Reputations in International Conflict and Cooperation

Ekrem T. Başer*

June 27, 2020

Abstract

A defining feature of International Relations (IR) is the pervasiveness of commitment problems. According to conventional wisdom, states can overcome their commitment problems by cultivating reputations in the eyes of the international audience. Yet, IR has a static perspective on reputations, because our reputational theories take states' preferences as fixed. As such, IR lacks theories of endogenous reputation building and spending. Further, we do not know how current reputations shape incentives to maintain them: should reputations be more valuable for states with better or worse reputations? Here, I analyze a formal model covering both conflictual and cooperative interactions, where I allow states' preferences to change. When preferences are known to change, the international audience doubts states with good reputations and extends the benefit of the doubt to states with poor reputations, because things might be different *this* time. This leads to reputation building and spending behavior in equilibrium. Further, states' willingness to invest in their reputations is decreasing in their current reputations. I discuss implications for the study of international conflict and cooperation. I also show that a number of the paradoxes highlighted by reputation critics in IR are in fact consistent with the reputation mechanism when state preferences are allowed to change.

*Political Science Department, University of Illinois Urbana-Champaign, baser2@illinois.edu. I would like to thank Çağlayan Başer, Rob Carroll, Stephen Chaudoin, Nuole Chen, Xinyuan Dai, Brian Gaines, Chris Grady, Alice Iannantuoni, Jim Kuklinski, Bob Pahre, Charla Waeiss, Matt Winters and seminar participants at University of Illinois, Washington University St. Louis, 2018 Empirical Implications of Theoretical Models Institute, 2019 MPSA Linkages in IR and CPE Workshop, 2019 SPSA Positive Political Theory–Conflict Panel, and 2020 Formal Models of International Relations Conference for helpful comments and discussions. I also would like to thank the Institute for Humane Studies for providing funding for this research through the Humane Studies Fellowship. All errors remain my own.

1 Introduction

A defining feature of International Relations (IR) is the pervasiveness of commitment problems. From preventive war to tariff policies, from deterrence to sovereign borrowing, states need to overcome commitment problems to achieve better outcomes. IR literature has highlighted the incentives to build and maintain reputations for future interactions as a primary mechanism through which states can make credible commitments in both international conflict and cooperation (Keohane 1984; Schelling 1966). This emphasis has long been echoed by world leaders, who frequently allude to the need to protect their countries' reputations when making decisions. Indian Prime Minister Vajpayee justified India's actions in the 1999 Kargil Crisis by appealing to the need for protecting India's reputation for resolve.[1] In 1790, Alexander Hamilton's policy of bailing out bankrupt states was explicitly motivated as a means to cultivate a reputation for the U.S. in international markets as a reliable debtor (Sargent 2012). In 1589, Philip II of the Spanish Habsburgs refused to make peace with England, for doing so would "imperil [Spain's] reputation (Parker 1994, 127)."

A state's *reputation* refers to beliefs held by the international audience about an unknown underlying characteristic pertaining to that state. This underlying characteristic can be dispositional, such as resolve or honesty, or it can be situational, such as the domestic political costs of settling debts and acquiescing to threats — provided that such costs are sufficiently persistent over time and across interactions. The audience forms these beliefs by learning from the data gathered via observing the state's past behavior, and in turn, uses these beliefs to predict the state's future behavior. A state can favorably alter its reputation by investing in it via sending relevant costly signals to the audience, or the state can spend its reputation by taking advantage of it, damaging the reputation in the process. Thus, reputation has asset-like characteristics: it improves or deteriorates depending on the effort its holder exerts into maintaining the reputation (Schelling 1966; Weisiger and Yarhi-Milo 2015). It follows that the *reputation mechanism*, the process describing how states are incentivized to cultivate reputations and, through them, achieve better outcomes, is inherently dynamic.

The problem is, contrary to the way it is often portrayed in the literature and by policymak-

1. http://expressindia.indianexpress.com/fe/daily/19990629/fec29033.html

ers, IR has a predominantly static view of the reputation mechanism, because our reputational theories tend to assume states' preferences and abilities are fixed.[2] The static view suggests that states either find it worthwhile to cultivate their reputations and never falter, or decide that today's opportunity costs of reputation building supersede the future dividends and never exert the effort. Relying on the static view, IR scholars often invoke the reputation mechanism to explain why states engage in behavior that is seemingly counter to their short-term interests, such as fulfilling their treaty obligations, meeting debt payments, or dismiss adversarial demands at the risk of costly conflict (Büthe and Milner 2008; Guzman 2008; B. A. Simmons 2000; Walter 2006). Yet, casual observation suggests that states frequently tarnish their hard-earned reputations. They lie, cheat, give in to threats, violate foreign investors' property rights, default on their debts, and otherwise renege on their promises. Moreover, states with blemished reputations are able to rebuild them. States are able to improve their credit after defaulting on their debts, attract investment after expropriating foreign assets, or can muster a reputation for resolve after revealing weakness. The static perspective is silent about why states might damage their reputations, and how damaged reputations can be rebuilt. More generally, we do not know whose behavior is constrained by reputational concerns: should reputations be more valuable for states with better or worse reputations?

In this paper, I analyze a dynamic formal model of reputations where equilibrium behavior is consistent with the asset-like characteristics of reputations: reputation building, spending, and rebuilding are observed in equilibrium. The model concentrates on the actions of a state which continually faces a commitment problem in its interactions with others who are unsure whether the state faces a commitment problem or not.[3] The state is able to leverage its reputation via sending costly signals to the international audience to secure better outcomes for itself. The underlying commitment problem can be one which hinders cooperation, such as the problem of assuring foreign investors that their property rights would be protected. Alternatively, the commitment problem can underpin conflictual interactions, such as the difficulty of credibly committing to engaging in costly conflict to fend off threats.[4]

2. Notwithstanding a few notable exceptions, such as Tomz (2007) and Sartori (2005).

3. In the game, commitment types are modeled as action and not payoff types. I am using action types for technical convenience, but interpreting them as capturing whether the state faces a commitment problem or not. See the model section for further discussion on this point.

4. The game does not alternate between cooperative and conflictual interactions. These cases represent different

I introduce a dynamic perspective on reputations by assuming that a state's preferences and abilities, captured by its "type," are not fixed but vary over time and across interactions. This variation can be thought to reflect domestic factors, such as shifts in public opinion, or international factors, such as global economic trends. Members of the international audience do not directly observe these changes, but they know that they are making inferences about a moving target. They know that a resolved state today might have reasons to be conflict-averse the next time, or a state that has nothing to gain from cheating today might think differently tomorrow. Reputations are thus formed and maintained in a way that reflects this shifting uncertainty. This is parallel to the Realist claim that states' intentions can never be fully known. As Mearsheimer (2001, 31) says: "A state's intentions can be benign one day, and hostile the next. Uncertainty about intentions is unavoidable."

The main result of the paper is that, when reputational incentives depend on shifting uncertainty, a state facing a commitment problem has diminished incentives to exert effort for its future reputation the better its current reputation. This is true regardless of whether the commitment problem underpins a conflictual or cooperative interaction. The reason is twofold. One, states know that tarnishing their reputation today does not prevent future reputation building. The fact that states' preferences might change, together with the audience's desire to avoid being taken advantage of or avoid costly conflict, mean that benefit of doubt will be extended in the future. That is, states know that spending their reputation does not prevent future rebuilding of their reputations. Two, the knowledge that states' preferences might change prevents the audience's suspicions from being completely cleared. In other words, shifting uncertainty puts an upper limit on reputations, and in turn, puts an upper limit on the amount of benefit improving reputations will bring. Together, these two points imply that as a state's reputation gets better, there is both less room for improvement and reduced returns on further cultivating the reputation. Then, states with better reputations have lower incentives to exert further effort compared to states with poorer reputations.

Members of the international audience are aware that, if a state is facing a commitment problem, it will exert less effort the better its reputation. Yet, the audience members are nevertheless more willing to engage in cooperative interactions and less willing to engage in conflictual inter-

regions of the parameter space and are analyzed separately.

actions as a state's reputation improves. This is because while a weak/unreliable state mistaken as a strong/reliable one will not exert further effort, this same behavior ensures that states which are thought to be strong/reliable by the audience with high probability tend to be that way. In conflict, this, combined with potential adversaries' desire to avoid issuing threats when the state will not yield, is enough incentive for the audience members to avoid conflictual interactions with states with better reputations. In cooperation, the constraining power of reputations on states with poor reputations is enough incentive for the audience to endure an odd betrayal by states with better reputations. In other words, reputation "works," in the sense of securing better outcomes for its holder.

There are a number of implications for the study of both conflict and cooperation in IR. First, the results suggests that the reputation mechanism as described here cannot be a perfect substitute to strong contractual institutions with enforcement power. In particular, there is less cooperation and more conflict in equilibrium compared to a counterfactual world where states can commit via binding contracts. This is in contrast to the static perspective on reputations, where reputational incentives are sufficient to yield what is achievable via binding contracts. This can be seen as the cost of operating under anarchy, and thus is complementary to Fearon (2018).[5]

Second, reputation dynamics in this paper feature endogenous reputation damaging behavior in equilibrium. For the cooperation literature in IR, this suggests that cooperation can break down even in the absence of domestic or international exogenous shocks forcing states' hands to terminate a cooperation regime. The ingredient required for this is the *belief* that states' preferences *can* change without the audience's knowledge, not that the change actually does take place. For instance, a common problem raised in the sovereign debt literature is that the theoretical models significantly underpredict the occurrence of sovereign default compared to what is observed in the data (e.g. Aguiar and Gopinath 2006). Reputation dynamics presented here suggest a possible reason for that problem. Reputations do constrain states, however, my results should reduce the confidence in IR about the degree of international cooperation supportable by reputational incentives alone.

Third, I show that, conditional on facing a commitment problem, states with better reputations

5. Note that this is not a statement on the long-run player's welfare. See Ekmekci, Gossner, and Wilson (2012) for a formal treatment of the welfare properties of reputation games with type replacements under very general conditions.

have reduced incentives to exert further effort for their reputations. This implies that building a reputation is more valuable for actors with no established reputations, such as new states and leaders, which is consistent with recent work in IR on the subject (Wolford 2007; Wu and Wolford 2018). That said, there should be important variation in how important reputation building will be for new actors depending on the initial beliefs of the international audience about them. Building a reputation is also more valuable for those with damaged reputations. For instance, Tomz (2007) shows that states which paid their debts despite expectations of default, such as Finland and Argentina during the Great Depression, had a significantly easier time accessing credit when they wanted to borrow again. I argue that such behavior is an integral part of reputation dynamics, and thus does not necessitate significant domestic political change as is argued in Tomz (2007). Further, because states facing commitment problems are more likely to take reputation-damaging actions when their reputations are better, this could also explain empirical patterns of surprising sovereign defaults and expropriations of foreign investors' property despite favorable economic conditions in the host country (Tomz and Wright 2007).

Fourth, there is a debate in IR about the reputation mechanism, where proponents and critics argue over whether reputations matter (for a summary of the debate see Crescenzi 2018; Tomz 2007, 14-36; Weisiger and Yarhi-Milo 2015). A number of the issues raised by reputation critics to highlight the shortcomings of the reputation mechanism are about how current reputations are related to reputation building incentives, indicating that the static perspective on reputations is at least partly to blame for fueling the controversy. Press (2005), for instance, argues that after states engage in reputation-damaging actions, such as backing down in a crisis, they sometimes work extra hard to counteract the reputation theory. In similar fashion, Mercer (2013) and Jervis (1997) argue that if a state can muster a reputation for being resolved or trustworthy, then it should be able to take advantage of this reputation and afford some behavior to the contrary. I show that these points, presented by their authors as paradoxes of the reputation mechanism, are in fact consistent with the rational reputation logic outlined here.

In IR, the closest works to this paper are Crescenzi (2018), Sartori (2005), and Tomz (2007). Crescenzi (2018), which also covers both conflict and cooperation, focuses on the fact that different interactions contain different types of information for a state's reputation, which depend on how similar the observed states and their relationships are to those of the observer. This paper

focuses on how reputation dynamics unfold if it is common knowledge that preferences change, and unlike in Crescenzi (2018), analyzes the underlying strategic behavior. Sartori (2005) presents a theory of diplomacy where effective diplomacy is possible through obtaining a reputation for honesty. Similar to the changing types premise here, in Sartori (2005) the value of the issue in dispute changes from interaction to interaction. That said, the focus in Sartori (2005) is on when "cheap-talk" is effective, whereas reputation dynamics here rest on costly signaling (Fearon 1994, 1997). Further, Sartori (2005) exogenously sets how long a tarnished reputation will persist before allowing states to rebuild their reputations, whereas here, improvements and decays in reputations are endogenous. Finally, Tomz (2007) provides a reputational theory of sovereign borrowing and default, where states' types are subject to change, much like in here. However, the theory in Tomz (2007) does not include a formal treatment or specification of equilibrium behavior. Further, type replacements in Tomz (2007) play a direct role in causing behavioral change: e.g. a state fulfilling its debt obligations would default once public opinion loses appetite for debt payments. In contrast, this paper focuses on how the possibility of change determines the behavior of states even when their preferences remain stable.

In economic theory, following the seminal Cripps, Mailath, and Samuelson (2004), which showed reputations are a short-term phenomenon in the absence of a mechanism to replenish the underlying uncertainty, a number of papers focusing on long-run reputation dynamics have been published (Board and Meyer-ter-Vehn 2013; Bohren 2013; Ekmekci, Gossner, and Wilson 2012; Faingold and Sannikov 2011; Liu 2011; Liu and Skrzypacz 2014; Phelan 2006; Wiseman 2008).[6] Among these, Ekmekci, Gossner, and Wilson (2012), Phelan (2006), and Wiseman (2008) are explicitly focusing on type replacements, and are thus the closest to this paper. Ekmekci, Gossner, and Wilson (2012) concentrates on the welfare properties of reputation games with replacements. They focus on how payoff bounds established by Fudenberg and Levine (1989, 1992) change with the introduction of replacements. In contrast, I focus on the equilibrium behavior and how it changes with the introduction of replacements. Wiseman (2008) analyzes an infinitely repeated chain-store game where the monopolist's type alternates between weak and strong. Phelan (2006) looks at how reputational concerns influence the behavior of a government with a changing type

6. A small set of studies prior to Cripps, Mailath, and Samuelson (2004) worked with changing types as well (Cole, Dow, and English 1995; Holmstrom 1999; Mailath and Samuelson 2001)

deciding on how much to tax a continuum of citizens. Formally, the game in Phelan (2006) can be considered a special case of the cooperation condition presented here.[7] The analysis in this paper is based on Phelan (2006)'s equilibrium characterization and his uniqueness proof, therefore the formal contribution of this paper is limited to showing that Phelan (2006)'s results extend to the conflict condition in a straightforward manner. Substantively, however, none of these papers discuss international politics, or examine cooperative and conflictual interactions together.[8]

In what follows, I first discuss the importance of reputational incentives in the presence of anarchy. Next, I discuss why incorporating changing preferences is important to understand reputation dynamics. This is followed by the introduction of the formal model. The analysis is structured so that the reader can contrast reputation dynamics when state preferences are fixed and when they are allowed to change. Before concluding, I discuss the aforementioned implications for the study of conflict and cooperation in general, and the IR literature on reputations in particular.

2 Reputations and Commitment Problems

Scholars of international politics have long considered a state's reputation to be one of its most important assets. In his account of the Peloponnesian War, Thucydides refers to reputation as something worth protecting at all costs. Schelling (1966, 124) argues that reputation is "one of the few things worth fighting over." To Keohane (1984), the existence of reputational incentives is a core reason why international relations need not be as conflict ridden as argued by structural realists (Mearsheimer 2001; Waltz 1979).

The emphasis on reputational incentives, both by scholars and policy-makers, should be understood within the context of anarchy. Even the most minimal definition of anarchy, inability to write enforceable contracts, would imply that states operating in the international arena face particularly acute commitment problems.[9] States want to assure creditors that they will settle their

7. There is an important difference, however. In Phelan (2006), the government can costlessly signal if no citizen produces. Here, signalling is always costly for the long-run player.

8. Ekmekci, Gossner, and Wilson (2012)'s framework is general enough to cover any reputation game, including but not limited to the cooperation and conflict conditions presented here — although there is no discussion of this in their paper. However, as mentioned above, the authors are interested in using entropy techniques to examine equilibrium payoff bounds, not behavior.

9. Numerous arguments exist over how to define anarchy and what it implies for state behavior. For prominent

debts, they want to assure foreign companies that their investments will be protected, they would like to assure allies that they will help when in need, they would like to assure adversaries that, if threatened, they would not back down. However, states' allies, adversaries, and future partners, hence, the international audience, recognize that there are incentives to renege on commitments. Even when all parties would have been better-off had states were able to make credible commitments, everyone gets stuck in Pareto-inferior outcomes. Hence, anarchy makes international politics tragic.[10]

This is where reputations come in. In the absence of enforceable contracts, reputations act as a vehicle through which states can make credible commitments. Reputational incentives emerge because important underlying factors pertaining to states' preferences and abilities are private information. These characteristics can be dispositional, such as resolve, or situational, such as the economic need to ensure access to international lending markets. States know that their interactions are monitored by the international audience, whose members form beliefs about these unknown factors by learning from states' past behavior, which are in turn used to make predictions about future behavior. These beliefs held by the audience, which I label reputations, establish a link between interactions of today and tomorrow.

The possibility of establishing a reputation in the eyes of the audience and reaping its long-term benefits incentivizes self-seeking states to send costly signals that are otherwise contrary to their short-term interests. These costly signals should be of the type that could favorably alter the audience's beliefs towards the state, such as cutting government spending to demonstrate good debtor behavior, or weapon tests and military maneuvers to signal resolve. Given the existence of reputational incentives, commitments gain credibility. It is as if a self-seeking state declares to the international audience that: "I am committing to this behavior today, but I know that you —my partner, my adversary— might be skeptical that I will follow through on my commitment. Notice that I will engage in many interactions that are similar to the one at hand, in the future. If I do not follow up on my commitment today, others will take note, my reputation will suffer, and I will be

examples, see Lake (1996), Mearsheimer (2001), Milner (1991), Powell (1994), and Wendt (1992). The minimal definition I provide, however, should not be controversial. That said, the ability to write enforceable contracts would be limited whenever those who hold power cannot commit to not using their power (Acemoglu 2003). In this sense, anarchy is not at all *sui generis* to IR, but IR is a predominantly anarchical domain.

10. See Fearon (1995, 2018), Hirshleifer (1995), and Powell (2006) as examples of research which frame the costs of anarchy in IR via commitment problems.

worse-off, because of it. As you see, I cannot let this happen."

In other words, under anarchy, reputation is a primary mechanism through which states can circumvent commitment problems. This point has been made early on by IR theorists who argued that anarchy need not be so tragic: if states can make credible commitments due to their reputations being at stake, we should be optimistic about international cooperation (Keohane 1984). This means international law can be self-enforcing (Guzman 2008; B. A. Simmons 2000); diplomacy can be more than just cheap talk (Sartori 2005); states can attract foreign investment via good behavior (Büthe and Milner 2008); international lending markets can flourish, notwithstanding the incentives to default (Cole and Kehoe 1998; Tomz 2007). On the conflict side, this means states have a further incentive to be resilient when facing threats from adversaries. If states can build a reputation for toughness or resolve, perhaps it is worth paying the price of escalating the conflict today to deter future threats (Schelling 1960; Sechser 2010; Weisiger and Yarhi-Milo 2015). This logic can also be applied to domestic contexts whenever credible commitment problems arise. As Walter (2006) argues, states do not want to concede to separatist rebels, because they are afraid that this will encourage future separatists. Similar dynamics are said to be at play in governments' calculus of whether to repress dissent as well. Given dissent, a government's failure to follow through on its threat of repression can embolden future groups to challenge its rule as well (Licht and Allen 2018; Ritter 2014).

It is important to note that another mechanism borne out of repeated interactions, the threat of retaliation, can also alleviate commitment problems. The threat of facing retaliation, either by another state or through community punishment, can make states' commitments credible (Axelrod and Keohane 1985; Fudenberg and Maskin 1986; Rubinstein 1979). There are two potential issues with this mechanism. One, the threat of punishment may not be credible due to the costs, including opportunity costs, of punishment for the punisher. Two, if the punishment requires coordination by multiple states to be effective, then a plethora of problems might stifle the effort, not the least of which is free riding. That said, as Olson (1965) and Ostrom (1990) demonstrate, small societies can be very effective at collective action, and the society of states is fairly small. Regardless of the feasibility of retaliatory solutions to commitment problems, behavior in these settings neither require uncertainty nor does it feature forming and updating of beliefs, and thus are distinct from the reputation mechanism. As such, this mechanism is out of scope for this paper.

3 Reputations and Shifting Uncertainty

Reputational incentives require uncertainty. If there is complete information, if all states are fully informed about the costs and benefits that accrue to each others' actions, there would be no room for self-seeking states to influence others' beliefs in their favor, and reputational incentives would cease to exist. Extant theoretical arguments in the IR literature stipulate that states should invest in their reputations, because if a state spends its reputation by engaging in myopic actions, such as showing weakness in the face of adversity or defaulting on its debt, that state reveals its preferences to the audience — it reveals its "type" (Alt, Calvert, and Humes 1988; Dafoe and Caughey 2016; Sechser 2010; Tingley and Walter 2011). Members of the international audience recognize that they would not observe such reputation tarnishing behavior if the state did not face a commitment problem. The audience infers that the state must be weak, an untrustworthy ally, or an unreliable borrower, and future relationships reflect this revised opinion. Since the audience's uncertainty about the state's "type" disappears, so does the reputational incentives. Faced with this prospect, the state should either (i) decide to exert effort to boost its reputation and never falter, or (ii) decide that bearing the cost of exerting effort is not worth the downstream benefits, and never build its reputation. Hence, conceptualized this way, reputation dynamics are anything but dynamic.

There are two problems with this picture. First, in the real world, states frequently engage in reputation-tarnishing behavior: they lie, they cheat, they default, they cave in. Beyond appealing to exogenous shocks, our reputation theories have little to say about why a state would tarnish a reputation, after spending the effort to build it. Second, once states engage in such reputation-tarnishing behavior, we often observe that they are able to rebuild their reputations, if gradually. They are able to find international lenders to borrow from, able to attract investment, or muster a reputation for resolve. If states reveal their "true types" after engaging in myopic behavior, like the static reputation mechanism suggests, how are they able to rebuild their reputations?

A potential answer could be about imperfect monitoring. Sometimes the international audience is unable to observe state actions directly or has difficulty interpreting the meaning of the actions they do observe (Jervis 1976). If the international audience only observes outcomes that are imperfectly correlated with the actions which generate them, or incorrectly interpret the signals

they receive some of the time, then it might be possible that the audience will ignore reputation damaging signals every now and then, attributing them to factors outside of the states' control. Cripps, Mailath, and Samuelson (2004) show that in such settings characterized by both uncertainty about agents' preferences and imperfect monitoring, it is impossible to sustain long-run reputational incentives. The intuition is that, over time, reputations become so entrenched that eventually the temptation to take advantage of others' favorable beliefs become irresistible. Over the long run, however, such behavior is informative for the audience, which will eventually figure out the preferences of that state. Uncertainty will then vanish, destroying reputational incentives. This means imperfect monitoring cannot lead to reputation spending and rebuilding behavior while also sustaining reputational incentives. Instead, there needs to be a mechanism that would continually replenish the uncertainty at the heart of reputational incentives (Cripps, Mailath, and Samuelson 2004; Ekmekci, Gossner, and Wilson 2012).

Criticizing formal IR work on reputations, Mercer (1996) identifies one such crucial source of persistent uncertainty in international affairs:

> Although existing formal work on reputation assumes that the actor's type is "perfectly correlated across games," in principle there is no reason why an actor's type in a new situation could not be imperfectly correlated with its type in previous situations. This would be a step toward a formal understanding of general reputations (38).

Here, Mercer (1996) identifies a core mechanism through which long-run reputation dynamics can emerge. In the real world, states' "types" are indeed not fixed, as is assumed in reputation models in IR. They continually evolve based on variables such as changes in leadership, shifts in public opinion, and state of the economy. While some of these variables are observable by the international audience, many more are not and thus need to be inferred from behavior. This is why intelligence services heavily invest in information gathering for governments, and there exists a political risk analysis market serving the global investment community. Further, thinking of states as having changing types as opposed to fixed types also captures the structural Realist assertion that states' intentions can never be fully known (Waltz 1979; Mearsheimer 2001). If anarchical international system is associated with persistent doubt about others' intentions, as structural Realists argue, we need shifting uncertainty. This premise is the core component of the reputation model I present in the next section.

4 Model

The model is an infinitely repeated game of adverse selection in discrete time; hence it is based on the adverse selection approach to modeling reputations commonly employed in political science and economics (Alt, Calvert, and Humes 1988; Fudenberg and Levine 1989; Kreps and Wilson 1982; Mailath and Samuelson 2001; Tingley and Walter 2011). In each period, a long-run player, state L, faces a new short-run player, which I will generically refer to as state S. Short-run players, S, enter the game for only a single period, after which they are replaced by a new S in the next period. Since the primary goal is to examine how a state's incentives to invest in its reputation change as a function of its reputation, the model structure is designed to emphasize the long-run player L's incentives, while anonymizing its partners and adversaries. As such, this model is aimed to capture general reputations, as opposed to reputations built for and perceived by a specific partner or adversary. In this sense, this model is closer to the idea of reputations as mentioned in Crescenzi (2018), Sartori (2002), B. A. Simmons (2000), Tomz (2007), and Weisiger and Yarhi-Milo (2015), where states invest in their reputations as perceived by the relevant international audience. By contrast, the reputations captured in Kydd (2007) and Wolford (2007) are different in kind, as the focus is on the interactions between two states, such as the U.S. and the Soviet Union trying to establish trust between each other. I describe the stage game played between states L and S in the next section.

4.1 Stage Game

The stage game depicts an interaction between two states $i \in \{S, L\}$, where state S decides whether to engage with L or not, and L decides whether to invest in its reputation or not. The players move simultaneously. Let $A = A_S \times A_L$ be the set of all action profiles, where A_i represents actions available to player i. Let $A_S = \{E, O\}$, where $a_S = E$ stands for entering and $a_S = O$ for staying out, and $A_L = \{I, D\}$, where $a_L = I$ is exerting effort or investing in reputation and $a_L = D$ stands for no effort or divesting in reputation. Let ΔA_i denote the set of all mixed strategies available to player i, with $g_i(a_i)$ representing an element of ΔA_i. Note that $g_L(a_L = I)$ can be interpreted as L's effort level which yields the signal I with probability $g_L(a_L = I)$ and signal D with the complementary probability. I will use $a_i \in A_i$ to represent $g_i(a_i) = 1$.

The stage game payoff to player i is $U_i(g_S, g_L) = \Sigma_{(a_S,a_L)\in A} u_i(a_S, a_L) g_S(a_S) g_L(a_L)$. I will assume the following about $u_i \colon A \to \mathbb{R}$.

Assumption 1 (*Effort is costly*)**.** $u_L(a_S, D) - u_L(a_S, I) > 0$ for all $a_S \in A_S$

This means, regardless of the action taken by S, exerting effort is costly for L. Note that this cost can be different depending on whether S engages with L or not.

Assumption 2 (*Independent outside option*)**.** $u_S(O, I) = u_S(O, D) = d$

If S decides to stay out and not interact with L, the outside option of S does not depend on what L does.

Assumption 3. Payoffs are described by either the cooperation or the conflict condition as defined below:

A. *Cooperation condition:*

i. $u_L(E, a_L) - u_L(O, a_L) > u_L(a_S, D) - u_L(a_S, I)$ for all $a_L \in A_L, a_S \in A_S$

ii. $u_L(E, D) - u_L(O, D) \geq u_L(E, I) - u_L(O, I)$

iii. $u_S(E, I) > d > u_S(E, D)$

B. *Conflict condition:*

i. $u_L(O, a_L) - u_L(E, a_L) > u_L(a_S, D) - u_L(a_S, I)$ for all $a_L \in A_L, a_S \in A_S$

ii. $u_L(O, D) - u_L(E, D) \geq u_L(O, I) - u_L(E, I)$

iii. $u_S(E, I) < d < u_S(E, D)$

In the cooperation condition, the first point indicates that L's gain from when S plays enter instead of stay out is greater than the cost L has to pay to invest in its reputation: L wants S to enter. We can interpret this as L prefers borrowing money from S to not, receiving foreign investment from S to not, or more generally, prefers that S trusts L. If this assumption is violated, that is, if the payoff from inducing S to enter is not worth the cost of effort L has to bare, L would never want to invest in its reputation. Therefore, this assumption makes it possible for L to be responsive to reputational incentives. The second point in the cooperation condition means that

when L does not invest in its reputation, the gain from inducing S to enter instead of staying out is weakly greater. For instance, L's gain from borrowing S's money when L has to cut government spending can at most equal L's gain when it does not have to cut government spending. The third point indicates that S wants to enter if L will exert effort, but would prefer staying out otherwise. Importantly, this means S wants L to invest in its reputation.

In the conflict condition, the first point indicates that L's gain from when S stays out instead of entering is greater than the cost L has to pay to invest in its reputation. That is, L wants S to stay out; L prefers not to be threatened. The second point indicates that when L does not invest in its reputation, the gain from inducing S to stay out instead of entering is weakly greater. For instance, L is weakly better-off if S decides to stay out and not threaten L without requiring L to signal strength such as via troop mobilizations or building up arms, compared to if S staying out requires L to mobilize troops. The third and final point means that S wants to threaten L if L will acquiesce to its demands, but not otherwise. This means S strictly prefers L to not invest in its reputation.

Together, these assumptions imply that state L is facing a commitment problem. If Assumption 3A —the cooperation condition— is true, then this commitment problem is akin to those discussed in literatures such as sovereign debt, foreign direct investment, and international trade. For instance, a state would like to attract foreign investment, but foreign actors are willing to invest only if they trust the state to exert costly effort to protect the investors' property rights.[11] If Assumption 3B —the conflict condition— applies, then the commitment problem is akin to the ones referred to in the deterrence or interstate crisis literatures. For instance, a state prefers not to be challenged by another party, while the other party would prefer challenging only if they think the state will back down.[12] In both cases, L would be better-off if it could commit to exerting effort (e.g. protect property rights or refuse to back down) ex-ante but it cannot, due to anarchy. If the cooperation condition applies, in the unique Nash equilibrium (NE) of the stage game, S stays out and L does not exert any effort. If the conflict condition applies, then S enters and L again does not

11. The "product-choice" or "quality" games often analyzed in reputation models in Economics are covered by assumptions 1,2, and 3A.

12. This is the perfect monitoring version of Selten's chain-store game, or the entry-deterrence game, analyzed by Kreps and Wilson (1982), where the Monopolist's off-the-equilibrium path decision when the challenger stays out is still observed. To put it differently, in the version here, the monopolist would have the ability to signal when it is not challenged by an entrant.

exert any effort in the unique NE of the stage game. Figure 1 displays two stage game examples: one where the cooperation condition applies and one where the conflict condition applies.

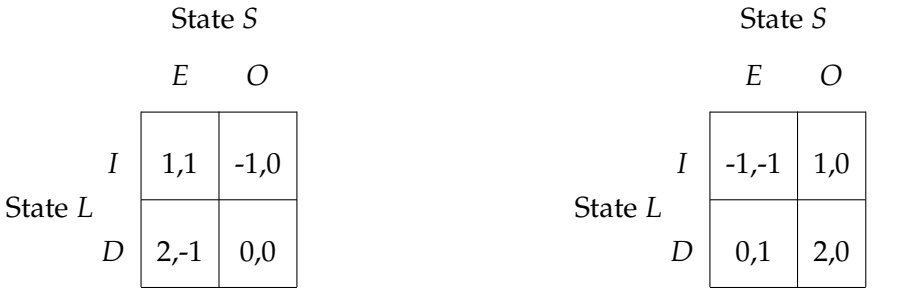

Figure 1: Stage game examples. The panel on the left is described by the cooperation condition. The unique NE is (*O*,*D*). The panel on the right is described by the conflict condition. The unique NE is (*E*,*D*)

4.2 Repeated Game and Information

The stage game described above is repeatedly played by L with a new S in each period, for infinitely many periods. State L discounts future payoffs by a factor of $\delta \in (0,1)$.

Reputational incentives are introduced in the model via private information state L has about its "type." Let $\Theta = \{N, C\}$ be the type space which contains two types: a commitment type (C) who is an action type programmatically exerting effort or playing I with certainty, and a strategic normal type (N) with stage game payoffs as described in the previous section.[13] I will focus on the normal type of L's behavior throughout the paper, and if the type is not specified, L always refers to the normal type of L. State S does not know which type of L it is facing, but has beliefs over the type space. Let μ_t be the prior probability S assigns to L being a commitment type at time $t \in \{0, 1, ...\}$, with the prior at the beginning of the game being μ_0. I define L's reputation as follows:

Definition 1 (Reputation)**.** The prior probability state S assigns to state L being the commitment

13. As long as it is common knowledge that L can be a commitment type with positive probability, the introduction of more commitment types, or a "bad type" which always takes reputation damaging actions should not change the results qualitatively. If the type space only involves bad and normal types, however, the results can be very different, as is shown in Mailath and Samuelson (2001). Therefore, the consequential assumption is the existence of (or the belief about the existence of) a type committed to investing in its reputation, and the results presented here apply to real world situations when this assumption is reasonable.

type at time t, μ_t, is the *reputation* of L.

The existence of commitment types in this model allow for but does not necessitate that reputations pertain to a dispositional trait of state L. In fact, I prefer to interpret this modeling choice as a proxy for S's uncertainty about L's stage game payoffs, similar to Fudenberg and Levine (1992). Specifically, I am referring to S being uncertain regarding whether L does or does not face a commitment problem at a given time. The choice of modeling this uncertainty via nonstrategic commitment types makes the analysis more tractable, especially when effort costs get arbitrarily small. In order for a strategic state which is maximizing its payoffs to exhibit the same kind of behavior as a nonstrategic commitment type, its payoffs should make investing in reputation to be a weakly dominant strategy in the infinitely repeated game. This is satisfied if Assumption 1 does not hold (effort is not costly), and thus if L does not face a commitment problem in the stage game. The equivalance would also require further assumptions on the evolution of beliefs: S should not think that L is more likely to be a commitment type when it observes D.[14] On the connection between action and payoff types, see Weinstein and Yildiz (2016) for the proof that any infinitely repeated game with commitment types is strategically equivalent to a game with incomplete information about the stage-game payoffs.[15]

The core premise underlying the arguments of this paper is that states' preferences are not fixed, but change over time. This is captured by allowing L's type to change at the beginning of each period, following a simple Markov process. Importantly, while S knows that L's type might change, it does not directly observe the actual replacements. If L is a commitment type in period t, it will switch to a normal type in the next period with $1-\lambda$ probability, and with probability λ, it will remain a commitment type. If L is a normal type in period t, in $t+1$ it will be a commitment type with probability ϵ, and will remain a normal type with probability $1-\epsilon$. I will maintain that L's initial reputation is $\mu_0 = \epsilon$. The transition probabilities are common knowledge. The type transition matrix described above can be summarized as follows:

14. Kreps and Wilson (1982, 263) appeal to to this criterion for beliefs. Cho and Kreps (1987) provide a formal treatment of this restriction for sequential equilibria in signaling games, though the games they analyze do not have the richer dynamic structure as in here.

15. Weinstein and Yildiz (2013) prove this for finitely repeated games.

$$
\begin{array}{c} \\ \theta_t = Normal \\ \theta_t = Commitment \end{array}
\begin{array}{c}
\begin{array}{cc} \theta_{t+1} = Normal & \theta_{t+1} = Commitment \end{array} \\
\begin{pmatrix} 1-\epsilon & \epsilon \\ 1-\lambda & \lambda \end{pmatrix}
\end{array}
$$

The effect of type changes is twofold. One, it modifies how L views the future, making its effective discount factor $\delta(1-\epsilon)$. Two, and more importantly, the fact that today's normal type of L can be a commitment type tomorrow prevents the continuation value of the game for L to ever equal the infinite repetition of the stage game NE, and in so doing, introduces dynamism to the interaction necessary to generate building and spending behavior in equilibrium. I will assume the following about the transition probabilities:

Assumption 4 (Infrequent type changes)**.** The following are true about the transition probabilities:

- $\epsilon < \frac{d-u_S(E,D)}{u_S(E,I)-u_S(E,D)}$
- $\lambda > \frac{d-u_S(E,D)}{u_S(E,I)-u_S(E,D)}$

This assumption makes sure that, if S is certain of L's type today, the type replacements are rare enough that tomorrow's S will be able to act upon this knowledge. Suppose the first point is violated, i.e. $\epsilon \geq \frac{d-u_S(E,D)}{u_S(E,I)-u_S(E,D)}$. This means, even when L damages its reputation by playing D in period t and S is sure that L is a normal type at the end of period t, at $t+1$, the probability that L is a commitment type will be high enough to induce S to enter in cooperation, and stay out in conflict. If the second point is violated, i.e. $\lambda \leq \frac{d-u_S(E,D)}{u_S(E,I)-u_S(E,D)}$, then even having a perfect reputation at the end of period t will not be enough to induce S to enter in cooperation and stay out in conflict at period $t+1$. In either case, L derives no benefit in exerting effort, and thus reputational incentives disappear.

4.3 Strategies and Equilibria

Repeated games, such as the one described here, are notorious for their indeterminacy due to having multiple equilibria. Here I restrict attention to Markovian strategies with L's reputation serving as the state variable.

First, let's define strategies according to the Markov state variable μ; L's reputation. A Markov strategy for the normal type of L is a function $\sigma_L(\mu)\colon [0,1] \to [0,1]$ which takes the reputation of L and returns a mixed action, the probability for playing I, which is interpreted as L's effort level. The commitment type of L always plays I with certainty, thus exerts maximal effort. Similarly, a Markov strategy for S is a function $\sigma_S(\mu)\colon [0,1] \to [0,1]$ which takes the reputation of L and returns a mixed action; a probability for S to enter.

Since each S interacts with L for only a single period, S cares about payoffs from only the current period. Let p be the effort level S expects across L's two possible types given entry; that is, $p = \mu + (1-\mu)\sigma_L(\mu)$. If S stays out, it receives its outside option d, and if it enters, it expects to receive $pu_S(E,I) + (1-p)(u_S(E,D))$. Then, there is a cutoff effort level p^* which will induce indifference between entering and staying out. Define p^* as:

$$p^* = \frac{d - u_S(E,D)}{u_S(E,I) - u_S(E,D)}$$

This expression takes values between 0 and 1 by Assumption 3. In the cooperation condition, for any reputation level μ, S's strategy σ_S is optimizing if $\sigma_S > 0$ implies $\mu + (1-\mu)\sigma_L \geq p^*$ and $\sigma_S < 1$ implies $\mu + (1-\mu)\sigma_L \leq p^*$. In the conflict condition, the reverse is true: $\sigma_S > 0$ implies $\mu + (1-\mu)\sigma_L \leq p^*$ and $\sigma_S < 1$ implies $\mu + (1-\mu)\sigma_L \geq p^*$. Then, in either condition $0 < \sigma_S < 1$ implies $\mu + (1-\mu)\sigma_L = p^*$. Note that, if L's reputation is sufficiently high, S will enter in cooperation and stay out in conflict regardless of the effort level of the normal type of L, in particular, even when the normal type of L exerts no effort ($\sigma_L = 0$). If L's reputation is $\mu > p^*$ this implies $\mu + (1-\mu)\sigma_L > p^*$, which means $\sigma_S = 1$ in cooperation and $\sigma_S = 0$ in conflict. Then we can consider the following as the cutoff reputation, where if L's reputation exceeds this threshold, S strictly prefers entering in cooperation and staying out in conflict:

$$\mu^* = \frac{d - u_S(E,D)}{u_S(E,I) - u_S(E,D)}$$

Upon observing the realization of L's action at time t, S updates its prior belief μ_t about L's type according to Bayes' Rule. This posterior is then adjusted according to the type transition probabilities to arrive at the prior belief μ_{t+1} at time $t+1$. Let $\Phi(\mu|a_L) = \mu_{t+1}$ for the realized action $a_L \in \{I, D\}$:

$$\Phi(\mu|I) = \lambda\Big[\frac{\mu}{\mu + (1-\mu)\sigma_L}\Big] + \epsilon\Big[1 - \frac{\mu}{\mu + (1-\mu)\sigma_L}\Big] \tag{1}$$

$$\Phi(\mu|D) = \epsilon \tag{2}$$

The first expression describes the reputation of L at the beginning of time $t+1$, once S observes I at time t. This expression is strictly decreasing in σ_L, that is, the less effort S expects L to exert, the higher L's gain in reputation will be if the observed signal is I. The second expression indicates that if state S is certain that L is the normal type at the end of period t, L starts the next period with the reputation $\mu_{t+1} = \epsilon$. Therefore ϵ represents the lower bound for how bad L's reputation can get. Similarly, if S is certain that L is the commitment type at the end of period t, L starts the next period with the reputation $\mu_{t+1} = \lambda$, which represents the upper bound for L's reputation.

L's behavior today influences its future payoffs via its reputation. Let $V(\mu)$ represent the continuation value of the game for L when its reputation is μ, for a given strategy profile $(\sigma_S(\mu), \sigma_L(\mu))$:

$$\begin{aligned} V(\mu) = \sigma_L\Big(\sigma_S u_L(E,I) + (1-\sigma_S)u_L(O,I) + \delta(1-\epsilon)V(\Phi(\mu|I))\Big) \\ + (1-\sigma_L)\Big(\sigma_S u_L(E,D) + (1-\sigma_S)u_L(O,D) + \delta(1-\epsilon)V(\epsilon)\Big) \end{aligned} \tag{3}$$

The first term at the right-hand side of the equation captures L's continuation payoff if its realized action in the current period is I, which means L's reputation in the next period will be $\Phi(\mu|I) > \mu$. The second term captures L's continuation payoff if the realized action is D, which means L's reputation in the next period will be $\Phi(\mu|D) = \epsilon$. State L will decide on its effort level σ_L by comparing its continuation payoffs described by these two terms in equation (3). Then, L's strategy σ_L is optimizing if $\sigma_L > 0$ implies that the first term is weakly greater than the second, $\sigma_L < 1$ implies that the second term is weakly greater than the first, and thus $0 < \sigma_L < 1$ implies that the first and the second term equal each other, indicating indifference on L's part.

The pair $(\sigma_S(\mu), \sigma_L(\mu))$ is called a *Markov Perfect Equilibrium* (MPE) if σ_S and σ_L respect the optimization problems of S and L respectively and are best-responses to each other for all μ.

4.4 Why Markovian Strategies?

In addition to significant gains in mathematical tractability, there are substantive reasons for requiring behavior to be Markovian in this setting. Focusing on Markovian strategies based on the reputation of L accords well with the perception in the literature that reputation is an asset which can be built, maintained, and spent (Schelling 1966). This restriction affords the analysis to most clearly focus on reputation dynamics: how L's reputation, formed via S's learning and updating based on observed behavior, conditions the behavior of both L and S. It eliminates strategies that depend on the history of play outside of the belief channel, such as the retaliatory strategies discussed earlier which do not capture the phenomenon of interest.

Markovian equilibria have also been shown to have desirable properties from the perspective of psychology and bounded rationality, in that they are robust to small perturbations of stage game payoffs, relaxation of infinite recall, forgetting the period index, among others (Bhaskar, Mailath, and Morris 2013; Faingold and Sannikov 2011; Mailath and Morris 2002, 2006). IR scholars, including reputation critics, have previously raised concerns about whether the high degree of cognitive sophistication signaling games demand of their agents is appropriate for real-world policymakers (Downs and Jones 2002; Jervis 2002; Mercer 1996). The much lower cognitive load required by Markovian strategies should alleviate some of these concerns.

5 Results

5.1 Repeated game with no uncertainty

First, consider the case where there is no uncertainty regarding L's type. That is, S knows that L is the normal type with certainty and it will stay that way: $\mu_0 = \epsilon = 0$. Recall that the unique NE of the stage game involves S staying out in the cooperation case (Assumption 3A) and entering in the conflict case (Assumption 3B), while L exerts no effort in both cases. Trivially, and given the strategies described in the previous section, the unique MPE of the repeated game will feature the indefinite repetition of the stage-game NE. This is because L has no incentive to exert costly effort, since doing so will not lead to an improvement in its reputation, and thus no improvement in future payoffs. Given this, S has no incentive to enter in the cooperation condition, or stay out

in the conflict condition. Therefore, under no uncertainty and no reputational incentives, there is no cooperation (we never observe the outcome (E, I)). Also, while there is no costly conflict (i.e. we never observe (E, I)), there is also nothing L can do to prevent facing threats from S. As such, L bares the brunt of anarchy. This reiterates the point that uncertainty is necessary for reputational incentives, and without reputational incentives, anarchy makes commitment problems acute.[16]

5.2 Repeated game with uncertainty and fixed types

Now suppose that S is uncertain about L's type, that is, $\mu_0 > 0$. However, L's type is drawn at the beginning of the game and remains fixed throughout, which means $\epsilon = 0$ and $\lambda = 1$. The fact that L's type remains fixed implies that, if L reveals itself to be the normal type by playing D at any point, the equilibrium play in the continuation game starting the next period will revert to the unique MPE of the no uncertainty case, which is the infinite repetition of the unique Nash equilibrium of the stage game. This means $V(\mu = 0) = \frac{u_L(O,D)}{1-\delta}$ in the cooperation case and $V(\mu = 0) = \frac{u_L(E,D)}{1-\delta}$ in the conflict case. Recognizing these prospects, L will either decide that mimicking the commitment type is worth it and will never play D to avoid facing the continuation value $V(\mu = 0)$, or L will decide that mimicking the commitment type is not worth it and will never exert any effort. Given Assumptions 1-3, whether L will find pretending to be the commitment type worthwhile will depend on whether L is forward looking enough. In both conflict and cooperation conditions, the MPE has the following familiar structure:

Proposition 0. *Let $\epsilon = 0$, $\lambda = 1$, and $0 < \mu_0 < \mu^*$. There exists a threshold $\bar{\delta} \in (0, 1)$, where if $\delta \leq \bar{\delta}$ the unique MPE of the game is the repetition of the stage-game Nash equilibrium for all μ. If $\delta > \bar{\delta}$, then in the unique MPE of the game L exerts maximum effort $\sigma_L(\mu) = 1$ in both the cooperation and conflict conditions; while S plays E with certainty $\sigma_S(\mu) = 1$ in the cooperation condition and O with certainty $\sigma_S(\mu) = 0$ in the conflict condition for all $\mu > 0$. At $\mu = 0$, $\sigma_L(\mu = 0) = 0$ and $\sigma_S(\mu = 0) = 0$ in the cooperation condition and $\sigma_S(\mu = 0) = 1$ in the conflict condition.*

I state this proposition without proof, since this is the standard reputation result which shows that even a tiny amount of uncertainty about L's type $\mu_0 > 0$ can generate reputational incentives

16. There are, of course, many non-Markovian equilibria depending on trigger strategies which can yield Pareto superior outcomes for L and S in cooperation, and better outcomes for L in conflict. As discussed above, I am assuming away such behavior to focus on reputation dynamics and how those alleviate commitment problems.

that would significantly change the equilibrium play to L's benefit (Kreps and Wilson 1982; Milgrom and Roberts 1982). Note that, in the presence of reputational incentives and when L's type is fixed, L is able to secure in equilibrium the payoff it would receive if anarchy did not apply, that is $V(\mu) = \frac{u_L(E,I)}{1-\delta} > \frac{u_L(O,D)}{1-\delta}$ in the cooperation condition and $V(\mu) = \frac{u_L(O,I)}{1-\delta} > \frac{u_L(E,D)}{1-\delta}$ in the conflict condition. More importantly, the amount of cooperation observed in equilibrium is the same as in the case where L could credibly commit to playing I via writing an enforceable contract. Similar to the no-uncertainty case, there is no conflict on the equilibrium path, however L is better-off for being able to commit to always play I via reputational incentives, since it never faces threats in equilibrium (S never plays enter). If the stage game is represented by the examples in Figure 1, then for $\delta > \bar{\delta} = \frac{1}{2}$, the MPE of the game will feature infinite repetition of (E, I) in cooperation and (O, I) in conflict.

Note that, on the equilibrium path, the actions taken by players do not change. At each interaction, the behavior of both L and S are constant, and the outcome of each stage game is identical. Further, the equilibrium-path behavior features no reputation building or spending, that is, if the game starts with a reputation of μ_0 for L, this reputation remains constant for all periods t. This is because the normal type of L perfectly mimics the commitment type of L at all times, since failing to do so will destroy its reputation forever. Thus, the audience learns nothing about L's type when they observe L exerting effort. This is stated in the following corollary to Proposition 0:

Corollary 0.1. *On the equilibrium path of the MPE described in Proposition 0:*

- *The actions of L and S remain constant.*
- *L's reputation remains constant and equal to the prior μ_0 at the beginning of the game.*

The corollary follows trivially from Proposition 0. This reiterates the point that, when L's type is assumed to be fixed, reputation dynamics are not dynamic at all; they are static. They fail to display the asset-like characteristics of reputations discussed in the IR literature, where reputations can be built, maintained, and spent. Nevertheless, Proposition 0 is the standard reputation result on which IR scholars rely on, implicitly or explicitly, when they assume that a state will fulfill its commitments because of reputational reasons (e.g. Büthe and Milner 2008; B. A. Simmons 2000; Walter 2006).

Moreover, as mentioned earlier, this equilibrium is very fragile. Cripps, Mailath, and Samuelson (2004) show that when L's actions are not perfectly monitored by S, eventually there comes a point where L's reputation is so established that the temptation to deviate and not exert any effort becomes irresistible for L. Over time, such deviations by L from the commitment action enables S to statistically identify L's type, destroying the underlying uncertainty and thus reputational incentives. They argue that in the absence of a mechanism to replenish the uncertainty which lies at the heart of the reputation mechanism, reputations are a short-run phenomenon. With this in mind, next I analyze reputation dynamics when L's type is allowed to change.

5.3 Reputations with changing types

Consider the general case where S is uncertain about L's type, and it is common knowledge that L's type can change. Maintaining Assumptions 1-4, Proposition 1 below describes the MPE of the game, which is also the main result:

Proposition 1. *Let $\mu^k = \Phi(\mu^{k-1}|I)$ for $k \in \{0, 1, ..., N, N+1, ...\}$, where $\mu^0 = \Phi(\mu|D) = \epsilon$, $\mu^{N-1} = \Phi(\mu^{N-2}|I) < \mu^*$ and $\mu^N = \Phi(\mu^{N-1}|I) > \mu^*$, denote L's reputation after k successive realizations of I. Define $A = \frac{u_L(E,D)-u_L(E,I)}{u_L(E,D)-u_L(O,D)}$, $B = \frac{u_L(O,D)-u_L(O,I)}{u_L(E,D)-u_L(O,D)}$, and $\bar{\delta} = max\{\frac{A-B}{(1-\epsilon)}, \frac{|B|}{(1-\epsilon)}\}$. Then if $\delta > \bar{\delta}$, then for all μ, the pair $(\hat{\sigma}_S, \hat{\sigma}_L)$ describes a Markov Perfect Equilibrium:*

- *For both the cooperation and conflict conditions, the following describes L's equilibrium strategy:*

$$\hat{\sigma}_L(\mu) = \frac{\mu^* - \mu}{1 - \mu} \quad \forall \mu < \mu^* \quad \text{and} \quad \hat{\sigma}_L(\mu) = 0 \quad \forall \mu \geq \mu^*$$

- *If the cooperation condition applies, then for $\mu \geq \mu^*$ (that is, $\mu^k \geq \mu^N$), we have $\hat{\sigma}_S(\mu) = 1$. For $\mu < \mu^*$ (that is, $\mu^k < \mu^N$) we have:*

$$\hat{\sigma}_S(\mu^k) = \left(\frac{\Sigma_{i=0}^{k}(\frac{A-B}{\delta(1-\epsilon)})^i}{\Sigma_{i=0}^{N}(\frac{A-B}{\delta(1-\epsilon)})^i}\right) - \frac{B}{\delta(1-\epsilon)}\left(\frac{\Sigma_{i=k}^{N-1}(\frac{A-B}{\delta(1-\epsilon)})^i}{\Sigma_{i=0}^{N}(\frac{A-B}{\delta(1-\epsilon)})^i}\right)$$

- *If the conflict condition applies: then for $\mu \geq \mu^*$ (that is, $\mu^k \geq \mu^N$), we have $\hat{\sigma}_S(\mu) = 0$. For $\mu < \mu^*$*

(that is, $\mu^k < \mu^N$) we have:

$$\hat{\sigma}_S(\mu^k) = -\frac{B}{\delta(1-\epsilon)}\left(\frac{\Sigma_{i=k}^{N-1}(\frac{A-B}{\delta(1-\epsilon)})^i}{\Sigma_{i=0}^{N}(\frac{A-B}{\delta(1-\epsilon)})^i}\right)$$

The proof is in the appendix. Here, I will walk the reader through the logic behind the equilibrium behavior. First recall that if L's reputation is high enough ($\mu \geq \mu^*$), S will enter in cooperation and stay out in conflict, regardless of the strategy of the normal type of L. For a conflict example, suppose S knows for certain that, if L is a weak target, it will definitely acquiesce when threatened. Even here, if S believes L to be strong with sufficiently high probability, it will not threaten L, to avoid costly conflict. Now suppose that, if the weak type of L will acquiesce with certainty, L's reputation is not strong enough to keep S from threatening L ($\mu < \mu^*$). Then, the weak type of L knows that if it is willing to exert enough effort to compensate for this reputation gap, it can still induce S to stay out. In equilibrium, "enough effort" corresponds to the amount of effort or the probability of playing I that would make S indifferent between staying out and entering. Then S will think: "I doubt L is the strong type, which bodes well for me to demand concessions. That said, I know that the weak type of L is willing to exert effort to appear strong. This means, regardless of whether L is weak or strong, I will have to face resistance. Therefore, perhaps I should stay out." In cooperation, the same considerations apply, except that S wants L to exert effort and L wants S to enter. Therefore S will enter with certainty when $\mu \geq \mu^*$ and L will adjust its effort to induce S to enter when $\mu < \mu^*$, or more correctly, induce indifference on S between entering and staying out.

To summarize, when $\mu < \mu^*$, the normal type of L will adjust its effort level in a way that would balance its reputation in order to induce indifference on S between entering and staying out. If $\mu \geq \mu^*$, the normal type of L loses the incentive to exert any effort, since no effort is required for S to act in the way L prefers. Why would L prefer damaging its reputation when it exerted so much costly effort to push it beyond μ^*? To understand this, we need to first discuss S's strategy.

Recall that, as Assumption 1 states, effort is costly for the normal type of L. If exerting effort will not bring any downstream benefits to L, then L would never play I, as can be seen in the no-uncertainty case. In other words, L needs to get something in return if it is to exert effort. S will adjust its strategy such that it will induce L to take the desired action (I in cooperation, D in

conflict). In equilibrium, this takes the form of inducing L to be indifferent between I and D. The strategy of S, $\hat{\sigma}_S$, provided in Proposition 1 implies that:

$$0 < \hat{\sigma}_S(\mu^0) < \hat{\sigma}_S(\mu^1) < ... < \hat{\sigma}_S(\mu^N) = 1 \qquad \text{(Cooperation)}$$

$$1 > \hat{\sigma}_S(\mu^0) > \hat{\sigma}_S(\mu^1) > ... > \hat{\sigma}_S(\mu^N) = 0 \qquad \text{(Conflict)}$$

Where μ^k for $k \in \{0, 1, ..\}$ is as defined in Proposition 1. This means S is increasingly willing to enter in cooperation and less willing to enter in conflict, as L's reputation improves. At each step, the increase in the probability that S will enter in cooperation, and increase in the probability that S will stay out in conflict, is the reward L requires to continue exerting effort for its reputation.

Why are these "rewards" not sufficient to make L want to protect its reputation, once its reputation reaches the threshold μ^*? To see this, note that once $\mu \geq \mu^*$, that is, once $\mu^k \geq \mu^N$, state S will play enter in cooperation and stay out in conflict with certainty. In turn, if L continues playing I, the continuation game looks exactly the same for $k \geq N$. In order for L to continue exerting effort, L should prefer bearing the cost of effort indefinitely without getting rewarded by S with a higher probability of entering or staying out to spending its reputation today and starting from $V(\mu^0)$ tomorrow.

When types are fixed, L would prefer exerting effort indefinitely in this situation, because then spending reputation means switching to the MPE of the no-uncertainty case forever, which is strictly worse than indefinitely exerting the cost of effort. However, when types are known to change, even when L destroys its reputation today, tomorrow S will think that L is a commitment type with probability ϵ, which enables reputation rebuilding. S will still enter with some probability ($\hat{\sigma}_S(\mu^0) > 0$) in cooperation, and stay out with some probability in conflict ($\hat{\sigma}_S(\mu^0) < 1$), making the continuation value $V(\mu^0)$ strictly better than reverting to the MPE of the no-uncertainty case. Once L's reputation reaches the threshold μ^*, S is unable to "reward" L further, at which point L finds it irresistible to spend its reputation today and start rebuilding tomorrow.

On the equilibrium path, the evolution of L's reputation features stretches of reputation building followed by spending behavior, as such, reputations will tend to appear cyclical. Figure 2 displays the evolution of L's reputation in a scenario consistent with the equilibrium dynamics

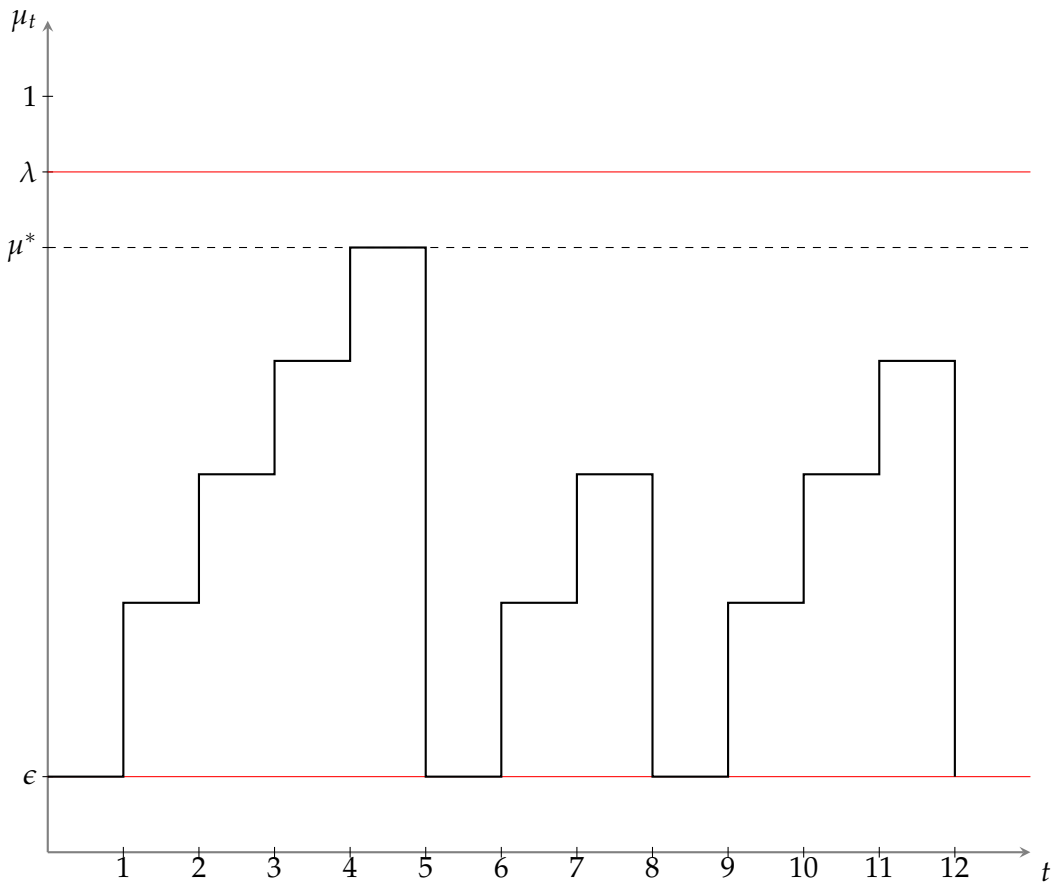

Figure 2: Evolution of L's reputation in a scenario consistent with the equilibrium dynamics. Improvements and drops in reputation are the result of S's updating upon observing I and D respectively.

discussed here. However, note that because both S and L play mixed strategies when $\mu < \mu^*$, destruction of L's reputation can happen at any point.

In the stage game examples presented in Figure 1 Assumption 2 holds with equality, in which case the equilibrium dynamics are simpler. In particular, there is a one-step transition period between the period immediately after L damages its reputation where $\mu = \mu^0 = \epsilon$, and the period at which L damages its reputation again with certainty. Suppose the transition probabilities obey Assumption 4 and $\delta > 0.5$. Then the equilibrium will feature two phases I label "doubt" and "trust," and the play will continually alternate between the two phases. In the doubt phase, S believes L to be likely weak or unreliable, and thus enters with low probability in cooperation and high probability in conflict, while L is willing to exert some effort to build its reputation. In the trust phase, S believes L to be likely resolved or reliable, and is willing to enter with certainty

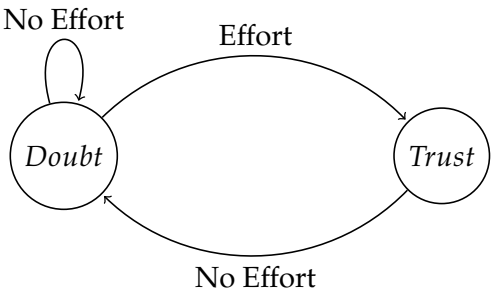

Figure 3: Illustration of the equilibrium dynamics, when stage game payoffs are described by the examples in Figure 1. "No Effort" and "Effort" corresponds to the realization of D and I respectively in a given period.

in cooperation and stay out with certainty in conflict, whereas L wants to take advantage of its favorable reputation and exerts no effort. Figure 3 displays this equilibrium structure.

The discussion so far indicates that the temptation of L to take advantage of its reputation is increasing in cooperation as the probability that S enters is increasing, and in conflict it is increasing as the probability that S stays out is increasing. At the same time, S is willing to enter with higher probability in cooperation and stay out with higher probability in conflict the more times it observes L investing in its reputation. Then, L's incentives to invest in its reputation is decreasing in its reputation. The following corollary to Proposition 1 states this formally:

Corollary 1.1. *In the MPE of the game described in Proposition 1:*

- *The probability that the normal type of L exerts effort is weakly decreasing in its reputation for all μ.*
- *The probability that the normal type of L exerts effort is strictly decreasing in its reputation for $\mu < \mu^*$.*

The result follows directly from looking at the first order condition of $\hat{\sigma}_L$ with respect to μ in Proposition 1. This suggest that the most pronounced behavioral impact generated by reputational incentives happens at lower reputation levels.

Furthermore, because L is able to induce its desired behavior on S via leveraging its reputation, L benefits from the presence of reputational incentives. In other words, compared to the counterfactual situation where L is unable to make credible commitments, L is strictly better-off under the reputational equilibrium. However, because type changes eventually introduce an irresistible temptation for L to take advantage of its reputation, we observe less cooperation and more costly

conflict (i.e. the outcome (E, I)) on the equilibrium path compared to the counterfactual situation where L is able to make credible commitments via writing an enforceable contract. This is stated in the next proposition:

Proposition 2. *On the equilibrium path of the MPE described in Proposition 1, there is less cooperation and more conflict, defined as the stage game outcome (E, I), compared to (i) the unique MPE of the game with fixed types, and (ii) the counterfactual world where L can credibly commit to exerting effort via binding contracts.*

From the definition of $\hat{\sigma}_S$ and $\hat{\sigma}_L$, the probability of observing (E, I) at any period in either the conflict or cooperation condition is $0 < \hat{\sigma}_S\hat{\sigma}_L < 1$ for $\mu < \mu^*$ and $\hat{\sigma}_S\hat{\sigma}_L = 0$ for $\mu \geq \mu^*$. This means, in the cooperation condition, $\hat{\sigma}_S\hat{\sigma}_L < 1$ for all μ, which means there is less cooperation compared to the no-anarchy and fixed-types cases. In the conflict condition, since $0 < \hat{\sigma}_S\hat{\sigma}_L$ for $\mu < \mu^*$, there is more conflict compared to the no-anarchy and fixed-types cases.

Finally, I conjecture that the MPE described in Proposition 1 is the unique MPE of the game. The arguments for uniqueness follow Lemmas 2-7 and Theorem 8 of Phelan (2006), but those arguments do not directly apply, because in Phelan (2006) the government (long-run player) can costlessly signal when no citizen (the short-run players) produces any output (similar to S staying out), whereas here signaling is always costly by Assumption 1. A potential strategy to prove uniqueness can be as follows: In any MPE of the game, following an observation of D, S will always enter with positive probability in cooperation and will not enter with certainty in conflict. If it did, this would mean that S thinks the weak type of L will exert less effort than is stated in the MPE discussed above, which in turn means that L should be able to improve its reputation much faster by playing I. This brings the period at which $\mu > \mu^*$ closer compared to the MPE above, and should produce a contradiction where S does not enter (in cooperation) or stay out (in conflict) with certainty notwithstanding $\mu > \mu^*$.

Similarly, following an observation of D, S will not enter with certainty in cooperation and it will enter with positive probability in conflict. If this was not true, then after destroying its reputation, L could still reap the benefit of its desired behavior from S without having to exert any effort, which contradicts S's optimization. Therefore both S and L should play a mixed strategy in the period after D is observed. The next step would be to show via a similar argument using induc-

tion that when $0 < k < N$ both S and L should play mixed strategies. Since the Bellman equations describing L's continuation values (in the appendix) are full rank and thus have a unique solution, the behavior in any MPE of the game should be the same as the MPE in Proposition 1.

5.4 A Note on Assumption 3

Equilibrium dynamics in Proposition 1 depend on the following assumption:

$$u_L(E, D) - u_L(O, D) \geq u_L(E, I) - u_L(O, I) \qquad \text{(Cooperation)}$$

$$u_L(O, D) - u_L(E, D) \geq u_L(O, I) - u_L(E, I) \qquad \text{(Conflict)}$$

This is Assumption 3Aii for the cooperation condition and 3Bii for the conflict condition, which means L's gain from inducing S to perform L's desired behavior (enter in cooperation, stay out in conflict) is weakly greater when L does not exert any effort. Alternatively, this can also mean that L's cost of signaling is weakly higher when S enters in cooperation and stays out in conflict. If this assumption is violated, and if my conjecture about the uniqueness of the equilibrium is correct, then there is no MPE where strategies of L and S depend only on L's reputation. In other words, strategies should in some way depend on the history of play outside of the channel of L's reputation. This assumption is violated, for instance, if L is not allowed to signal when S stays out, such as in the classic chain-store game. Wiseman (2008) analyzes an infinitely repeated chain-store game where the long-run player's type is subject to change, which provides clues about what equilibrium dynamics would 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. The strategies then 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 eqilibrium

strategies are thus 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 investing in its reputation is an increasing function of its current reputation, unlike in Corollary 1.1. 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 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.

6 Discussion

6.1 Reputation as an Enforcement Mechanism under Anarchy

In IR scholarship, especially with regard to international cooperation, existence of reputational incentives is thought to constrain states similar to the way contractual institutions with enforcement power would constrain agents in the domestic realm. In other words, reputation is a key "enforcement" mechanism in IR (Keohane 1984; B. Simmons 2010). The preceding analysis shows that, when states' preferences are subject to change, reputation dynamics feature both reputation building and spending behavior in equilibrium. Importantly, if my uniqueness conjecture is correct, there is no MPE where states always invest or divest in their reputations. Therefore, as Proposition 2 suggests, if anarchy means that states have to rely on reputational incentives to make their commitments credible, then the level of cooperation which can be sustained under anarchy is strictly less compared to what can be achieved via contractual institutions with enforcement power. This stands in contrast with the prevailing static perspective on reputations in IR, where reputational incentives can support first-best outcomes in cooperation. On the conflict side, the same result implies that anarchy is associated with more conflict if reputational incentives are generated by changing as opposed to fixed preferences. These results can be seen as complementary to Fearon (2018).

A comparison of different mechanisms through which commitment problems can be resolved

in the international arena is beyond the scope of this paper. That said, folk theorem type arguments suggest that reciprocity or community punishment, the second major "enforcement" mechanism in IR, could be used to achieve first best outcomes in cooperation (Rubinstein 1979; Fudenberg and Maskin 1986). Although implementing reciprocity-based mechanisms involve coordination and free riding problems, research suggests that small societies, of which the international society is an example, can effectively solve their collective action problems (Olson 1965; Ostrom 1990). Therefore, one can speculate based on the results of this paper that international institutions could potentially achieve better compliance levels by using their resources toward strengthening reciprocity mechanisms over generating reputational incentives, such as via rating systems.[17] That said, reputational incentives do constrain and are likely less costly to generate from the perspective of an institution with limited resources. A fruitful area for future research would be to look at this trade-off from a mechanism design perspective.

6.2 Who is Constrained by Reputations?

Reputational incentives may not be able to sustain the same level of cooperation as can be sustained under contractual institutions with enforcement power, but they do constrain. This is implied by the comparison of equilibria under the no-uncertainty case and under reputations with changing types. Who, then, is constrained by reputational incentives?

The question of who is constrained by reputations is irrelevant if states' preferences and abilities are assumed to be immutable, which tends to be either implicitly or explicitly assumed in IR scholarship. As is stated in Corollary 0.1, if preferences are fixed, then the behavior is static: states either always invest in their reputations or they never do. When preferences are known to change, Corollary 1.1 indicates that states with poor reputations are more likely to invest in their reputations compared to states with better reputations.

In international cooperation, reputational incentives improves the welfare of all parties involved by making mutually beneficial outcomes attainable. However, since the constraining power of reputational incentives is greater for states with poor reputations and weaker for states with good reputations, this improvement in welfare is produced by the agents who are at the left

17. See Kelley (2017) for an example of an international rating system devised by the U.S. State Department ranking states based on their policies on human trafficking.

tail of the reputation distribution. To put it differently, reputation mechanism is a double-edged sword in cooperation. While it makes mutually beneficial outcomes attainable, this comes at the cost of letting states with good reputations to take advantage of the trust their reputations generate.

In international conflict, the substantive picture is different. From a guns-butter trade-off perspective, a state exerting effort for its reputation can be seen as diverting resources to economically unproductive (beyond the instrumental value) avenues for its society such as armaments (e.g. Powell 1993). The results in this paper suggest that states with poor reputations, e.g. those which are perceived as weak, are more willing to bare this burden. Further, because changing preferences induces permanent doubt, states cannot fend-off threats permanently, as their mettle will be periodically tested.

If reputational incentives lose their constraining power on states with better reputations, this mechanism could underpin instances of surprising behavior in international conflict and cooperation. This covers both surprising displays of resolve or surprising defeats and unexpected cooperative behavior and betrayals of trust. For instance, many expressed confusion when Norway, a reputed stalwart for environmental protection led by Prime Minister Gro Harlem Brundtland "The Green Queen," announced the resumption of whale hunting in 1992.[18] This happened notwithstanding the protests of the International Whaling Commission, of which Norway was a party. A study on Norwegian public diplomacy commissioned by the Norwegian Foreign Ministry, suggests that Norway's favorable reputation regarding environmental protection might have played an enabling role (Leonard and Small 2003). Similarly, reputational incentives could rationalize why, in the years running up to its crushing defeat in 1870, the French army had the most stellar reputation in the world while at the same time has been commonly described as widely complacent by historians — the latter point being obvious only ex-post.[19] Prior to the Franco-Prussian War of 1870, there is evidence that even high ranking Prussian generals thought that the French army might have been better than that of Prussia.[20]

18. https://nyti.ms/298uR8G

19. The French army's reputation was built over numerous successful overseas campaigns, in Crimea, and in the Italian Wars. This can also be observed in the fact that French officers were in high demand among states who wanted to modernize their militaries, such as Japan and Turkey, both of which switched to German officials following 1870. For a long list of reasons why French army was inferior than that of Prussia before 1870, see e.g. Wawro (2003, chapter 2).

20. For instance, Prinz Friedrich Karl, a prominent Prussian aristocrat and general who also served in the 1870-71 Franco-Prussian War, admits that the French might have a better army than the Prussians in the afterword of an 1860

Regarding unexpected cooperative behavior, Tomz (2007) shows that states which paid their debts despite the expectation among international lenders toward default, had a significantly easier time accessing credit when they wanted to borrow again. Tomz (2007) cites examples such as Finland and Argentina during the Great Depression. Another example is when the fledgling US government assumed debt obligations of its bankrupt states in 1790, with the explicit goal of cultivating a favorable reputation for the US in international lending markets. In the words of Alexander Hamilton, the architect of the 1790 bailout, stopping debt repayments "cannot but have a malignant influence upon our public and mercantile credit... Every gust that arises in the political sky is the signal for measures tending to destroy [our] ability to pay or to obstruct the course of payment (quoted in Irwin 2011, 111)." The results in this paper suggest that such behavior is an integral part of reputation dynamics, and thus does not necessitate significant domestic political change as is argued in Tomz (2007).

Finally, if states with good reputations have the least incentive to further invest in them, it might also be the case that they have greater incentives to develop technologies to hide their undesirable behavior from their audiences. This parallels, for instance, findings in Kono (2006) where democracies, which otherwise have liberal trade policies, tend to erect non-tariff barriers to trade, because they are more difficult to detect. Similarly, Bazillier, Hatte, and Vauday (2017) show that multinational firms with well-established reputations for environmental responsibility are more likely to have locations in countries with laxer environmental regulations. The relationship between reputations and obfuscation of "bad" behavior is a potentially fruitful area for future research.

6.3 The Role of Shifting Uncertainty

The idea that states' preferences and abilities are subject to change is the main ingredient behind all of the results presented in this paper. In the model, these changes are not directly observed by the international audience. Rather, the audience continually suspects that states' preferences might have changed unbeknown to them. This shifting uncertainty can be interpreted as reflecting the underlying instability of state preferences over time, paralleling the Realist claim that states' intentions can never be known (Mearsheimer 2001; Waltz 1979). Alternatively, it can be interpreted as

pamphlet *Ueber die Kampfweise der Franzosen*.

reflecting the changes in what is at stake across different interactions, and thus the comparability of observed behavior from one case to another.

Proposition 0 describes the equilibrium when there is uncertainty about state preferences, but when those preferences are assumed to be fixed. This is what I label as the static perspective on reputations, and can be thought of as reflecting maximum stability in state preferences ($\epsilon = 0$, $\lambda = 1$). Here reputational incentives have the most constraining power because once a state's reputation is tarnished, there is no chance for rebuilding. Contrast this with what happens when Assumption 4 is violated, that is, when state preferences are too volatile. In that case, there is no value in the data pertaining to yesterday's interactions, because today's interactions will be so different than yesterday's that there is no point on acting upon what is learned from those data. Therefore, too much instability destroys reputational incentives.

Further, note that although reputations have the most constraining power on state behavior when preferences are fixed, this equilibrium has been shown to be fragile. In particular, (Cripps, Mailath, and Samuelson 2004) show that reputational incentives cannot be sustained in the presence of imperfect monitoring, as such, persistent reputational incentives require persistent uncertainty. I have not shown the effect of relaxing the perfect monitoring assumption on the equilibrium behavior presented in Proposition 1, when preferences are allowed to change. However, the reason why the static equilibrium is fragile with imperfect monitoring is because of the inability to rebuild reputations when preferences are fixed. Therefore the shifting uncertainty assumption in this paper should prevent the equilibrium presented here from unraveling under imperfect monitoring. If this claim holds, then this would mean that both too much and too little stability would destroy reputational incentives. If so, this would suggest that the very reason why structural Realists like Mearsheimer (2001) believe anarchy to be tragic —the inherent inscrutability of other states' intentions— is also what gives rise to permanent reputational concerns. In other words, what makes power politics tragic also carries with it the seeds of cooperation.

6.4 Reputation Debate in IR

The prevailing static perspective on reputations in IR is silent about how incentives to invest in reputations change as a function of reputations, because when preferences are fixed so is state be-

havior. Lacking a theoretical understanding about how the value of building reputations changes with current reputations has led some scholars to question the analytical purchase of the reputation mechanism in IR by highlighting a set of so-called paradoxes. Jervis (1997) argues that statesmen who are aware of the reputation mechanism will work extra hard —say, after defeats— to prove the theory wrong. If those with worse reputations can work harder, what information reputations convey, if any? Therefore, whether reputation mechanism works in IR should depend on whether leaders believe it to be true (Jervis 1997; Monteiro 2012). In similar fashion, Press (2005) says while reputation-based deterrence theories would expect others to think an actor with a blemished reputation to be weak or distrustful, we often find the opposite. According to his *Never Again Theory*, when a country backs down from a crisis, it can actually have more credibility in the eyes of the others, as backing down again will be even more costly.

Jervis (1997, 270) further argues that if an actor builds a reputation for being resolved, cooperative, or honest, then it should be able to afford some behavior to the contrary. Mercer (2013, 224) takes it a step further: "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."

All of the arguments raised by reputation critics mentioned above are in fact consistent with the reputation mechanism, when we relax the assumption that state preferences are immutable. According to Corollary 1.1, states with better reputations are less likely to exert further effort for their reputations. This means, consistent with Jervis (1976) and Press (2005) a state with a blemished reputation should work harder to rebuild its reputation. This also means that a state with a stellar reputation is more likely to take advantage of its reputation compared to a state with a worse reputation, consistent with Jervis (1976) and Mercer (2013). Further, given that, except L's preferences (which is private information), everything in the model is common knowledge and that in equilibrium L and S play best responses to each other, the theory relies on policymakers being aware of the reputation mechanism, addressing Jervis (1997) and Monteiro (2012).

Note that this specifies the behavior of a state which is facing a commitment problem (the normal type of L), because those are the actors the behavior of which is impacted by reputational

incentives. In other words, this is the behavior of a —say— weak state, deciding how much effort to expend in order to appear strong depending on the degree to which the others think it is strong. From the perspective of others, the uncertainty about whether the state is weak or strong is never resolved — in fact, the weak state adjusts its effort level to sustain this indeterminacy. If the state has a poor reputation, the audience correctly believes that the state is likely not strong, but they also know that if the state is weak, it will work hard to cultivate its reputation. This latter part is the reason why Press (2005)' *Never Again Theory* is consistent with the reputation mechanism. Similarly, if a state has a good reputation, the audience correctly believes that the state is likely strong, but if the state happens to be weak, it will take advantage of this reputation. This is why Mercer (2013)'s point is not a paradox, but is consistent with the reputation mechanism here.

7 Conclusion

The prevalence of commitment problems is a defining characteristic of the international domain. States' inabilities to convince partners and adversaries that they will follow through on their promises leads to both inefficient conflicts and creates difficulties in attaining mutually beneficial cooperation. IR scholars have long identified reputational concerns as the primary vehicle through which states resolve their commitment problems. Yet, the field has a static perspective on reputations, because our reputational theories take states' preferences as fixed. As such, IR lacks theories of reputation building and spending. Further, we do not know how current reputations shape incentives to maintain them: should reputations be more valuable for states with better or worse reputations?

In this paper, I offer a dynamic model of reputations covering both conflictual and cooperative interactions, where I relax the assumption that states' preferences are immutable. I explain why states tarnish their hard-earned reputations, how tarnished reputations are rebuilt, and more generally, how states' current reputations shape their incentives to build future reputations. When states' preferences are allowed to change over time and across interactions, this restricts how much an audience can infer states' future behavior based on observations of their past behavior. Knowing that preferences change leads the international audience to doubt those with good reputations and extend the benefit of the doubt to those with poor reputations, because the audience has rea-

son to believe that things might be different *this* time. Doubt limits the usefulness of reputations for their holders, and benefit of the doubt opens up the possibility of rebuilding tarnished reputations. Then, states which are motivated by reputational concerns exert less effort when their reputations improve and more effort when their reputations deteriorate.

These results have a number of implications for the study of conflict and cooperation in IR. First, if states have to rely on the reputation mechanism to make credible commitments under anarchy, then the amount of cooperation which can be sustained under anarchy is strictly less and the amount of conflict we observe is strictly more, compared to if states could credibly commit via writing enforceable contracts.

Second, reputational incentives do not constrain all states equally; those with worse reputations are constrained more than those with better reputations. In cooperation, this means that the reputation mechanism is a double-edged sword. While it makes mutually beneficial outcomes attainable, this comes at the cost of letting states with good reputations to take advantage of the trust their reputations generate. In conflict, if a state investing in its reputation means diverting resources to economically unproductive avenues for its society as the guns-butter tradeoff perspective would suggest, then states with poor reputations, e.g. those which are perceived as weak, are more willing to bare this burden.

Third, providing a theoretical understanding for how reputational incentives change as a function of current reputations addresses a number of problems raised by IR scholars about the reputation mechanism. These include the *Never Again Theory* of Press (2005), *Domino Theory Paradox* of Jervis (1997), and the argument in Mercer (2013) that states with better reputations should afford some behavior to the contrary. I show that these points, presented by their authors as paradoxes, are shortcomings of the static perspective on reputations. They are, however, consistent with the rational reputation logic outlined in this paper.

References

Acemoglu, Daron. 2003. "Why not a Political Coase Theorem? Social Conflict, Commitment, and Politics." *Journal of Comparative Economics* 31 (4): 620–652.

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.

Axelrod, Robert, and Robert O. Keohane. 1985. "Achieving Cooperation under Anarchy: Strategies and Institutions." *World Politics* 38 (01): 226–254.

Bazillier, Rémi, Sophie Hatte, and Julien Vauday. 2017. "Are environmentally responsible firms less vulnerable when investing abroad? The role of reputation." *Journal of Comparative Economics* 45 (3): 520–543.

Bhaskar, V., George J. Mailath, and Stephen Morris. 2013. "A Foundation for Markov Equilibria in Sequential Games with Finite Social Memory." *Review of Economic Studies* 80 (3): 925–948.

Board, Simon, and Moritz Meyer-ter-Vehn. 2013. "Reputation for Quality." *Econometrica* 81 (6): 2381–2462.

Bohren, J Aislinn. 2013. "Stochastic Games in Continuous Time : Persistent Actions in Long-Run Relationships."

Büthe, Tim, and Helen V. Milner. 2008. "The Politics of Foreign Direct Investment into Developing Countries: Increasing FDI through International Trade Agreements?" *American Journal of Political Science* 52 (4): 741–762.

Cho, In-Koo, and David M. Kreps. 1987. "Signaling Games and Stable Equilibria." *Quarterly Journal of Economics* 102 (2): 179–221.

Cole, Harold L., James Dow, and William B. English. 1995. "Default, Settlement , and Signalling : Lending Resumption in a Reputational Model of Sovereign Debt." *International Economic Review* 36 (2): 365–385.

Cole, Harold L., and Patrick J. Kehoe. 1998. "Models of Sovereign Debt: Partial Versus General Reputations." *International Economic Review* 39 (1): 55.

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

Cripps, Martin, George J. Mailath, and Larry Samuelson. 2004. "Imperfect Monitoring and Impermanent Reputations." *Econometrica* 72 (2): 407–432.

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

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

Ekmekci, Mehmet, Olivier Gossner, and Andrea Wilson. 2012. "Impermanent types and permanent reputations." *Journal of Economic Theory* 147 (1): 162–178.

Faingold, Eduardo, and Yuliy Sannikov. 2011. "Reputation in Continuous-Time Games." *Econometrica* 79 (3): 773–876.

Fearon, James D. 1994. "Domestic Political Audiences and the Escalation of International Disputes." *American Political Science Review* 88 (03): 577–592.

———. 1995. "Rationalist Explanations for War." *International Organization* 49 (3): 379–414.

———. 1997. "Signaling Foreign Policy Interests." *Journal of Conflict Resolution* 41 (1): 68–90.

Fearon, James D. 2018. "Cooperation, Conflict, and the Costs of Anarchy." *International Organization* 72 (03): 523–559.

Fudenberg, Drew, and David K. Levine. 1989. "Reputation and Equilibrium Selection in Games with a Patient Player." *Econometrica* 57 (4): 759.

———. 1992. "Maintaining a Reputation when Strategies are Imperfectly Observed." *The Review of Economic Studies* 59 (3): 561.

Fudenberg, Drew, and Eric Maskin. 1986. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information." *Econometrica* 54 (3): 533–554.

Guzman, Andrew T. 2008. *How International Law Works: A Rational Choice Theory.* New York: Oxford University Press.

Hirshleifer, Jack. 1995. "Anarchy and its Breakdown." *Journal of Political Economy* 103 (1): 26–52.

Holmstrom, Bengt. 1999. "Managerial Incentive Problems: A Dynamic Perspective." *Review of Economic Studies* 66 (1): 169–182. arXiv: arXiv:1011.1669v3.

Irwin, Douglas A. 2011. "Revenue or Reciprocity? Founding Feuds over Early U.S. Trade Policy." In *Founding Choices: American Economic Policy in the 1790s,* edited by Douglas A. Irwin and Richard Sylla, 89–120. Chicago: University of Chicago Press.

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

———. 1997. *System Effects: Complexity in Political and Social Life.* Princeton: Princeton University Press.

———. 2002. "Signaling and Perception: Drawing Inferences and Projecting Images." In *Political Psychology,* edited by Kristen R. Monroe, 293–312. Mahwah: Lawrence Erlbaum.

Kelley, Judith. 2017. *Scorecard Diplomacy: Grading State to Influence Their Reputation and Behavior.* Cambridge University Press.

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

Kono, Daniel Y. 2006. "Optimal Obfuscation: Democracy and Trade Policy Transparency." *American Political Science Review* 100 (3).

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

Kydd, Andrew H. 2007. *Trust and Mistrust in International Relations.* Princeton: Princeton University Press.

Lake, David A. 1996. "Anarchy, Hierarchy, and the Variety of International Relations." *International Organization* 50 (01): 1.

Leonard, Mark, and Andrew T. Small. 2003. *Norwegian Public Diplomacy.* London: Foreign Policy Centre.

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.

Liu, Qingmin. 2011. "Information acquisition and reputation dynamics." *Review of Economic Studies* 78 (4): 1400–1425.

Liu, Qingmin, and Andrzej Skrzypacz. 2014. "Limited records and reputation bubbles." *Journal of Economic Theory* 151 (1): 2–29.

Mailath, George J., and Stephen Morris. 2002. "Repeated Games with Almost-Public Monitoring." *Journal of Economic Theory* 102 (1): 189–228.

———. 2006. "Coordination Failure in Repeated Games with Almost-Public Monitoring." *Theoretical Economics* 1 (3): 311–340.

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

Mearsheimer, John J. 2001. *The Tragedy of Great Power Politics.* New York: W. W. Norton.

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. "Predation, reputation, and entry deterrence." *Journal of Economic Theory* 27 (2): 280–312.

Milner, Helen V. 1991. "The Assumption of Anarchy in International Relations Theory: A Critique." *Review of International Studies* 17 (1): 67.

Monteiro, Nuno P. 2012. "We Can Never Study One Thing: Reflections on Systems Thinking in IR." *Critical Review* 24 (3): 343–366.

Olson, Mancur. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups.* Cambridge: Harvard University Press.

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action.* New York: Cambridge University Press.

Parker, Geoffrey. 1994. "No Title." Chap. The Making in *The Making of Strategy: Rulers, States, and War.* Cambridge.

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

Powell, Robert. 1993. "Guns, Butter, and Anarchy." *American Political Science Review* 87 (1): 115–132.

———. 1994. "Anarchy in International Relations Theory: The Neorealist-Neoliberal Debate." *International Organization* 48 (2): 313–344.

———. 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.

Ritter, Emily Hencken. 2014. "Policy Disputes, Political Survival, and the Onset and Severity of State Repression." *Journal of Conflict Resolution* 58 (1): 143–168.

Rubinstein, Ariel. 1979. "Equilibrium in Supergames with the Overtaking Criterion." *Journal of Economic Theory* 21 (1): 1–9.

Sargent, Thomas J. 2012. "Nobel Lecture: United States Then, Europe Now." *Journal of Political Economy* 120 (1): 1–40.

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

———. 2005. *Deterrence by Diplomacy.* Princeton: Princeton University Press.

Schelling, Thomas C. 1960. *The Strategy of Conflict.* Cambridge: Harvard University Press.

———. 1966. *Arms and Influence.* New Haven: Yale University Press.

Sechser, Todd S. 2010. "Goliath 's Curse: Coercive Threats and Asymmetric Power." *International Organization* 64 (4): 627–660.

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

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

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.

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

Waltz, Kenneth N. 1979. *Theory of International Politics.* New York: McGraw-Hill.

Wawro, Geoffrey. 2003. *The Franco-Prussian War: The German Conquest of France in 1870-1871.* New York: Cambridge University Press.

Weinstein, Jonathan, and Muhamet Yildiz. 2013. "Robust Predictions in Infinite-Horizon Games–an Unrefinable Folk Theorem." *Review of Economic Studies* 80 (1): 365–394.

———. 2016. "Reputation without Commitment in Finitely Repeated Games." *Theoretical Economics* 11:157–185.

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

Wendt, Alexander. 1992. "Anarchy is What States Make of It: The Social Construction of Power Politics." *International Organization* 46 (2): 391–425.

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

Appendix

Proof of Proposition 1

As mentioned before, the proofs rely on Phelan (2006)'s technique for characterizing the equilibrium. The game in that paper can be considered an example of what is described by the cooperation condition here. The differences between the game in that paper and the cooperation condition here are twofold: (i) his long-run player (a government) faces a unit measure of non-atomic citizens instead of a single short-run player at each period, which is immaterial for the formal arguments, and (ii) there, signaling is costless for the government when citizens do not produce anything, which coincides to $u_L(O,D) - u_L(O,I) = 0$, which violates Assumption 1 here, since here signaling is always costly for L. This difference is consequential for proving the uniqueness of the MPE, which is left as a conjecture in this paper. I show below that the arguments in (Phelan 2006) extend to the conflict condition discussed here. The extension is straightforward, but useful to demonstrate.

Throughout, I maintain that Assumptions 1-4 hold, and I set $\mu_0 = \epsilon$.

Specifying the Equilibrium

Proof. First, focus on L's strategy.

When $\mu < \mu^*$, L will adjust its effort level to make S indifferent between entering and staying out:

$$\mu u_S(E,I) + (1-\mu)(\hat{\sigma}_L u_S(E,I) + (1-\hat{\sigma}_L)u_S(E,D)) = d$$

which gives $\hat{\sigma}_L = \frac{\frac{d-u_S(E,D)}{u_S(E,I)-u_S(E,D)} - \mu}{1-\mu} = \frac{\mu^*-\mu}{1-\mu}$, where $\mu^* = \frac{d-u_S(E,D)}{u_S(E,I)-u_S(E,D)}$ which is $0 < \mu^* < 1$ by Assumption 3. When $\mu \geq \mu^*$, L will play D with certainty: $\hat{\sigma}_L = 0$. This concludes the specification

of L's equilibrium strategy for all μ.

Given $\hat{\sigma}_L$ defined above, and given that S observes I, L's reputation evolves according to the following expression:

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

Next, I need to show that L is able to push its reputation above μ^* in finite steps.

If $\frac{\lambda-\epsilon}{\mu^*} \geq 1$, then the expression above is linear in μ with a slope weakly greater than 1. Therefore, L can exceed μ^* in finite steps. If $\frac{\lambda-\epsilon}{\mu^*} < 1$, Assumption 4 guarantees that the fixed point of the above expression is always greater than μ^*. Thus, L's reputation can exceed μ^* in finite steps. Starting from $\mu = \mu_0 = \epsilon$, let N be the minimum number of steps required for μ to exceed μ^*. The case where L's reputation equals μ^* at the N^{th} step will not occur generically, therefore I will maintain that this will strictly exceed μ^*.

Next, focus on S's strategy.

Let μ^k represent L's reputation after k consecutive observations of the signal I, where $\mu^0 = \Phi^0 = \Phi(\mu|D) = \epsilon$, and $\mu^1 = \Phi^1 = \Phi(\Phi^0|I)$, and $\mu^2 = \Phi^2 = \Phi(\Phi^1|I)$, and so on. I will specify the behavior of S on this grid of beliefs $\mu \in \{\mu^0, \mu^1, ...\}$. Recall that L's strategy $\hat{\sigma}_L$ was constructed to induce indifference on S when $\mu < \mu^*$, and was set at $\hat{\sigma}_L = 0$ for $\mu \geq \mu^*$. Similarly, the strategy of S will induce indifference on L between I and D when $\mu < \mu^*$, and when $\mu \geq \mu^*$ its strategy will be set at $\hat{\sigma}_S = 1$ for the cooperation condition and $\hat{\sigma}_S = 0$ for the conflict condition. For notational simplicity I use $\hat{\sigma}_S^k = \hat{\sigma}_S(\mu = \mu^k)$. Let $V^k = V(\mu = \mu^k)$ be the continuation value of the game for the normal type of L, starting from the reputation μ^k. Then for $k \in \{0, 1, .., N-1\}$, the strategy of S will satisfy the conditions below.

$$V^k = (1 - \delta(1-\epsilon))\Big(\hat{\sigma}_S^k u_L(E, I) + (1 - \hat{\sigma}_S^k)u_L(O, I)\Big) + \delta(1-\epsilon)V^{k+1} \tag{4}$$

$$V^k = (1 - \delta(1-\epsilon))\Big(\hat{\sigma}_S^k u_L(E, D) + (1 - \hat{\sigma}_S^k)u_L(O, D)\Big) + \delta(1-\epsilon)V^{0} \tag{5}$$

Since L will play D with certainty once $\mu \geq \mu^*$, this implies that for $k \geq N$ we have different

expressions for V^k depending on whether the cooperation or conflict condition applies:

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

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

Regardless of which condition applies, there are $(N+1) \times V + N \times \hat{\sigma}_S^k = 2N+1$ unknowns and $2N+1$ equations laid out in the above linear system for both cooperation and conflict conditions. This linear system is thus full rank and has a unique solution. The solution for $\hat{\sigma}_S^k \in \{\hat{\sigma}_S^0, \hat{\sigma}_S^1, ..., \hat{\sigma}_S^{N-1}\}$ is as follows:

$$\hat{\sigma}_S(\mu = \mu^k) = \hat{\sigma}_S^k = \left(\frac{\Sigma_{i=0}^{k}(\frac{A-B}{\delta(1-\epsilon)})^i}{\Sigma_{i=0}^{N}(\frac{A-B}{\delta(1-\epsilon)})^i}\right) - \frac{B}{\delta(1-\epsilon)}\left(\frac{\Sigma_{i=k}^{N-1}(\frac{A-B}{\delta(1-\epsilon)})^i}{\Sigma_{i=0}^{N}(\frac{A-B}{\delta(1-\epsilon)})^i}\right) \qquad \text{(Cooperation)}$$

$$\hat{\sigma}_S(\mu = \mu^k) = \hat{\sigma}_S^k = -\frac{B}{\delta(1-\epsilon)}\left(\frac{\Sigma_{i=k}^{N-1}(\frac{A-B}{\delta(1-\epsilon)})^i}{\Sigma_{i=0}^{N}(\frac{A-B}{\delta(1-\epsilon)})^i}\right) \qquad \text{(Conflict)}$$

where $A = \frac{u_L(E,D)-u_L(E,I)}{u_L(E,D)-u_L(O,D)}$, and $B = \frac{u_L(O,D)-u_L(O,I)}{u_L(E,D)-u_L(O,D)}$. Assumptions 1 and 3 imply that $0 < |A| < 1$, $0 < |B| < 1$, and $0 < A - B < 1$. Let $\bar{\delta} = max\{\frac{A-B}{(1-\epsilon)}, \frac{|B|}{(1-\epsilon)}\}$. Then if $\delta > \bar{\delta}$, we have $0 < \hat{\sigma}_S^k < 1$ for $k \leq N-1$. For $k = N$, we have $\hat{\sigma}_S^N = 1$ in the cooperation condition and $\hat{\sigma}_S^N = 0$ in the conflict condition. When the left and right-hand side of the weak inequality specified in Assumption 3Aii (for cooperation) and 3Bii (for conflict) equal each other, that is, when L's cost for exerting effort does not depend on the actions of S, the strategies described above simplify to:

$$\hat{\sigma}_S^0 = \frac{B}{\delta(1-\epsilon)}, \qquad \text{and} \quad \hat{\sigma}_S^k = 1 \quad \text{for} \quad k > 0 \qquad \text{(Cooperation)}$$

$$\hat{\sigma}_S^0 = 1 + \frac{B}{\delta(1-\epsilon)}, \qquad \text{and} \quad \hat{\sigma}_S^k = 0 \quad \text{for} \quad k > 0 \qquad \text{(Conflict)}$$

This concludes the specification of equilibrium strategies.

The final step is to show that when $k \geq N$, that is when $\mu \geq \mu^*$, L strictly prefers not exerting any effort. The proof is by contradiction,

Formally, for $k \geq N$ we want to show that:

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

Suppose $k \geq N$ and $\mu^k > \mu^*$, but contrary to the claim above, L weakly prefers exerting effort:

$$(1-\delta(1-\epsilon))u_L(E,I) + \delta(1-\epsilon)V^{k+1} \geq (1-\delta(1-\epsilon))u_L(E,D) + \Delta V^0 \quad \text{(Cooperation)}$$
$$V^{k+1} \geq \frac{1-\delta(1-\epsilon)}{\delta(1-\epsilon)}\Big(u_L(E,D) - u_L(E,I)\Big) + V^0 \quad \text{(Cooperation)}$$
$$(1-\delta(1-\epsilon))u_L(O,I) + \delta(1-\epsilon)V^{k+1} \geq (1-\delta(1-\epsilon))u_L(O,D) + \Delta V^0 \quad \text{(Conflict)}$$
$$V^{k+1} \geq \frac{1-\delta(1-\epsilon)}{\delta(1-\epsilon)}\Big(u_L(O,D) - u_L(O,I)\Big) + V^0 \quad \text{(Conflict)}$$

I set $k = N-1$ in equation 4 and solve for V^N using 5, which produces the following equality:

$$V^N = \frac{1-\delta(1-\epsilon)}{\delta(1-\epsilon)}\hat{\sigma}_S^{N-1}\Big(u_L(E,D) - u_L(E,I)\Big) + \frac{1-\delta(1-\epsilon)}{\delta(1-\epsilon)}(1-\hat{\sigma}_S^{N-1})\Big(u_L(O,D) - u_L(O,I)\Big) + V^0$$

Given that for $k < N$ we have $0 > \hat{\sigma}_S^k > 1$, and by Assumption A3Aii (for cooperation) and A3Bii (for conflict), the above expression implies $V^N < V^{N+1}$. However, note that since $\hat{\sigma}_S^{k \geq N} = 1$ in the cooperation condition and $\hat{\sigma}_S^{k \geq N} = 0$ in the conflict condition, for $k \geq N$ it should be that $V^k = V^{k+1}$, because the continuation game starting at any prior $\mu \geq \mu^*$ should be identical. This produces a contradiction, therefore it must be that when $\mu \geq \mu^*$, L strictly prefers not exerting any effort.

This completes the proof of Proposition 1. □