_L(\mu'_t) - \hat{\sigma}_L(\mu_t)$, where $\hat{\sigma}_L$ is as defined in Proposition 1. Focus on the case where L was irresolute at $t-1$ and stays irresolute at t across scenarios. We will go through the cutoffs defined above for μ_t and μ'_t and determine the value of $\Delta(\mu)$ based on L's equilibrium strategy.

If $\epsilon > \mu$, then $\mu_t = \epsilon$ and $\mu'_t = \epsilon'$, then $\Delta(\mu) = \frac{\mu^* - \epsilon'}{1 - \epsilon'} - \frac{\mu^* - \epsilon}{1 - \epsilon}$ which is strictly negative.

If $\frac{\mu^* - \epsilon}{\lambda - \epsilon} \geq \mu \geq \epsilon$ then $\mu^* \geq \mu_t > \epsilon$ and $\mu^* > \mu'_t > \epsilon'$. This implies $\Delta(\mu) = \frac{\mu^* - \mu'_t}{1 - \mu'_t} - \frac{\mu^* - \mu_t}{1 - \mu_t}$. We have $\Delta(\mu) < 0$ if $\mu'_t > \mu_t$ (and thus if $\mu^{**} > \mu$) and $\Delta(\mu) > 0$ otherwise.

If $\frac{\mu^* - \epsilon'}{\lambda' - \epsilon'} \geq \mu > \frac{\mu^* - \epsilon}{\lambda - \epsilon}$ then $\mu_t > \mu^*$ and $\mu^* \geq \mu'_t > \epsilon'$. This implies $\Delta(\mu) = \frac{\mu^* - \mu'_t}{1 - \mu'_t} - 0$, which is strictly positive.

Finally, if $\mu > \frac{\mu^* - \epsilon'}{\lambda' - \epsilon'}$, then both μ_t and μ'_t are strictly greater than μ^*. L spends its reputation

with certainty in both scenarios, and thus $\Delta(\mu) = 0$. Then we have the following for $\Delta(\mu)$

$$\Delta(\mu) = \begin{cases} \frac{\mu^* - \epsilon'}{1-\epsilon'} - \frac{\mu^* - \epsilon}{1-\epsilon} & \text{if } \epsilon > \mu \\ \frac{\mu^* - \mu_t'}{1-\mu_t'} - \frac{\mu^* - \mu_t}{1-\mu_t} & \text{if } \frac{\mu^* - \epsilon}{\lambda - \epsilon} \geq \mu \geq \epsilon \\ \frac{\mu^* - \mu_t'}{1-\mu_t'} & \text{if } \frac{\mu^* - \epsilon'}{\lambda' - \epsilon'} \geq \mu > \frac{\mu^* - \epsilon}{\lambda - \epsilon} \\ 0 & \text{if } \mu > \frac{\mu^* - \epsilon'}{\lambda' - \epsilon'} \end{cases} \tag{14}$$

As per the above discussion, $\Delta(\mu)$ is strictly negative if $\mu^{**} > \mu$, strictly positive if $\frac{\mu^* - \epsilon'}{\lambda' - \epsilon'} \geq \mu > \mu^{**}$, and zero otherwise, as required.

□

6.3 On uniqueness

A uniqueness argument following Phelan (2006) would be: In any MPE, following a W action, R will always enter (stay out) with positive probability in cooperation (conflict). Otherwise this would mean that R thinks the irresolute L will play S with lower probability than is stated in our MPE, which means that the irresolute L should be able to improve its reputation much faster by playing S. This brings the period at which $\mu > \mu^*$ closer compared to the MPE above and should lead to a contradiction with R' optimization problem.

The problem is, the long-run player in Phelan (2006) does not pay any cost to take the strong action when the audience does not trust long-run player at all. The uniqueness proof there relies on the fact that the long-run player can costlessly push its reputation above μ^* to get a contradiction with R' optimization. This violates our Assumption 1. In contrast, here it is always costly for the irresolute L to play S, including after playing W in the previous period. This makes it difficult to show that the irresolute L's optimization allows playing S with positive probability in the period immediately after the irresolute L plays W. To speculate, additional constraints on how the irresolute L's costs vary by R' behavior may be necessary to handle the irresolute L's value function in this step.

Suppose this step is done, and that the irresolute L would be willing to pay the cost of pushing its reputation above μ^*, relying on the fact that R expects very little out of the irresolute L, and

thus the irresolute L can improve its reputation much faster compared to the MPE in Proposition 1. Then we can arrive at the same contradiction for R' optimization when L's reputation is $\mu > \mu^*$. This would in turn imply that R should play E and L should play S with interior probability in the period immediately after a W outcome is observed. The next step would be to show via an induction argument that when $0 < k < N$ both R and L should play mixed strategies. Since the Bellman equations describing L's continuation values above are full rank with a unique solution, the behavior in any MPE of the game should be the same as in Proposition 1.

6.4 On equilibrium dynamics if L cannot signal resolve when R stays out

In the model I present, I assume that L has the ability to take costly actions to signal its resolve to potential partners and adversaries even if it does not face explicit threats or is extended gestures of cooperation. I argued in the main text that this is a more realistic setup for dynamics I am trying to capture in this paper. Here I speculate about what happens if, as in the standard chain-store game, L cannot signal its resolve when R stays out.

The first thing to note is that this significantly complicates the formal analysis, as both L and R then will need to take into account how L's reputation drifts given no news. In fact, if my conjecture about the uniqueness of the equilibrium in Proposition 1 is correct, there may not be any MPE where strategies depend only on L's reputation. Wiseman (2008) analyzes an infinitely repeated chain-store game (akin to the conflict environment here), where the long-run player's type is subject to change, which provides clues about what equilibrium dynamics might look like in that case.

Wiseman (2008, Theorem 4) proves the existence of an equilibrium which features a cycle that repeats itself *ad infinitum*, and this equilibrium is the limit of the unique sequential equilibrium of the finitely repeated version of the game when the number of periods approaches infinity. Each cycle includes a stretch of periods where the short-run player enters with positive probability, followed by a stretch of periods where the short-run player stays out with certainty. Consistent with my suggestion that there may not be an MPE in that setting, the strategies in that equilibrium depend on both the long-run player's reputation and at which stage the play is in the current cycle. Since the long-run player is not allowed to signal when the short-run player stays out in Wiseman

(2008), both players need to take into account what "no news" means, and equilibrium strategies are more complex functions of the transition probabilities.

At each stage of the cycle in that equilibrium, except the last period, the long-run player's probability of strong action is an increasing function of its current reputation, unlike in Proposition 1. That said, in each successive period bringing the play closer to the end of the cycle, the long-run player's willingness to work for its reputation is strictly decreasing, and in the last period, the long-run player divests with certainty. Therefore, consistent with the qualitative results in here when $\mu^k < \mu^N$, the cycle in Wiseman (2008)'s equilibrium features a progressively decreased willingness by the long-run player to invest in its reputation as the play approaches to the part of the cycle where the short-run player is deterred with certainty. In the final period of each cycle, similar to the $\mu^k \geq \mu^N$ case here, the long-run player divests with certainty.