CREDIBLE REFORMS: A ROBUST IMPLEMENTATION APPROACH

TYRION LANNISTER*

ABSTRACT. I study the problem of a government with low credibility, which decides to make a reform to remove ex-post time inconsistent incentives due to lack of commitment. The government has to take a policy action, but has the ability to commit to limiting its discretionary power. If the public believed the reform solved this time inconsistency problem, the policy maker could achieve complete discretion. However, if the public does not believe the reform to be successful, some discretion must be sacrificed in order to induce public trust. With repeated interactions, the policy maker can build reputation about her reformed incentives. However, equilibrium reputation dynamics are extremely sensitive to assumptions about publics beliefs, particularly after unexpected events. To overcome this limitation, I study the *optimal robust policy* that implements public trust for all beliefs that are consistent with common knowledge of rationality. I focus on robustness to all extensive-form rationalizable beliefs and provide a characterization. I show that the robust policy exhibits both partial and permanent reputation building along its path, as well as endogenous transitory reputation losses. In addition, I demonstrate that almost surely will eventually convince the public that she does not face a time consistency problem and she is able to do this with an exponential arrival rate. This implies that as we consider more patient policy makers, the payoff of robust policies converge to the complete information benchmark. Finally, I explore how further restrictions on beliefs alter optimal policy and accelerate reputation building.

JEL CLASSIFICATION CODES: E58, D81, D82, D83

1. INTRODUCTION

In the mind of policy makers, a reputation for credibility is a delicate and hard-won. Policy shifts are discussed with great care and concerns regarding how the public will react. By contrast, formal models of reputation employing insights from repeated games typically assume a perfect degree of certainty and coordination. The purpose of this paper is to build

Date: October 2016.

I am grateful to Daron Acemoglu, George-Marios Angeletos, Abhijit Banerjee, Robert Townsend, Ivan Werning, and Muhamet Yildiz who provided support, guidance, and extremely helpful suggestions. I am also thankful for the essential feedback provided by Fernando Alvarez, Dilip Abreu, Gabriel Carroll, Arun Chandrasekhar, Sebastian Di Tella, Juan Dubra, Glenn Ellison, Alvaro Forteza, Ben Golub, Horacio Larreguy, Hugo Hopenhayn, Anton Kolotilin, Pablo Kurlat, Stephen Morris Plamen Nenov, Juan Passadore, Juuso Toikka, Xiao Yu Wang, Alex Wolitzky, Luis Zermeño and Jan Zilinsky .

on this literature in order to model reputation in a way that reflects the uncertainty faced by policy makers.

Following the seminal papers of Kydland and Prescott [1977] and Barro and Gordon [1983], macroeconomic theory has extensively studied the so-called time inconsistency problem of government policy. In essence, all time inconsistency problems consist of an authority who needs some agents in the economy (e.g., consumers, financial sector) to trust her with a decision that will be taken on their behalf. For example, in the canonical example of monetary policy, a policy maker needs the public to trust her announcements of inflation policy. The fundamental problem resides in the fact that even if the policy maker is allowed to announce what policy she plans to take, the agent's decision to trust decision ultimately depends on their *perception* or beliefs about the (ex-post) incentives the policy maker will face *after* they trust her. This creates a wedge between the ideal policies the authority would want to implement and the ones that she can credibly promise. In the context of our inflation example, when agents know that the government will have ex-post incentives to boost employment by increasing inflation and reducing real wages, this will result in an inefficiently high equilibrium inflation.

The literature has dealt with this problem in two ways. Some authors have argued that the government should *forfeit some flexibility* through a formal arrangement (e.g., inflation targeting/caps and tax restrictions). Others have argued that the government should *modify its incentives*, either by credibly expressing reputational concerns through repeated interactions (Stokey [1989, 1991] and Chari and Kehoe [1990, 1993a,b]) or by delegating the decision making to an agency with different incentives that limit its time inconsistency bias. Examples of such delegation include appointing a conservative central banker (Rogoff [1985]) or making the monetary authorities subject to a formal or informal incentive contracts (Lohmann [1992], Persson and Tabellini [1993], Walsh [1995] and Svensson and Woodford [2004]). If full commitment to contingent policies is not available and flexibility is socially desirable, these "*incentive reforms*" may become a desirable solution. The key difference between both solutions is that policies enforced by incentive reforms are very sensitive to the assumption that the public knows exactly what the reformed incentives are. If the public believed that with a high probability, the government still has a time inconsistency problem, then the situation would remain unsolved. I will model this uncertainty as the public having incomplete information about the policy makers incentives, as in Barro [1986], Phelan [2006], Vickers [1986] and more recently in Sibert [2003] and Hansen and McMahon [2016], but I will also allow the public to have uncertainty about the governments expectations for the continuation game (i.e. uncertainty about policy maker's strategies and

her higher order beliefs). The main goal of our paper will be to investigate if, through repeated interactions, the government can convince the public about its reformed incentives, and convince them that she does not have a time inconsistency problem.

Using equilibrium analysis to answer this question typically relies on rather strong common knowledge assumptions as to how agents play, their priors on the government's type, as well as how all parties revise their beliefs after unexpected events. In the particular case of repeated games, predictions of a particular equilibrium may be extremely sensitive to assumptions as to how agents update their expectations about the continuation game *on all potential histories* that might be observed. These are complicated, high dimensional objects, of which the policy maker may have little information about. Persson and Tabellini [1990] stress the difficulties in using reputational models for normative analysis (even without incomplete information about the policy maker's incentives), which depends on the degree of coordination among market participants, a problem made even worse in a repeated game setting. This may be due to the difficulty in eliciting both the higher order beliefs from the public as well as contingent beliefs on nodes that may never be reached.

The approach I use is conceptually related to the robust mechanism design literature (Bergemann and Morris [2005, 2009]).The policy is required to implement trust along its path for all feasible agent beliefs within a large class. The class of beliefs that are deemed feasible is crucial to our exercise, since a larger set makes the analysis more robust, and a smaller set makes it trivial; the feasible set I consider is discussed below. The goal is to implement the publics trust for any belief system they may hold, consistent with strong common certainty of rationality: every agent knows that the other agents are rational, they know everyone knows this, and so on. This is the rationale behind rationalizability, which consists on an iterative deletion of dominated strategies. The present analysis requires beliefs to be such that agents do not to question the government's rationality, unless proven otherwise, which is given by the solution concept of *strong* or *extensive-form rationalizability* of Pearce [1984], Battigalli and Siniscalchi [2002].

I show that this policy exhibits endogenous transitory gains and losses of reputation. Moreover, the policy almost surely achieves permanent separation (i.e. public is convinced about the success of the reform from then on) it does so with an exponential arrival rate. As the discount factor of the policy maker increases, the expected payoff of this robust policy approximates the full commitment first best benchmark. This policy will also be the *maxmin* strategy for the policy maker regardless of their particular beliefs and hence provides a lower bound both for payoffs as well as the speed of separation of any strategy that is consistent with extensive form rationalizability.

To understand the intuition behind the results, suppose the public hypothesizes facing an old, time inconsistent policy maker (if the reform has not taken place) and has observed her taking an action that did not maximized her spot utility. To fix ideas, suppose she took an action that, if she was the time inconsistent type, gave her 10 utils. Instead, she could have reacted in a manner that gave her 25 in spot utility instead. The implied opportunity cost paid would only be consistent with her being rational if she expected a net present value of at least 15 utils, and therefore the opportunity cost paid would have been a profitable investment. This further implies that, if she was still the time inconsistent policy maker, her beliefs and planned course of action from tomorrow onward must deliver (from the policy maker's point of view) more than 15 utils, which is a constraint that rationality imposes on the government's expected future behavior. However, if the maximum feasible net present value attainable by a time inconsistent government was 10 utils, the public should then infer that the only possible time inconsistent type that they are facing is an irrational one. However, if such a history was actually consistent with the reformed, time consistent policy maker (e.g. she had an opportunity cost of 2) then the public should henceforth be fully convinced that they are facing the reformed government. The implied spot opportunity cost paid by the time inconsistent type, therefore, will be a measure of reputation that places restrictions on what the public believes the policy maker will do in the future. I emphasize that this will be independent of their particular beliefs, and it relies only on an assumption of rationality. I also show that this is, in fact, the *only* robust restriction that strong common certainty of rationality imposes, making the implied opportunity cost paid the *only* relevant reputation measure in the robust policy. Moreover, I show that the optimal robust policy can be solved as a dynamic contracting problem with a single *implied promise keeping constraint*, analogous to Thomas and Worrall [1988], Kocherlakota [1996] and Alvarez and Jermann [2000] in the context of optimal risk sharing with limited commitment, which makes the analysis of the optimal robust policy quite tractable.

The rest of the paper is organized as follows. Section 2 describes two macroeconomic applications of time inconsistency, monetary policy and capital taxation, which are informed by our theoretical results. Section 3 provides a brief literature review. Section 4 introduces the stage binary action repeated game, and introduces the concepts of weak and strong rationalizability. Section 5 defines the concept of robust implementation and studies robust implementation for all weak and strong rationalizable outcomes. In section 6 I study the basic properties of the optimal robust strategy and the reputation formation process as well as the limiting behavior as policy makers become more patient. In Section 7 I study an alternative model, where the policy maker can make ex ante transfers to agents in exchange

for their trust. This model allows for more generality in the environment studied, while at the same time can be solved almost in closed form, which can be used to illustrate the main tradeoffs studied before. In Section 8 I study some extensions to our model and discuss avenues for future research. Finally, Section 9 concludes.

2. Examples

I start with some time-inconsistency examples from the literature and use them to motivate my model and analysis. I focus on two of the most commonly studied environments in macroeconomic literature: *capital taxation* and *monetary policy*. I will illustrate that even if the policy maker undertakes a reform that solves for her time inconsistent bias, when agents have imperfect knowledge about the government objectives, a time inconsistency problem of government policy arises.

2.1. Capital Taxation.

I use a modified version of Phelan [2006] and Lu [2012], where the time inconsistent type is a benevolent government, instead of just opportunistic. Consider an economy with two type of households: workers (w) and capitalists (k). For each type, there is a continuum of identical households, of measure one. Capitalist households have an investment possibility and can invest $q \in [0, \overline{q}]$ units in a productive technology with a constant marginal benefit of 1 and a constant marginal cost of I. Workers do not have access to this technology and can only consume their own, fixed endowment of $e > 0$.

There is also a public good that can be produced by a government that has a marginal value of z_k to capitalist households and z_w to workers, where $z = (z_k, z_w)$ is a joint random variable. The government taxes a portion $\tau(z)$ of capital income after the shock is realized, in order to finance the production of $r(z)$ units of public good. Given the expected policy $\{\tau(z), r(z)\}_{z \in Z}$, workers and capitalists household utilities are given by

$$U_w = e + \mathbb{E}_z [r(z) z_w] \tag{2.1}$$

$$U_k = (1 - \tau^e) q - Iq + \mathbb{E}_z [r(z) z_k] \tag{2.2}$$

where $\tau^e = \mathbb{E}[\tau(z)]$. A leading example is the case where the "public good" is simply redistribution from the capitalists to the workers. In this case $z_w > 0$ and $z_k = 0$.

The optimal investment decision for a capitalist is to invest $q_i = \overline{q}$ if $1 - \tau^e < I$, and 0 otherwise, since they do not internalize their marginal effect on the production of public good. As a benchmark, we will first solve for the policy $\{\tau_k(z), r_k(z)\}_{z \in Z}$ that maximizes only the

capitalist households expected utility, subject to the government's budget constraint:

$$\max_{q\in[0,\overline{q}],\tau_k(\cdot),r_k(\cdot)} (1-\mathbb{E}[\tau_k(z)])\,q - Iq + \mathbb{E}[r_k(z)\,z_k] \text{ s.t. } r_k(z) \leq \tau_k(z)\,q \text{ for all } z \tag{2.3}$$

Given q, the optimal policy involves full expropriation $(\tau_k(z)=1, r_k(z)=q)$ when $z_k \geq 1$ and zero taxes otherwise, which induces an expected tax rate of $\tau^e = \Pr(z_k \geq 1)$. If

$$I < \Pr(z_k \leq 1) \tag{2.4}$$

then households expecting policy $\{\tau_k(z), r_k(z)\}_{z\in Z}$ will choose $q=\overline{q}$. However, this will not be the policy chosen by a benevolent government that also values workers. After the households investment decision, and the state of nature has been realized, the government chooses public good production $\tilde{r}$ and tax rate $\tilde{\tau}$ to solve:

$$\max_{\tilde{r},\tilde{\tau}} \tilde{r}(z_k + \alpha z_w) + (1-\tilde{\tau})\,q \text{ s.t. } \tilde{r} \leq \tilde{\tau} q \tag{2.5}$$

where $\alpha \geq 0$ is the relative weight that the government puts on workers welfare.

Defining $z_g := z_k + \alpha z_w$, the marginal value of the public good between capitalists and the government will typically be different, unless $\alpha = 0$. Solving 2.5 gives $\tau_g^e = \Pr(z_k + \alpha z_w > 1)$. I will assume that $I > \Pr(z_k + \alpha z_w \leq 1)$, so capitalist households will optimally decide not to invest $(q=0)$ and no public good production will be feasible. Finally, I assume that the parameters of the model are such that a benevolent government would want to commit to the capitalist's most preferred policy $\{\tau_k(z), r_k(z)\}_{z\in Z}$ if she was given the possibility. [1]

To solve the "time inconsistency" problem, I first explore the possibility of introducing a cost to raise taxes. This means that if taxes are increased, the government has to pay a cost of $c > 0$. The government would then optimally choose taxes $\tau = 0$ and increase them only when needed. She solves $\max_{\tilde{r},\tilde{\tau}} \tilde{r}(z_k + \alpha z_w) - 1\{\tilde{\tau} > 0\}\,c$ subject to $\tilde{r} \leq \tilde{\tau} q$. In this case, the expected tax rate is now $\tau^e(c) = \Pr\left(z_k + \alpha z_w > \frac{c}{q}\right)$. By setting $c = \overline{c}$ to solve $1 - \tau^e(c) = I$, the time inconsistent government, by credibly distorting its tax policy, can now induce households to invest.

Another way to deal with the problem is to make an institutional reform and delegate the public good provision to a different policy maker, who has incentives aligned with the capitalist households. The new policy maker type now solves $\max_{\tilde{r},\tilde{\tau}} \tilde{r}(z_k + \alpha_{new} z_w) + (1-\tilde{\tau})\,q$ subject to $\tilde{r} \leq \tilde{\tau} q$. By introducing a "pro-capitalist government" with $\alpha_{new} = 0$, the capitalists most desirable policy $\{\tau_k(z), r_k(z)\}_{z\in Z}$ would be credibly implemented without the need of setting a cost to increase capital taxes. Under some parametric assumptions, it will

[1] This happens if $\Pr(z_k > 1)\,\mathbb{E}(z_k + \alpha z_w \mid z_k > 1) + \Pr(z_k \leq 1) > 0$.

be socially desirable for the benevolent government (without taken into account the commitment cost paid) to delegate policy making to the "pro-capitalist" type that does not need to impose tax increase costs to convince households to invest[2].

However, if households were not convinced that they are indeed facing a reformed, pro capitalist government, they will need some assurance (i.e. some restrictions to ex post increase taxes) in order to *trust* that the government will not expropriate their investments too often. In Phelan [2006], the analog to a pro-capitalist type is a commitment type (as in Fudenberg and Levine [1989]) that always picks the same tax rate. In Lu [2012], the government can make announcements, and can be either a committed type (i.e. one bound by the announcement) or a purely opportunistic type that may choose to deviate from the promised policy, which is analog to the benevolent type in our setting.

Formally, Let $\pi \in (0,1)$ be the probability that capitalist households assign to the new government to actually be a pro-capitalist type. Then, if there is complete flexibility to increase taxes, the expected tax rate would be

$$\tau^e(\pi) = \pi \Pr(z_k > 1) + (1-\pi)\Pr(z_k + \alpha z_w > 1) \tag{2.6}$$

Condition 2.6 implies that for sufficiently low π, we would have $1-\tau^e(\pi) < I$ and capitalists will decide not to invest. Thus, as long as capitalists perceive that the new government might still be time inconsistent (modeled by a low π), it will be necessary to set some cost to raise taxes in order to induce capitalists to invest, even though the government is now a pro capitalist type.

2.2. **Monetary Policy.** I use the framework and notation in Obstfeld and Rogoff [1996][3]. I assume that total output (in logs) y_t depends negatively on the real wage and some supply side shock z_t, according to

$$y_t = \overline{y} - [w_t - p_t(z_t)] - z_t \tag{2.7}$$

where $\overline{y}$ is the flexible price equilibrium level, z_t is a supply shock with $\mathbb{E}(z_t) = 0$ and $p_t(z_t)$ is the nominal price level at time t set by the monetary authority. In equilibrium, nominal wages are set according to $w_t = \mathbb{E}_{t-1}[p_t(z_t)]$, to match expected output to its natural level $\overline{y}$. A benevolent monetary authority observes the shock z_t and decides the inflation level in order to minimize deviations of output with respect to a social optimal output y^* and

[2]This happens if $\Pr(z>1)\mathbb{E}(z_k+\alpha z_w \mid z_k>1)+\Pr(z_k\leq 1) > \Pr\left(z_k+\alpha z_w > \frac{c}{q}\right)\mathbb{E}\left(z_k+\alpha z_w \mid z_k+\alpha z_w > \frac{c}{q}\right)+\Pr(z_k+\alpha z_w \leq 1)$

[3]Section 9.5, pp 634-657.

deviations of inflation from a zero inflation target:

$$\mathcal{L} = \frac{1}{2}\left(y_t - y_t^*\right)^2 + \frac{\chi}{2}\pi_t^2. \tag{2.8}$$

I assume that $k := y_t^* - \overline{y} > 0$. This measures the wedge between the output level targeted by authorities and the natural level of output, which are different due to market inefficiencies, even under flexible prices.[4] Defining inflation as $\pi_t(z_t) := p_t(z_t) - p_{t-1}$, and using equation 2.7, together with the wage setting rule, the loss function simplifies to

$$\mathcal{L}(\pi, \pi^e, z) = \frac{1}{2}\left[\pi - \pi^e - z - k\right]^2 + \frac{\chi}{2}\pi^2, \tag{2.9}$$

where $\pi^e = \mathbb{E}[\pi(z)]$ are the expectations formed by the private sector about inflation (which should be correct under rational expectations). The *full commitment benchmark*, exists when the monetary authority can commit, ex-ante, to a state contingent inflation policy $\pi(z)$, to solve

$$\min_{\pi(\cdot),\pi^e} \mathcal{L}(\pi, \pi^e, z) \text{ s.t: } \pi^e = \mathbb{E}[\pi(z)]$$

with solution

$$\pi_c(z) = \frac{z}{1+\chi} \text{ and } \pi^e = \mathbb{E}_z[\pi_c(z)] = 0. \tag{2.10}$$

In contrast, when the monetary authority cannot commit to a state contingent policy, conditional on π^e and z, she chooses π to solve:

$$\min_{\pi\in\mathbb{R}} \mathcal{L}(\pi, \pi^e, z) \iff \pi_{nc}(z) = \frac{\pi^e + z + k}{1+\chi} \tag{2.11}$$

By taking expectations on both sides of 2.11 we get $\pi^e = \frac{k}{\chi}$, which I will refer to the *time inconsistency bias*. Equilibrium inflation is then

$$\pi_{nc}(z) = \frac{k}{\chi} + \pi_c(z) \tag{2.12}$$

Output $y(z)$ is identical in both cases, for all shocks. However, $\mathbb{E}\left[\pi_{nc}^2(z)\right] = \mathbb{E}\left[\pi_c^2(z)\right] + \frac{k^2}{\chi^2}$ so the outcome with no commitment is strictly worse than the full commitment benchmark.

An analogous binary model to the capital taxation environment can be written if wage setters can only choose between two levels, ω_L and ω_H. In such a model, the analog of implementing investment in the capital taxation model would be to convince agents that expected inflation is low enough, so that they choose $\omega = \omega_L$. Agents choosing high wages (because of high inflation expectations) is an undesirable equilibrium outcome, akin to capitalists not investing in the previous model.

[4] See Rogoff [1985] and Obstfeld and Rogoff [1996] for a discussion of such potential inefficiencies.

How can the monetary authority solve this problem? A first approach is to formally limit the flexibility of monetary policy by restricting the set of inflation levels from which the monetary authority can choose. Athey et al. [2005] show that this can be optimally done by choosing an *inflation cap* $\overline{\pi}$[5], such that $\pi(z) \leq \overline{\pi}$ for all z. Inflation policy is now

$$\pi(z \mid \overline{\pi}) = \min\left\{\frac{\pi^e(\overline{\pi}) + z + k}{1+\chi}, \overline{\pi}\right\}$$

where $\pi^e(\overline{\pi})$ solves the fixed point equation $\pi^e(\overline{\pi}) = \mathbb{E}_z\left[\min\left\{\frac{\pi^e(\overline{\pi})+z+k}{1+\chi}, \overline{\pi}\right\}\right]$.

An alternative approach, first suggested by Rogoff [1985] and studied by Persson and Tabellini [1990], Walsh [1995], Svensson and Woodford [2004] among others, is to introduce *institutional reforms to the monetary authority*. with the purpose of alleviating the time inconsistency bias by inducing changes in their preferences. Imagine first that the government can delegate the monetary policy to a policy maker type $\theta = new$, that wants to minimize a modified loss function

$$\mathcal{L}_{new}(\pi, \pi^e, z) = \frac{1}{2}\left[\pi - \pi^e - z - k^{new}\right]^2 + \frac{\chi^{new}}{2}\pi^2 \tag{2.13}$$

Rogoff [1985] suggests placing a "conservative central banker", that has $k^{new} = k$ but $\chi^{new} > \chi$, thereby placing a greater importance on inflation stabilization than society does. From 2.12 we see that increasing the weight χ makes the effective inflation bias smaller, and hence may alleviate the time inconsistency problem at the expense of a milder reaction to supply shocks, as evidenced by equation 2.11.

By setting $k^{new} = 0$ the optimal policy with no commitment for $\theta = new$ would implement the full commitment solution. This would correspond to having a monetary authority that believes there are no market inefficiencies and wants to stabilize output around its flexible price equilibrium level $\overline{y}$. The same outcome can be implemented if, instead of changing the preference parameters, we add a linear term to the loss function $\mathcal{L}_{new} = \mathcal{L} + \alpha[\eta\pi]$. Walsh [1995] and Persson and Tabellini [1993] argue that this can be done by offering a contract

[5]In their paper, Athey et al. [2005] solve for the optimal dynamic mechanism for a time inconsistent policy maker, that has private information about the state of the economy, which is i.i.d across periods, and show that any optimal mechanism exhibits a constant inflation cap in all periods. In a static setting, shocks can be thought as private information for the monetary authority, so an inflation cap would also be a characteristic of the more general mechanism design problem:

$$\max_{\pi(\cdot),\pi^e} \mathbb{E}_z\left[\mathcal{L}(\pi(z), \pi^e, z)\right]$$
$$s.t\ :\ \mathcal{L}[\pi(z), \pi^e, z] \geq \mathcal{L}[\pi(z'), \pi^e, z] \text{ for all } z, z' \in Z$$

to the central bank governor. This can be either a formal monetary contract[6] or an informal relational contract under which realized levels of inflation affect the continuation values for the monetary authority (e.g. the governor could be fired if inflation reaches sufficiently high levels, as in Lohmann [1992]). Here $\alpha > 0$ represents the relative weight of his self-interest payoffs relative to the social welfare. By picking $\eta = \frac{k}{\alpha}$ the full commitment inflation policy would be implemented.

While the institutional reform route may seem desirable, *these institutional reforms may not be perfectly observed by the private sector*. The public might not be convinced that the monetary authority is now more conservative or have a smaller time inconsistency bias than the previous one. Such a problem is likely to be particularly acute because these are changes in preferences, which involve either delegation or perhaps informal relational contracts that are imperfectly observed. If this was the case, restrictions such as inflation caps might still be necessary. For example, take the institutional reform with $k^{new} = 0$, and no inflation caps are set. If the public assigns probability $\mu \in (0, 1)$ to the incentive reform being successful, expected inflation would then be

$$\pi^e = (1 - \mu) \frac{k}{\chi} > 0$$

Thus, as long as the public perceives there might still be a time inconsistency bias, institutional reforms might not be enough, and inflation caps might be necessary to implement smaller inflation expectations. Barro [1986], Backus and Driffil [1984] and Vickers [1986] and more recently Sibert [2003] study models with unobservable types of monetary policy makers, investigating reputation building and signaling with commitment types, either by being committed to some specific policy, or being committed to a chosen announced policy (Lu [2012], King et al. [2012]). Hansen and McMahon [2016] study models with rational types such as the ones proposed, with "weak" (high k) and "strong" (low k) policy maker types.

3. Literature Review

The literature on time inconsistency of government policy is extensive, beginning with the seminal papers by Kydland and Prescott [1977] and Barro and Gordon [1983], where the idea of the commitment solution (i.e. choosing policy first) was first introduced. The reputation channel was first explored by Backus and Driffil [1984], Vickers [1986] and Barro

[6]Suppose the monetary authority minimizes $\hat{\mathcal{L}}_{old} = \mathcal{L}_{old} - \alpha u [\phi(\pi)]$ where $\phi(\pi)$ is a monetary reward function depending on realized inflation, and $\alpha > 0$ of monetary incentives relative to the monetary authorities "benevolent" incentives. See that by picking $\phi(\pi) = u^{-1}(-\eta\pi)$, a decreasing function of inflation, the contract will induce the linear component in 2.13, which coincides with the optimal contract in this setting.

[1986], who studied policy games in which a rational (albeit time inconsistent) government living for finitely many periods may find it optimal to imitate a "commitment type". This commitment type is an irrational one that plays a constant action at all histories. They show (following the arguments in Kreps et al. [1982], Kreps and Wilson [1982], Milgrom and Roberts [1982]) that for a long enough horizon, the unique sequential equilibrium of the game would involve the government imitating the commitment type for the first periods, and then playing mixed strategies, which imply a gradual reputation gain if she keeps imitating.

In an infinite horizon setting, Fudenberg and Levine [1989] show that a long-lived agent facing a sequence of short lived agents can create a reputation for playing as the commitment type. By consistently playing the commitment strategy, the long-lived agent can eventually convince the short lived agents that she will play as a committed type for the rest of the game. Celentani and Pesendorfer [1996] generalized this idea to the case of a government playing against a continuum of long-lived small players, whose preferences depend only on aggregate state variables. The atomistic nature of the small players allows them to use the results of Fudenberg and Levine [1989] to get bounds on equilibrium payoffs. Phelan [2006] studies the problem of optimal linear capital taxation, in a model with impermanent types, which can accommodate occasional losses of reputation. Rather than obtaining bounds, he characterizes the optimal Markovian equilibrium of the game, as a function of the posterior that the public has about the government's type.

A second strand of the literature on reputation focuses on a complete information benchmark with the goal of characterizing sustainable policies. These policies are the outcome of subgame perfect equilibria of the policy game, starting with Stokey [1989, 1991] and Chari and Kehoe [1990, 1993a,b]. In such environments, governments may have incentives to behave well under the threat of punishment by switching to a bad equilibrium afterwards. See Sargent and Ljungqvist [2004][7] for a tractable unified framework to study these issues.[8]

This paper studies reputation formation in terms of both payoff heterogeneity and equilibrium punishments. The main point of departure is that instead of designing the optimal policy for a time inconsistent policy maker (one who wants to behave as if she were time

[7]Chapter 16, pp 485-526

[8]This principle is also exploited in the relational contract literature (Levin [2003],Baker and Murphy [2002] ,Bull [1987]) where a principal announces a payment scheme after income is realized (the state-contingent policy) but has no commitment to it other than the one enforced by the threat of retaliation by the agent (not making effort, strike, quit, etc). Similar themes are studied in the literature on risk sharing with limited commitment (Thomas and Worrall [1988], Kocherlakota [1996], Ligon [1998] and Ligon et al. [2000]) where a transfer scheme conditional on the realization of income (the contingent policy) is enforced by threating agents who deviate of excluding them from the social contract.

consistent), I focus on the opposite case. I consider the problem of a trustworthy policy maker (with no time inconsistency bias) who nevertheless may be perceived as opportunistic by the agents. Her goal, therefore, is essentially to separate itself, if possible, from the time inconsistent, untrustworthy type. The most related papers in spirit to mine are Debortoli and Nunes [2010],King et al. [2012] and particularly Lu [2012] and Hansen and McMahon [2016]. Debortoli and Nunes [2010] study the optimal policy problem of a benevolent government that has access to a "loose commitment" technology, under which not all announcements can be guaranteed to be fulfilled. Lu [2012] explores the optimal policy of a committed government that worries she might be perceived as a government that cannot credibly commit to her announced policies. This paper also focuses on characterizing the optimal policy for the time consistent type (the committed type in her setting) instead of just studying the optimal policy of a time inconsistent type imitating a consistent one. King et al. [2012] apply these ideas in the context of the standard New Keynesian model, similar to our setup in subsection 2.2. Hansen and McMahon [2016] is a positive paper, which focuses mostly on the behavior of "strong" members of the board of a central bank, to signal their types at the beginning of their tenure. Ultimately these papers study equilibrium in an environment where all players involved know that the government is ex-ante either a type that can or cannot commit (which holds for all subsequent periods). They then study a particular equilibrium refinement that happens to select the best equilibrium for the able-to-commit type. They also show that other equilibrium refinements such as the intuitive criterion (e.g., Cho and Kreps [1987]) select a different equilibrium. In macroeconomics, the most related paper to mine that studies robustness to specific refinements is Pavan and Angeletos [2012]. They study the robust predictions of any equilibria in a global game setting with incomplete information.

The literature on robust mechanism design is fairly recent, starting with partial robust implementation in Bergemann and Morris [2005], and robust implementation in Bergemann and Morris [2009]. The latter focuses on finding conditions on environments and social choice functions such that they are implemented under implemented for all possible beliefs, if the only thing that the mechanism designer knows about the agent's beliefs is that they share common knowledge (or certainty) of rationality. When the environment is dynamic, different concepts of rationalizability may be used; for example Normal Form Interim Correlated Rationalizability (as in Weinstein and Yildiz [2012]) and Interim Sequential Rationalizability (Penta [2011, 2012]), among others. This paper focuses on the stronger assumption of common strong certainty of rationality (Battigalli and Siniscalchi [1999, 2002,

2003]) which is also equivalent to Pearce [1984] notion of "Extensive-form rationalizability". In a similar vein, the paper most related in spirit to mine is Wolitzky [2012]. He studies reputational bargaining in a continuous time setting where agents announce bargaining postures to which they might commit to with a given positive probability. He characterizes the minimum payoff consistent with mutual knowledge of rationality between players (i.e., one round of knowledge of rationality), and the bargaining posture that she must announce in order to guarantee herself a payoff of at least this lower bound. A crucial difference to my setting is the commitment technology, which ensures certain expected payoffs to the other party, regardless whether or not they think they are facing a rational opponent. I characterize optimal robust policy in a repeated setting in which one can guarantee themselves the best payoff that is consistent with (strong) common knowledge of rationality.

4. THE MODEL

I now introduce the framework and model, describing the stage game and showing multiplicity of equilibria in Section 4.1. I then setup the repeated game in Section 4.2 and develop the concept of system of beliefs in Section 4.3. In Section 4.4 I introduce weak and strong rationalizability and in Section 4.5 I argue why we must turn to robustness relative to equilibrium refinements.

4.1. **Stage Game.** There are two players: a policy maker d (she) and an agent p (he). Agent p represents the private sector (e.g. firms in monetary policy game, investors in capital taxation game). In the benchmark model with no commitment technology, p is asked to delegate a state-contingent decision to d (i.e. choose T), who after a state of nature $z \in Z$ is realized, has to choose a policy that affects both parties payoffs. We write $a \in \{0,1\}$ for $p's$ choice of whether to trust ($a = 1$) or not. If p does not trust d, both parties receive reservation utilities $\underline{u}_i$, $i \in \{p,d\}$. However, if p trusts, after the state z is drawn d has to chose from a menu of two alternative policies: $r \in \{0,1\}$. Sometimes we will refer to $r = 0$ as the "normal" policy, which is optimal most of the time and $r = 1$ "emergency" policy that needs to be taken in only some instances.

In the monetary policy game, for example, this model could be interpreted as wage setters choosing between two possible wage levels: high wages w_H and low wages w_L. We are interested in the policy problem of implementing low wages (interpreting $a = 1$ to setting $w = w_L$). Government policies react to monetary and supply shocks $z \in Z$, by choosing one of two inflation levels: low inflation ($\pi_L = r = 0$) and high inflation ($\pi_H = r = 1$). Analogously, in the capital taxation model of Section 2.1, the government decides whether to expropriate capital to finance public good provision (the emergency policy with $r = q$) or

to keep taxes at zero (the normal policy with $r = 0$). The government gets to make this decision only when capitalist households decide to invest $q_i = \overline{q}$ which corresponds to the public placing trust in the government as without any investment the government cannot finance public good provision in the first place. As such, $a = 1$ corresponds to "invest $\overline{q}$".

I will model shocks as $z = \left(U_p, U_{old}\right)$, where U_i is the relative utility of $r = 1$ with respect to $r = 0$, for both the public p and the decision maker d. I assume this random shock to be an absolutely continuous random variable over $Z := \left[\underline{U}, \overline{U}\right]^2 \subset \mathbb{R}^2$, with density function $f(z)$, and $\underline{U} < 0 < \overline{U}$. For example, in the capital taxation model, shocks z can be written as $z = \left(U_p, U_{old}\right) := \overline{q}(z_k - 1, z_b - 1)$. I endow the decision maker with a commitment cost technology: before p decides whether to trust or not, d chooses a cost $c \geq 0$ of taking the emergency action $r = 1$, so that ex-post utility is $u_d(c, r, z) = r(U_{old} - c)$. This can be interpreted as a *partial commitment* to the normal policy $r = 0$, which includes an escape clause to break the commitment, forcing the decision maker to suffer a cost of $c \geq 0$ utils (as in Lohmann [1992]). Although d cannot commit to a complete contingent rule, I assume that the commitment cost chosen is binding. In the capital taxation model, this corresponds to the cost of increasing taxes chosen by the time inconsistent government, while in the inflation setting model this would intuitively translate to the inflation cap $\overline{\pi}$, in that it is a partial commitment chosen by the monetary authority. I assume that, in a setting of complete information, the decision maker would need to set a positive commitment cost to induce the agent's trust:

$$\int_{U_{old}>0} U_p f(z)\, dz < \underline{u}_p < 0 \tag{4.1}$$

The first inequality shows that $c = 0$ cannot be part of any rationalizable outcome in the stage game. The second inequality shows the existence of a commitment cost that would induce the agent's trust (setting $c = \infty$). Furthermore, I assume that $\underline{u}_d < 0$, so a decision maker with incentives given by $U_{old}(z)$ would always prefer to set some commitment cost, ensuring p's trust, than having him choosing $a = 0$.

The key problem I study in this paper is the government's lack of credibility, when that government cannot enforce the private sector's trust without setting a strictly positive commitment cost. The main ingredient of the model is, as I stated before, the introduction of a reform of the decision maker incentives, which may or may not be believed by the private sector. Formally, a reform is a change in the ex-post incentives that the decision maker faces, by either delegating the decision to a different agent (such as the conservative central banker of Rogoff [1985]) or by designing a contract for the decision maker (as in Walsh

[1995], Persson and Tabellini [1993] and Svensson and Woodford [2004]). I model this reform by creating a new policy maker type, $\theta = new$, with ex-post payoffs given by

$$U_{new}(z) = U_p \tag{4.2}$$

i.e. the reformed decision maker has the same ex-post incentives as the public. Condition 4.1 implies that if p knew he was facing this type of decision maker, he would trust her even with $c = 0$, so that no commitment cost would be necessary. This corresponds to the "pro-capitalist" government in the capital taxation model, which removes the time inconsistency of government policy, or the reformed monetary authority, as in Walsh [1995] or the conservative central banker of Rogoff [1985].

Even though an incentive reform has been carried out, the public may remain unconvinced that the reform has been effective. For instance, investors might still believe that the government is not pro-capitalist enough, and will expropriate them too often; or that the new appointed central banker may not be a conservative type. I model this situation by introducing *payoff uncertainty* from the public side: p believes he is facing a reformed, time consistent decision maker $\theta = new$ with probability $\pi \in (0,1)$ and otherwise faces the old, time inconsistent type $\theta = old$.

I assume that if p was guaranteed his most desirable policy, he would place trust in d, but would not if she faced decision maker with no commitment. It also makes the assumption that p would trust d if she could somehow commit to play $r = 0$ with probability 1, and never reacting to shocks. [9] I also assume that $\underline{u}_d < 0$, so that both parties and types would benefit from this commitment (and hence losing all flexibility), if such a policy was enforceable. In the capital taxation problem, this would correspond to a ban on positive taxes while in our inflation example this would correspond to a commitment to zero inflation.[10] Although this would induce the public trust, this would come at the cost of losing all flexibility to optimally react to shocks. I allow $c = \infty$, the case where d decides to shut down the emergency action r, so that the commitment cost set is $C = \mathbb{R}_+ \cup \{\infty\}$.

This model follows an *executive approach* to optimal policy, where the decision maker herself decides the commitment cost and the policy rule. This contrasts with the *legislative approach* studied in Canzoneri [1985] and Athey et al. [2005] who instead solve for the

[9] In the context of the capital taxation model, by setting $\underline{u}_p = -(1-I)\,\mathbb{E}(z_k \mid z_b \geq 1)\,(\overline{q}-1)$, the left hand side inequality of 4.1 corresponds to the solution to 2.3 in the capital taxation problem, together with assumption 2.4. Notice also that $\underline{u}_p < 0$

[10] Although this seems to be to an extreme policy to be seen in practice, hyperinflation stabilization programs usually involve drastic measures, that resemble losing all flexibility to stabilize output. For example, Zimbabwe in 2009 decided to abandon its currency (and hence most of its monetary policy) within the context of a severe hyperinflation (which reached a peak of 79,600,000% per month in November of 2008).

optimal mechanism design problem from the point of view of p. In Section (8) I briefly explore this route and argue that in our setting, it would be detrimental to welfare, conditional on the government being of type $\theta = new$.

In this game, the old decision maker would then choose the commitment cost $c = \overline{c}$ that makes p indifferent between trusting and not trusting:

$$\overline{c} := \min \left\{ c \geq 0 : \int_{U_{old} > c} U_p f(z)\, dz \geq \underline{u}_p \right\} \tag{4.3}$$

One can show that $\overline{c}$ is both well defined and finite. Because the commitment cost choice is taken after the type has been realized, this is effectively a *signaling game*. The choice of commitment cost could in principle help p to infer d's type. As a prelude to what follows it is useful to characterize the set of Perfect Bayesian Equilibria (PBE) of this game. Formally, a PBE is a pair of distributions (dC_{old}, dC_{new}) over commitment costs $c \geq 0$, and a posterior probability $\hat{\pi}_p(c) \in [0,1]$ for all $c \geq 0$ such that all types of decision maker choose c_θ to maximize utility, and $\hat{\pi}_p(c)$ is consistent with Bayes rule (whenever applicable). Let $\underline{c}(\pi)$ be the minimum cost that induces p to trust d in any pooling equilibrium

$$\underline{c}(\pi) := \min \left\{ c \geq 0 : \pi \int_{U_p > c} U_p f(z)\, dz + (1-\pi) \int_{U_{old} > c} U_p f(z)\, dz \geq \underline{u}_p \right\}. \tag{4.4}$$

It is easy to see that $\underline{c}(\pi)$ is decreasing in π_0 and that $\underline{c}(\pi) < \underline{c}(0) = \overline{c}$ according to 4.3. Proposition 4.1 characterizes the set of all PBE of the stage game.

Proposition 4.1. *All PBE of the static game are pooling equilibria. For any* $\hat{c} \in \left[\underline{c}(\pi_0), \overline{c}\right]$ *there exists a PBE in which both types choose* $\hat{c}$ *as the commitment cost.*

Proof. See Appendix C □

Suppose now that, in light of the results of Proposition 4.1, the reformed decision maker $\theta = new$ is considering what commitment cost to choose, as a *normative* question; i.e. without knowing the beliefs of p, she needs choose c. If we are thinking about a policy prescription that is supported by some PBE of this game, the multiplicity of equilibria in Proposition 4.1 requires some equilibrium refinement. If the policy maker chooses the cost corresponding to the *best equilibrium* of the game (for herself), then Proposition 4.1 shows that $c = \underline{c}(\pi)$ should be played by either $\theta = new$ or $\theta = old$. I will argue that this is not a "robust" policy in the sense that expecting p to trust after observing $c = \underline{c}(\pi)$ relies on strong and sensitive assumptions about p 's beliefs, which may be imperfectly known by the decision maker.

First, observe that the best PBE policy is very sensitive to the prior. If $p's$ true prior were $\tilde{\pi} = \pi - \epsilon$ for some $\epsilon > 0$, then p would not trust after observing $c = \underline{c}(\pi)$. Second, even

if π were commonly known, after the commitment cost has been chosen, p updates beliefs to $\hat{\pi}_p(c) \in (0,1)$. The multiplicity of equilibria arises from the indeterminacy of beliefs following zero probability events generates a large set of potential beliefs that can arise in equilibrium. As such, p's behavior will depend on the complete specification of her updates beliefs for *all off-equilibrium costs*, not just the candidate equilibrium one. Therefore, small changes in the updating rule generates potentially very different behavior for p.[11] The main goal, then, is to find a cost $c \geq 0$ such that p trusts, regardless of the prior π and the updating rule $\pi_p(c)$ he might follow, as long as it is consistent with common knowledge of rationality. It is clear that by choosing $c = \infty$ and effectively removing all flexibility, a rational p would trust d independently of his beliefs. We can do better, however. If d chooses $c = \overline{c}$ then p will find it optimal to trust, irrespective of the updating rule $\hat{\pi}_p(\cdot)$ and the prior π. Is easy to see that (as will be shown later) that in fact, $c = \overline{c}$ is the *only* robust policy, when the only assumption we make about $p's$ beliefs is that they are consistent with strong common certainty of rationality; i.e. p believes he is facing a rational d if her observed past behavior is consistent with common knowledge of rationality.

Since the difference between the reformed and the time inconsistent type is about their ex-post incentives, Proposition 4.1 gives a negative result about the signaling potential of commitment costs: types cannot separate in any equilibrium of the stage game, by their choice of c. Hence, in the static game, the reformed decision maker cannot separate $\theta = old$ through her choice of commitment. Only by having repeated interactions can the reformed decision maker hope to convince p of the success of the reform, trying to signal her type through her reactions to the realized shocks. Throughout the remainder of the paper I will investigate whether robust policies, such as the one I found in the static game, can eventually convince p that $\theta = new$, regardless of his particular belief updating rule.

4.2. **Repeated game: Setup and basic notation.** I extend the stage game to an infinite horizon setting: $\tau \in \{0, 1, ...\}$. I assume that d is infinitely lived and that types are permanent; i.e. at $\tau = 0$ nature chooses $\theta = new$ with probability π_{new}. d has discounted expected utility with discount factor $\beta_\theta \in (0,1)$. For notational ease, I will assume $\beta_{old} = \beta_{new} = \beta$. I will specify when the results are sensitive to this assumption. Shocks are independent and identically distributed across periods: $z_\tau := \left(U_{p,\tau}, U_{old,\tau}\right) \sim_{i.i.d} f(z_\tau)$. I assume that there is a sequence of myopic short run players p_τ (or equivalently $\beta_p = 0$) which is a standard

[11] More generally, we apply Aumann and Brandenburger [1995] results to the interim normal form of this game, finding tight sufficient conditions for any particular Bayesian equilibria (not necessarily perfect) to be the expected solution outcome: **(a)** There is common knowledge of rationality, **(b)** the strategies of both $\theta = new$ and $\theta = old$ prescribed by the Bayesian equilibrium are common knowledge, and **(c)** the inference rule $\hat{\pi}_p(.)$ is also common knowledge

assumption in the reputation literature (Fudenberg and Levine [1989], Phelan [2006]). This will be without loss of generality for most applications to macroeconomic models applications.[12] At every period, d chooses $c_\tau \geq 0$ which is binding only for that period. The decision maker can change its choice freely in every period. I also assume that all past history of actions and shocks (except for $d's$ payoff type) is observed by all players at every node in the game tree. I will further assume that the structure of the game described so far is *common knowledge* for both players and that agents know their own payoff parameters.[13]

A stage τ *outcome* is a 4-tuple $h_\tau = (c_\tau, a_\tau, z_\tau, r_\tau)$ where c_τ is the commitment cost, $a_\tau \in \{0,1\}$ is the trust decision, and $r_\tau \in \{0,1\}$ is the contingent policy, where $r_\tau = 1$ if d chooses the emergency action, and $r_\tau = 0$ otherwise. A *history* up to time τ is defined as $h^\tau := (h_0, h_1, ..., h_{\tau-1})$. I will refer to a "partial history" as a history plus part of the stage game. For example, p moves at histories (h^τ, c_τ), after the commitment cost is chosen. The set of all partial histories will be denoted as $\mathcal{H}$, and $\mathcal{H}_i \subseteq \mathcal{H}$ is the set of histories in which agent $i \in \{p, d\}$ has to take an action.

A *strategy* for the policy maker is a function $\sigma_d : \mathcal{H}_d \to C \times \{0,1\}^Z$ that specifies, at the start of every period τ, a commitment cost c_τ and the contingent choice provided p trusts. Then, we can always write a strategy as a pair $\sigma_d(h^\tau) = (c^{\sigma_d}(h^\tau), r^{\sigma_d}(h^\tau, \cdot) : C \times Z \to \{0,1\})$, where the choice is a commitment cost $c^{\sigma_d}(h^\tau) \in C$ and a *policy rule function* $r^{\sigma_d}(h^\tau, c_\tau, z_\tau)$ of the shock, given commitment cost c_τ The superscript σ_d serves to remind the reader that these objects are part of a single strategy σ_d. Likewise, a strategy σ_p for p is a function $\sigma_p : \mathcal{H}_p \to \{0,1\}$ that assigns to every observed history, his trust decision; i.e. $\sigma_p(h^\tau, c_\tau) = a^{\sigma_p}(h^\tau) = 1$ if p trusts, and 0 otherwise. Write the set of strategies of each agent as Σ_i for $i \in \{d, p\}$ Also let $\Sigma = \Sigma_d \times \Sigma_p$ be the set of strategy profiles $\sigma = \left(\sigma_d, \sigma_p\right)$. If player $i \in \{d, p\}$ plays strategy σ_i, the set of histories that will be consistent with σ_i is denoted $\mathcal{H}(\sigma_i) \subset \mathcal{H}$. For a history $h \in \mathcal{H}$ we say a strategy σ_i *is consistent with* h if $h \in \mathcal{H}(\sigma_i)$. Let $\Sigma_i(h) = \{\sigma_i \in \Sigma_i : h \in \mathcal{H}(\sigma_i)\}$ be the set of strategies consistent with h.

Given a strategy profile $\sigma = \left(\sigma_p, \sigma_d\right)$ let $W_\theta(\sigma \mid h)$ be the expected continuation utility for $d's$ type $\theta \in \{old, new\}$ given history h

$$W_\theta(\sigma \mid h) := (1-\beta)\,\mathbb{E}\left\{\sum_{s=\tau}^{\infty} \beta^{s-\tau}\left[a_s r_s\left(U_{\theta,s} - c_s\right) + (1-a_s)\,\underline{u}_p\right] \mid h\right\} \tag{4.5}$$

[12] Celentani and Pesendorfer [1996] show that this assumption is without loss of generality when p is modeled as representative agent for a continuum of atomistic and anonymous patient agents. In particular, the capital taxation model of section 2.1 satisfies these assumptions when capitalist households have a common discount rate $\delta_k \in (0,1)$.

[13] These are the basic assumptions in Penta [2012].

where $c_s = c^{\sigma_d}(h^s)$, $a_s = a^{\sigma_p}(h^s, c_s)$ and $r_s = r^{\sigma_d}(h^s, c_s, z_s)$. Likewise, define $V(\sigma \mid h) = a_\tau \mathbb{E}_z\left(r_\tau U_{p,\tau}\right) + (1 - a_\tau)\,\underline{u}_p$ as the spot utility for agent p at history h.

4.3. **Systems of Beliefs.** Agents form beliefs about the payoff types of the other player, as well as the strategies that they may be planning to play. In static games, such beliefs are characterized by some distribution $\pi \in \Delta(\Theta_{-i} \times S_{-i})$ where Θ_{-i} is the set of types of the other agent and S_{-i} their strategy set. In our particular game, $\Theta_d = \{new, old\}$ and $\Theta_p = \{p\}$. In dynamic settings, however, agents may *revise their beliefs* after observing the history of play. This revision is described by a *conditional probability system*, which respects Bayes rule whenever possible. Formally, let $\mathcal{X}_i$ the Borel $\sigma-$algebra generated by the product topology[14] on $\Theta_{-i} \times \Sigma_{-i}$ and $\mathcal{I}_i = \left\{E \in \mathcal{X}_i : \text{proj}_{\Sigma_{-i}} E = \Sigma_{-i}(h) \text{ for some } h \in \mathcal{H}\right\}$ be the class of information sets for i. A *system of beliefs* π_i on $\Theta_{-i} \times \Sigma_{-i}$ is a mapping $\pi_i : \mathcal{I}_i \to \Delta(\Theta_{-i} \times \Sigma_{-i})$ such that:

(1) Given an information set $E \in \mathcal{E}_i$, $\pi_i(. \mid E)$ is a probability measure over $\Theta_{-i} \times \Sigma_{-i}$.[15]
(2) If $A \subseteq B \subseteq C$ with $B, C \in \mathcal{I}_i$, then $\pi_i(A \mid B)\,\pi_i(B \mid C) = \pi_i(A \mid C)$.

I write $\pi_i(E \mid h) = \pi_i(E \mid \Sigma_i(h))$ for $E \subset \Theta_{-i} \times \Sigma_{-i}$ for the probability assessment of event E conditional on history h. Denote $\Delta^{\mathcal{H}}(\Theta_{-i} \times \Sigma_{-i})$ to be the set of all systems of beliefs. Given $\pi_d \in \Delta^{\mathcal{H}}\left(\Theta_p \times \Sigma_p\right) = \Delta^{\mathcal{H}}\left(\Sigma_p\right)$ and strategy $\sigma_d \in \Sigma_d$, define $W_\theta^{\pi_d}(\sigma_d \mid h)$ as the expected continuation payoff for type $\theta \in \{old, new\}$ conditional on history h, under beliefs π_d:

$$W_\theta^{\pi_d}(\sigma_d \mid h) := \int W_\theta\left(\sigma_d, \hat{\sigma}_p \mid h\right) d\pi_d\left(\hat{\sigma}_p \mid h\right) \tag{4.6}$$

Analogously, given a system of beliefs π_p the expected utility of strategy σ_p conditional on history h is

$$V^{\pi_p}\left(\sigma_p \mid h\right) := \int V\left(\hat{\sigma}_d, \sigma_p \mid h\right) d\pi_p\left(\hat{\sigma}_d \mid h\right) \tag{4.7}$$

For a given system of beliefs π_i we write $\sigma_i \in SBR_{\theta_i}(\pi_i)$ as the set of sequential best responses of type θ_i to beliefs π_i.[16]

4.4. **Weak and Strong Rationalizability.** In this subsection I introduce the notions of weak and strong rationalizability. The goal is to find strategies that are robust to changes in p's beliefs so that p is induced to trust as long as there is *common certainty of rationality*,

[14] A sequence $\{\sigma_{i,n}\}_{n \in \mathbb{N}}$ converges to σ_i in the product topology in Σ_i if and only if $\sigma_{i,n}(h) \to \sigma_i(h)$ for all $h \in \mathcal{H}_i$

[15] We endow $\Theta_{-i} \times \Sigma_{-i}$ with the Borel $\sigma-$algebra with respect to the product topology.

[16] A strategy σ_p is a *sequential best response to* π_p for all $(h^\tau, c_\tau) \in \mathcal{H}_p$ and all other strategies $\hat{\sigma}_p \in \Sigma_p$, we have $V^{\pi_p}\left(\sigma_p \mid h^\tau, c_\tau\right) \geq V^{\pi_d}\left(\hat{\sigma}_p \mid h^\tau, c_\tau\right)$. Likewise, σ_d is a *sequential best response to belief system* π_d *for type* $\theta \in \{new, bad\}$ if for all histories $h \in \mathcal{H}_d$ and all strategies $\hat{\sigma}_d \in \Sigma_d$ we have $W_\theta^{\pi_d}(\sigma_d \mid h) \geq W_\theta^{\pi_d}(\hat{\sigma}_d \mid h)$.

which means that all agents are rational, all agents are certain that all agents are rational and so on, ad infinitum. In static settings, beliefs that satisfy these common knowledge assumptions have their support over the set of *rationalizable strategies.* This set is characterized by an iterative deletion process described in Pearce [1984]. The set of strategies is refined by eliminating those which are not a best response to *some* beliefs about the other agents strategies, which are themselves best responses to some other beliefs, and so on.

However, the possibility of reaching zero probability events also creates different ways to extend the concept of rationalizability, which hinge upon on our notion of "certainty or rationality". An agent is *certain* about some event E if she believes that this event happens with probability 1.[17] We say that a history $h \in \mathcal{H}$ *is consistent with event* $E \subseteq \Theta_{-i} \times \Sigma_{-i}$ if there exist a strategy $\sigma_{-i} \in \text{proj}_{\Sigma_{-i}} E$ such that $h \in \mathcal{H}(\sigma_{-i})$. Abusing the notation somewhat, I will write $h \in \mathcal{H}(E)$ for such histories.

Definition 4.1 (*Weak Certainty of event* E)**.** A system of beliefs $\pi_i \in \Delta^{\mathcal{H}}(\Theta_{-i} \times \Sigma_{-i})$ is *weakly certain of event* $E \subset \Theta_{-i} \times \Sigma_{-i}$ if $\pi_i\left(E \mid h^0\right) = 1$

Definition 4.2 (*Strong Certainty in event* E)**.** A system of beliefs $\pi_i \in \Delta^{\mathcal{H}}(\Theta_{-i} \times \Sigma_{-i})$ is *strongly certain of event* $E \subseteq \Theta_{-i} \times \Sigma_{-i}$ if $\pi_i(E \mid h) = 1$ for all $h \in \mathcal{H}(E)$

To illustrate the difference between both concepts, suppose p has a belief system π_p, that is certain of some event E, and is also certain about a smaller event $F = \{(new, \sigma_{new}), (old, \sigma_{old})\} \subset E$. That is, he is certain about what strategy each type of player d chooses (which is the required assumption in the construction of a Bayesian equilibrium in pure strategies). However, π_p may be an incorrect prediction of d's behavior. Take a history h in which p realizes that the observed history is *not* consistent with the strategies in F but it is nevertheless consistent with event E: i.e. $\{\sigma_{new}, \sigma_{old}\} \cap \Sigma_d(h) = \emptyset$ but $h \in \mathcal{H}(E)$. If π_p is *weakly certain* of event E, then after the unexpected move by d, no restrictions are imposed on the updated beliefs from history h on. In particular, he is not required to remain certain about event E, even if the observed history is consistent with it. On the other hand, if π_p is *strongly certain* about event E, he would realize his beliefs about event F were wrong, but his updated beliefs would remain certain about event E. In a way, the concept of strong certainty is similar to an agent that knows that event E is true, and her updated beliefs should respect it as a "working hypothesis" (Battigalli and Siniscalchi [2002])

[17] When the event E is also true we say that the type *knows* E. This admits the possibility that an agent believes with probability one an event that is indeed false. In static games, because the game ends right after the payoffs are realized, there is no substantive difference between *certainty* and *knowledge*. In dynamic games the situation is more subtle, since an agent's beliefs may be proven wrong (or refuted) by the observed path of play. Because of this feature, the literature has focused on the concept of certainty (Ben-Porath [1997], Battigalli and Siniscalchi [1999], Penta [2011, 2012]) instead of knowledge, for dynamic games.

These two different notions of certainty will give rise to two different notions of rationalizability. Define the set of sequentially rational outcomes $R_i \subset \Theta_i \times \Sigma_i$ as

$$R_i = \left\{(\theta_i, \sigma_i) : \ \sigma_i \in SBR_{\theta_i}(\pi_i) \text{ for some } \pi_i \in \Delta^{\mathcal{H}}(\Theta_{-i} \times \Sigma_{-i})\right\}. \tag{4.8}$$

The set R_i gives all the strategies and payoff types such that σ_i is the sequential best response to some system of beliefs.

I will now formally follow the iterative procedure of Battigalli and Siniscalchi [2003]. For a given set $E \subset \Theta_{-i} \times \Sigma_{-i}$ write $\mathbf{W}_i(E) \subset \Delta^{\mathcal{H}}(\Theta_{-i} \times \Sigma_{-i})$ to be the set of of beliefs π_i that are weakly certain of E. Analogously, define $\mathbf{S}_i(E) \subset \mathbf{W}_i(E)$ for the set of beliefs that are strongly certain of it. I will denote $WCR_i^k \subset \Theta_i \times \Sigma_i$ and $SCR_i^k \subset WCR_i^k$ as the sets of type-strategy pairs for agent i that are consistent with k rounds of mutual weak (strong) certainty of rationality. For $k = 0$, define

$$WCR_i^0 = SCR_i^0 = R_i.$$

For $k > 1$, define iteratively:

$$WCR_i^k := \left\{(\theta_i, \sigma_i) : \begin{cases} (1): & (\theta_i, \sigma_i) \in WCR_i^{k-1} \\ (2): & \exists \pi_i \in \mathbf{W}_i\left(WCR_{-i}^{k-1}\right) : \sigma_i \in SBR_{\theta_i}(\pi_i) \end{cases}\right\} \tag{4.9}$$

$$SCR_i^k := \left\{(\theta_i, \sigma_i) : \begin{cases} (1): & (\theta_i, \sigma_i) \in SCR_i^{k-1} \\ (2): & \exists \pi_i \in \mathbf{S}_i\left(SCR_{-i}^{k-1}\right) : \sigma_i \in SBR_{\theta_i}(\pi_i) \end{cases}\right\}. \tag{4.10}$$

We start with beliefs that are weakly (strongly) certain of event $E = R_{-i}$ and then we proceed with an iterative deletion procedure, in which the set agent i is weakly (strongly) certain about is the set $E = WCR_i^{k-1}$ and similarly for strong certainty. Finally, the sets of *weak and strong rationalizable outcomes* is defined as

$$WCR_i^\infty = \bigcap_{k \in \mathbb{N}} WCR_i^k \tag{4.11}$$

$$SCR_i^\infty = \bigcap_{k \in \mathbb{N}} SCR_i^k \tag{4.12}$$

The sets WCR_i^∞,$SCR_i^\infty \subset \Sigma_i$ are the sets of strategies for i that are consistent with him having weak (strong) common certainty of rationality. I will denote $\mathcal{B}_i^{WR}$ and $\mathcal{B}_i^{SR}$ as the sets of weak and strong rationalizable beliefs for p, respectively

$$\mathcal{B}_i^{WR} := \Delta^{\mathcal{H}}\left(WCR_{-i}^\infty\right) \text{ and } \mathcal{B}_i^{SR} := \Delta^{\mathcal{H}}\left(SCR_{-i}^\infty\right). \tag{4.13}$$

I will say that a strategy-belief pair (σ_i, π_i) is θ_i-*strong rationalizable* (or simply $p-$strong rationalizable for the case $i = p$) whenever $\pi_i \in \mathcal{B}_i^{SR}$ and $\sigma_i \in SBR_{\theta_i}(\pi_i)$. A history h is θ_i-strong rationalizable whenever $h \in \mathcal{H}(\sigma_i)$ for some weak rationalizable pair (σ_i, π_i). I will refer to such pairs as a θ_i-*strong rationalizations* of h. I also define the analogous notions for weak rationalizability.

4.5. **Discussion.** In the above we have characterized the multiplicity of equilibria in the static game and have established the setup of the repeated game including belief systems and notions of rationalizability. The next section will turn to robust implementation. I briefly connect the concepts now, arguing that equilibrium refinements are not robust to a variety of perturbations we might think of.

First, and most importantly for our applications is robustness to *strategic uncertainty*. Recall that the static game had multiple equilibria. The infinitely repeated game setting only exacerbates this problem. This suggests that in order for either to form predictions or to make policy recommendations, some equilibrium refinement is needed, such as selecting the best equilibrium for the decision maker . In some contexts, this may be reasonable: e.g. if agents could meet and agree upon a desired outcome before the game started and are able to decide both the expected behavior by all agents, the punishments that should be sanctioned to deviators, subject to the constraint that these should be self-enforceable. However, in this environment the public has no reason to agree with the time inconsistent type and, as such, selecting the optimal equilibrium seems suspect (since both types of decision maker have different best equilibria they would like to implement). A second limitation of equilibrium refinements is that they are very sensitive to common knowledge assumptions about the payoff structure of the game. If we allow the set of feasible payoff structures to satisfy a richness condition, [18]and we pick a Nash equilibrium of the game and the belief systems that support it, then arbitrarily small perturbations on the beliefs may pick *any* other weak rationalizable outcome as the unique equilibrium of the perturbed game (e.g., Weinstein and Yildiz [2012, 2007], Penta [2012]). Under these assumptions, the only concept that is robust to small perturbations of beliefs is weak rationalizability, and hence only predictions that hold for all weak rationalizable strategy profiles are robust to these perturbations. [19]However, the richness assumption may be too a stringent condition for our

[18]Formally, for every strategy σ_i there exist a type $\hat{\theta}_i(\sigma_i) \in \hat{\Theta}_i$ such that σ_i is conditionally dominant for type $\hat{\theta}_i(\sigma_i)$ at every history consistent with it: i.e. $W_{\theta_i}(\sigma_i, \sigma_{-i} \mid h) > W_{\theta_i}(\hat{\sigma}_i, \sigma_{-i} \mid h)$ for all $\hat{\sigma}_i \in \Sigma_i, \sigma_{-i} \in \Sigma_{-i}, h \in \mathcal{H}_i(\sigma_i)$.

[19]Weinstein and Yildiz [2012] show that when we relax the restriction that all players know their own type at the beginning of the game (and never abandon this belief), then the only robust solution concept is normal form interim correlated rationalizability (ICR), extending their previous result on static games (Weinstein and Yildiz [2007]).

robustness exercise, since we are ultimately interested in modeling robustness to strategic uncertainty. Strong rationalizability, being a stronger solution concept may not be robust to all of these perturbations in payoff structures, but we briefly study some in Section 8, such as how to choose implementing strategies that are robust in richer payoff type spaces.

One of the main implications of strong rationalizability is that agents can be convinced at some histories that certain payoff types are not consistent with the history observed. Suppose that p reaches a history that is not consistent with both strong common certainty of rationality and $\theta = old$, but it is consistent with $\theta = new$. Strong common certainty of rationality implies that at these histories p must be certain that $\theta = new$ for all strong rationalizable continuation histories; it becomes *common knowledge* that $\theta = new$, and the game transforms in practice to a game of complete information.[20] When this happens, we will say that $\theta = new$ has achieved *full or strong separation* from $\theta = old$. This is one of the key ingredients of robust reputation formation: the reformed decision maker can gain reputation by taking actions that $\theta = old$ decision maker would never take, or that at least would be very costly for her.

5. ROBUST IMPLEMENTATION

This section introduces the notion of robust implementation to a given set of restrictions on p's beliefs (subsection 5.1) and solves for the robust implementing policies for two important benchmarks: weak and strong rationalizable beliefs. Focusing on strong rationalizable implementation, I characterize the optimal strong rationalizable implementation by solving a recursive dynamic contracting problem with a single *implied promise keeping constraint*. Moreover, for histories where robust separation has not occurred, the relevant reputation measure for d is the *implied spot opportunity cost or sacrifice* for $\theta = old$ of playing $r_{\tau-1}$, so only the immediate previous period matters in terms of building partial reputation. I show that on the outcome path of the optimal robust policy, $\theta = new$ gets both partial gains and (endogenous) losses of reputation until robust separation is achieved. After this, the game essentially becomes one with complete information.

5.1. **Definition.** The decision maker has some information about $p's$ beliefs or may be willing to make some assumptions about them. She considers that $p's$ possible beliefs lie

[20] If at some continuation history p observes behavior that is inconsistent with $\theta = new$ playing a strongly rationalizable strategy, p abandons the assumption of strong common certainty of rationality, which then allows him to believe that $\theta = old$ after all. When this happens, we apply the "best-rationalization principle" as in Battigalli and Siniscalchi [2002]. It states that whenever p arrives at such a history, she will believe that there are at least $k-$rounds of strong common certainty of rationality, with k the highest integer for which the history is consistent with k rounds of strong rationalizability.

in some subset $\mathcal{B}_p \subset \Delta^{\mathcal{H}}(\Theta_d \times \Sigma_d)$. Write $SBR_p\left(\mathcal{B}_p\right) = \bigcup_{\pi_p \in \mathcal{B}_p} SBR_p\left(\pi_p\right) \subset \Sigma_p$ as the set of all sequential best responses to beliefs in $\mathcal{B}_p$. We will say that a strategy σ_d is a *robust implementation of trust in* $\mathcal{B}_p$ when it induces p to trust d at all $\tau = 0, 1, 2, ...$, provided d *knows* that **(1)** $p's$ beliefs are in $\mathcal{B}_p$ and **(2)** p is sequentially rational.

Definition 5.1 (*Robust Implementation*)**.** A strategy $\sigma_d \in \Sigma_d$ robustly implements trust in $\mathcal{B}_p$ if, for all histories $(h^\tau, c_\tau) \in \mathcal{H}_p(\sigma_d)$ we have

$$a^{\sigma_p}(h^\tau, c_\tau) = 1 \text{ for all } \sigma_p \in SBR_p\left(\mathcal{B}_p\right)$$

Under the assumptions on the stage game, for a given belief system π_p , its sequential best response will be generically unique. Therefore, if d knows both that p is rational and that he has beliefs π_p, then she can predict the strategy that p will choose.

5.2. **Weak Rationalizable Implementation.** I begin with the most lax notion of rationalizability at our disposal – weak rationalizability; i.e. $\mathcal{B}_p = \mathcal{B}_p^{WR}$. Here I show that this notion of rationalizability delivers a rather stark and, in some sense, negative result: only by eliminating the emergency action entirely can d robustly implement trust. Since the public cares only about the decision maker's strategy, the multiplicity of commitment costs that are consistent with common certainty of rationality allows for the following weak rationalizable beliefs: believe d is rational only if she takes one of these specified decisions but it is actually thought to be irrational if she takes any other. Then it becomes impossible to implement trust in both belief types, unless d gets rid of the emergency action altogether.[21]

Proposition 5.1. *The unique robust implementing policy in* $\mathcal{B}_p = \mathcal{B}_p^{WR}$ *involves* $c_\tau = \infty$ *(i.e.prohibiting the emergency action) every period.*

Thus, the result is that weak rationalizability is too weak a concept to be used for our purposes. After an unexpected commitment cost choice, p could believe d to be *irrational* and never trust d again unless r is removed. This sort of reasoning does not take into account a restriction that, say, if p could find some other beliefs under which d would be rational, then this now becomes p's working hypothesis. This is precisely the notion of strong rationalizability, which I explore below.

5.3. **Strong Rationalizable Implementation.** The next three subsections present the main results of the paper in which I characterize the *optimal strong rationalizable implementation*. I will show that an optimal robust implementing strategy corresponds to the sequential

[21]This argument extends to any game of *private values* with multiple weak rationalizable outcomes. A Bayesian game is of private values if the utility function for each agent depends only on their own payoff parameter, and not about the other agents payoffs.

best response to some strong rationalizable system of beliefs. In that sense, an optimal robust implementing strategy will be equivalent to finding the most pessimistic beliefs that d could have about p's behavior, that is consistent with common strong certainty of rationality. Most importantly, I will also show that any optimal robust strategy will be in fact the *minmax* strategy for d, delivering the best possible utility that d can guarantee at any continuation history, regardless of her system of beliefs.

To simplify notation, denote $\Sigma_i^{SR} = \text{proj}_{\Sigma_i} SCR_i^{\infty}$ for the set of extensive form rationalizable strategies for agent i. Abusing the notation somewhat I will also write $\Sigma_\theta^{SR} = \text{proj}_{\Sigma_d} \left\{ \left(\hat{\theta}, \hat{\sigma}_d \right) \in SCR_d^{\infty} : \hat{\theta} = \theta \right\}$ to represent the set of extensive form rationalizable strategies for type. The goal is to characterize *optimal robust strategies: i.e.* robust strategies that maximize the expected (ex-ante) utility for d, at $\tau = 0$. As a warm up, I solve for the optimal robust strategy in the stage game.

Example 5.1 (**Optimal Robust Strategy in the static game**)**.** Take the stage game of Section 4.1. Then, there exist only two robust commitment cost choices: $c = \overline{c}$ and $c = \infty$, which implies that the optimal robust implementation of trust is $c = \overline{c}$, for either $\theta \in \{new, old\}$. This follows from the argument in the proof of Proposition 4.1: irrespective of the system of beliefs $\pi_p(c)$, if p is certain that he is facing a rational d, he will find it optimal to trust. For $c < \overline{c}$ there always exist strong rationalizable beliefs that induces p not to trust , and for $c > \overline{c}$, p cannot expect to be facing a rational decision maker, since it would have been a dominant strategy just to play $c = \overline{c}$, that induces p to trust and gives d a strictly higher payoff, regardless of her type.

In the repeated game setting, in order to guarantee $p's$ trust we need to make the utility of trust to be greater than the outside option value $\underline{u}_p$ for all strong rationalizable beliefs. Since p is myopic and does not care directly about the commitment cost paid, the only relevant object to determine his expected payoff is the way he expects d to react to shocks at time τ. Define then the set of all strong rationalizable policy functions

$$\mathbf{R}(h^\tau, c_\tau) = \left\{ r(\cdot) = r^{\sigma_d}(h^\tau, c_\tau, \cdot) \text{ for some } \sigma_d \in \Sigma_d^{SR}(h^\tau, c_\tau) \right\} \tag{5.1}$$

Define also $\mathbf{R}_\theta(h^\tau, c_\tau) \subset \mathbf{R}(h^\tau, c_\tau)$ as those policy functions that are $\theta-$rationalizable. Is easy to show that $a^{\sigma_p}(h^\tau, c_\tau)$ for all strong rationalizable strategies if and only if [22] $\int r(z_\tau) U_p(z_\tau) f(z_\tau) dz_\tau \geq$

[22] Because of Fubini's theorem, we can write $\mathbb{E}^{\pi_p}\left[r^{\sigma_d}(h^\tau, z) U_p \mid h^\tau, c_\tau \right] = \mathbb{E}_z \left\{ \mathbb{E}^{\pi_d}_{\tilde{\sigma}_d} \left[r^{\tilde{\sigma}_d}(h^\tau, z_\tau) \mid h^\tau, c_\tau \right] U_p \right\}$ which corresponds to the expected value over a mixed strategy $\hat{\sigma}_d$ with expected policy $\mathbb{E}\left[r^{\hat{\sigma}_d}(h^\tau, c_\tau) \right] = \mathbb{E}_{\tilde{\sigma}_d}\left[r^{\tilde{\sigma}_d}(h^\tau, z_\tau) \mid h^\tau, c_\tau \right]$. Then, the minimum rationalizable payoff of trusting is the one that assigns probability 1 to the worst rationalizable policy function $r(\cdot)$ from the viewpoint of p, on that history

$\underline{u}_p$ for all $r(\cdot) \in \mathbf{R}(h^\tau, c_\tau)$, which can be rewritten in a single condition as:

$$\underline{V}(h^\tau, c_\tau) := \min_{r(\cdot)\in\mathbf{R}(h^\tau,c_\tau)} \int r(z_\tau) U_p(z_\tau) f(z_\tau) dz_\tau \geq \underline{u}_p \tag{5.2}$$

i.e. the worst rationalizable payoff for p must be higher than the reservation utility. In Appendix **??** we show that $\mathbf{R}(h^\tau, c_\tau)$ and $\mathbf{R}_\theta(h^\tau, c_\tau)$ are compact sets and the objective function in the minimization problem of 5.2 is continuous in the product topology, which makes $\underline{V}(h^\tau, c_\tau)$ a well defined object.

The optimal robust strategy $\sigma^*_{new} = \{c^*(\cdot), r^*(\cdot)\}$ for type $\theta = new$ is the strategy that solves the following programming problem:

$$W^*_{new} = \max_{\{c^*(\cdot),r^*(\cdot)\}} \mathbb{E}\left\{(1-\beta)\sum_{\tau=0}^{\infty}\beta^\tau\left[U_p(z_\tau) - c^*(h^\tau)\right] r^*(h^\tau, c^*(h^\tau), z_\tau)\right\} \tag{5.3}$$

subject to

$$\underline{V}(h^\tau, c^*(h^\tau)) \geq \underline{u}_p \text{ for all } h^\tau \in \mathcal{H}\left(\sigma^*_\theta\right) \tag{5.4}$$

and analogously for W^*_{old}. The goal for the rest of the paper is to characterize the solution to 5.3. optimal robust strategy for the reformed payoff type $\theta = new$. Note that restriction 5.4 fully incorporates the robustness restriction into the programming problem. Theorem **??** shows that Σ^{SR}_θ is a compact set, and so are the subsets $\Sigma^{SR}_\theta(h) \subset \Sigma^{SR}_\theta$ of history consistent strategies, for all θ. This implies that existence of payoff functions $\underline{W}_\theta, \overline{W}_\theta : \mathcal{H}_d \to \mathbb{R}$ such that, for all $h \in \mathcal{H}_d$ and $\theta \in \{new, old\}$

$$\underline{W}_\theta(h) \leq W^{\pi_d}_\theta(\sigma_d \mid h) \leq \overline{W}_\theta(h) \text{ for all } \pi_d \in \mathcal{B}^{SR}_\theta, \sigma_d \in SBR_\theta(\pi_d). \tag{5.5}$$

I will refer to $\underline{W}_\theta(\cdot)$ and $\overline{W}_\theta(\cdot)$ as the *best* and *worst strong rationalizable payoffs* for type θ. I will also write $\underline{\mathbb{W}}_\theta := \underline{W}_\theta\left(h^0\right)$ and $\overline{\mathbb{W}}_\theta = \overline{W}_\theta\left(h^0\right)$ for the ex-ante worst (and best) rationalizable payoff, from $\tau = 0$ perspective. The first result relates these bounds to robust implementation. Any optimal robust policy is extensive form rationalizable (i.e. it corresponds to the best response of some rationalizable beliefs) and delivers the worst rationalizable payoff $\underline{W}_\theta(h)$ at all histories and types $\theta \in \{new, old\}$.

Lemma 5.1. *Let σ^*_θ be the optimal robust strategy for type θ. Then $\sigma^*_\theta \in \Sigma^{SR}_\theta$, with rationalizing belief $\underline{\pi}_\theta \in \mathcal{B}^{SR}_\theta$. Moreover, for all histories $h \in \mathcal{H}_d$*

$$W_\theta\left(\sigma^*_d \mid h\right) = \underline{W}_\theta(h),$$

i.e., the optimal robust policy delivers the worst strong rationalizable payoff at all histories.

Lemma 5.1 implies a very important corollary. The optimal robust strategy is the minmax strategies for type $\theta \in \{old, new\}$ (as in Mailath and Samuelson [2006]) across all beliefs that are consistent with common strong certainty of rationality. The beliefs $\underline{\pi}_\theta$ corresponds to the min-max beliefs for type θ, the most pessimistic beliefs that type θ can have about the strategy that p may be playing. Therefore, at history h^τ, the optimal robust policy gives the *best payoff that type θ can guarantee herself*, regardless of her beliefs, as long as p plays some strong rationalizable strategy. This further implies that the value of program 5.3 at any history satisfies

$$W_\theta^* = \underline{\mathbb{W}}_\theta \tag{5.6}$$

Note here that the worst rationalizable payoff does not coincide with the payoff of the worst Bayesian equilibrium of the extensive form game. Common strong certainty of rationality, strictly speaking, is neither a stronger nor weaker solution concept than Bayesian equilibrium.[23].

5.4. **Observed Sacrifice and Strong Rationalizable Policies.** The program 5.3 may seem complicated, because of the potentially complex history dependence of the set of strong rationalizable policies $\mathbf{R}_\theta(h^\tau, c_\tau)$. Since $\mathbf{R}(h^\tau, c_\tau) = \mathbf{R}_{new}(h^\tau, c_\tau) \cup \mathbf{R}_{old}(h^\tau, c_\tau)$, characterizing these sets will determine the shape of $\underline{V}(h^\tau, c_\tau)$. I will derive the restrictions that strong rationalizability, together with the observed history, impose on the set of policy functions $r(\cdot)$ that p may expect, and show that we only need to know the previous period implied opportunity cost paid by type θ, to be able to characterize the set $\mathbf{R}_\theta(h^\tau, c_\tau)$. In this sense, the set of strong rationalizable policies $\mathbf{R}_\theta(h^\tau, c_\tau)$ is Markovian, with a state variable that is observable by all agents in the game.

Consider a history $(h^\tau, c_\tau) \in \mathcal{H}_p$ observed by agent p. Suppose first that p hypothesizes that d is of payoff type θ, and that history h^τ is such that $r_{\tau-1} = 0$ and $U_{\theta,\tau-1} - c_{\tau-1} > 0$, so that d played the normal action in the previous period, but she would have preferred to play the emergency action, if she was of type θ. Let $h^\tau(r = 1)$ be the continuation history had d chosen $r_{\tau-1} = 1$ instead. Then, a $\theta-$rationalizable pair (σ_d, π_d) is consistent with the observed h^τ if and only if

$$\beta W_\theta^{\pi_d}(\sigma_d \mid h^\tau) \geq (1-\beta)(U_{\theta,\tau-1} - c_{\tau-1}) + \beta W_\theta^{\pi_d}(\sigma_d \mid h^\tau(r=1)). \tag{5.7}$$

[23]Applying Aumann and Brandenburger [1995] to the interim normal form game, for a particular Bayesian equilibria to be the predicted outcome of the game, we need the common prior assumption (i.e. both players know $\pi = \Pr(\theta = new)$) together with weak common *knowledge* of rationality and beliefs (i.e. weak common certainty, plus the requirement that the beliefs are correct). While the common certainty of rationality is weaker than strong certainty, this characterization implies a much stronger condition. Agents have common knowledge about the *strategies* that each other will play and these beliefs must be correct.

To interpret condition 5.11, define first $S_{\theta,\tau-1} := U_{\theta,\tau-1} - c_{\tau-1} > 0$ as the *sacrificed spot utility* for type θ of playing $r_{\tau-1} = 0$ instead of $r_{\tau-1} = 1$. Also, let

$$\mathbf{NPV}_{\theta}^{\pi_d}\left(\sigma_d \mid h^{\tau}\right) := \frac{\beta}{1-\beta}\left[W_{\theta}^{\pi_d}\left(\sigma_d \mid h^{\tau}\right) - W_{\theta}^{\pi_d}\left(\sigma_d \mid h^{\tau}\left(r=1\right)\right)\right] \tag{5.8}$$

denote the *net present continuation value* under pair (σ_d, π_d) of having played $r_{\tau-1} = 0$. This formulation gives a very intuitive characterization of condition 5.7:

$$S_{\theta,\tau-1} \leq \mathbf{NPV}_{\theta}^{\pi_d}\left(\sigma_d \mid h^{\tau}\right) \tag{5.9}$$

i.e. it would have been optimal for type θ to "invest" an opportunity cost of utils yesterday (by not following the spot optimal strategy) only if she expected a net present value that would compensate her for the investment. We can further refine condition 5.7 by first showing that

$$W_{\theta}^{\pi_d}\left(\sigma_d \mid h^{\tau}\left(r=1\right)\right) \geq \underline{W}_{\theta}\left(\sigma_d \mid h^{\tau}\left(r=1\right)\right) \geq \underline{\mathbb{W}}_{\theta} \tag{5.10}$$

Combining 5.10 with 5.7 implies a simple *necessary* condition for $\theta-$rationalizability: if (σ_d, π_d) $\theta-$ rationalizes (h^{τ}, c_{τ}), then

$$W_{\theta}^{\pi_d}\left(\sigma_d \mid h^{\tau}\right) \geq \frac{1-\beta}{\beta} S_{\theta,\tau-1} + \underline{\mathbb{W}}_{\theta} \tag{5.11}$$

Condition 5.11 also holds for any other history (h^{τ}, c_{τ}), where we generalize the definition of sacrificed utility as

$$S_{\theta,\tau-1} := \max_{\tilde{r}\in\{0,1\}} \left(U_{\theta,\tau-1} - c_{\tau-1}\right)\tilde{r} - \left(U_{\theta,\tau-1} - c_{\tau-1}\right) r_{\tau-1} \tag{5.12}$$

5.11 puts restrictions on $\theta-$rationalizing pairs (σ_d, π_d) (and hence over policy functions) based only on the previous period outcome, disregarding the information in the observed history up to $\tau - 1$. A striking feature of strong rationalizability is that in fact, 5.11 is also *sufficient*: whether a policy function $r(\cdot)$ pair is strong rationalizable or not depends only on the observed past sacrificed utility. Proposition 5.2 states the core result of this paper.

Proposition 5.2. *Let $(h^{\tau}, c_{\tau}) \in \mathcal{H}_p$ be $\theta-$rationalizable. Then $r(\cdot) \in \mathbf{R}_{\theta}(h^{\tau}, c_{\tau})$ if and only if there exists a measurable function $w : Z \to \left[\underline{\mathbb{W}}_{\theta}, \overline{\mathbb{W}}_{\theta}\right]$ such that*

$$\left(1-\beta\right)\left[U_{\theta}\left(z_{\tau}\right) - c_{\tau}\right] r\left(z_{\tau}\right) + \beta w\left(z_{\tau}\right) \geq \left(1-\beta\right)\left[U_{\theta}\left(z_{\tau}\right) - c_{\tau}\right]\hat{r} + \beta\underline{\mathbb{W}}_{\theta} \tag{5.13}$$

for all $\hat{r} \in \{0,1\}, z_{\tau} \in Z$, and

$$\int \left\{\left(1-\beta\right)\left[U_{\theta}\left(z_{\tau}\right) - c_{\tau}\right] r\left(z_{\tau}\right) + \beta w\left(z_{\tau}\right)\right\} f\left(z_{\tau}\right) dz_{\tau} \geq \frac{1-\beta}{\beta} S_{\theta,\tau-1} + \underline{\mathbb{W}}_{\theta} \tag{5.14}$$

Condition 5.13 is analogous to the Abreu et al. [1990] notion of *enforceability*. A policy $r(\cdot)$ will be "enforceable" at some history only if we can find a continuation payoff function that enforces it on the set of strong rationalizable payoffs $\left[\underline{\mathbb{W}}_\theta, \overline{\mathbb{W}}_\theta\right]$. This argument employs the same tools and insights as those in Abreu et al. [1990]. Condition 5.14 is the translation of condition 5.11 into this notation. It resembles a promise keeping constraint in a dynamic contracting problem: the expected value of following a rationalizable strategy σ_d at this history (given by the right hand side of 5.14) must be greater than the value implied by the implied opportunity cost paid in the previous period, which can be thought of as the utility "promised" by some rationalizable pair (σ_d, π_d). Its proof closely resembles the well known "optimal penal codes" argument in Abreu [1988]: any strong rationalizable outcome can be enforced by switching to the worst rationalizable payoff upon observing a deviation from the prescribed path of play. This means that without loss of generality, we can check whether a policy $r(\cdot)$ is $\theta-$rationalizable if it is implementable whenever type θ thinks that if she deviated, she will have to play the optimal robust policy from then on.

Proposition 5.2 requires the history (h^τ, c_τ) to be $\theta-$rationalizable. In order to be able to use this characterization, we need to determine whether (h^τ, c_τ) is also rationalizable. Because of Lemma 5.1, we know that all histories reached by the optimal robust policy for $\theta = new$ are $new-$rationalizable. Along its path, the observed history may or may not be $old-$rationalizable as well. Determining whether a history is $old-$rationalizable is equivalent to determining whether we have achieved *robust separation*: i.e. if a history is $new-$rationalizable but is not $old-$rationalizable, then p should be certain he is facing $\theta = new$ in the continuation path of the optimal robust policy. Let

$$S_\theta^{\max} := \frac{\beta}{1-\beta}\left(\overline{\mathbb{W}}_\theta - \underline{\mathbb{W}}_\theta\right) \tag{5.15}$$

be the maximum sacrifice level for type θ, that is consistent with common strong certainty of rationality. Proposition 5.3 gives necessary and sufficient conditions for robust separation

Proposition 5.3. *Take a $new-$rationalizable history $(h^\tau, c_\tau) \in \mathcal{H}_p$. Then, it is also $old-$rationalizable if and only if $S_{old,k} \leq S_{old}^{\max}$ for all $k \leq \tau - 1$*

This proposition characterizes completely the conditions for strong separation from type $\theta = old$, along the path of any strong rationalizable strategy, in particular the optimal robust one. The first result that can be inferred from Proposition 5.3 is that *robust separation can never be achieved by the commitment cost decision*. (see Lemma C.2 in Appendix C), and hence $\theta = new$ can only separate from $\theta = old$ based only on how she reacted to the observed shocks. The second result provides a recursive characterization of robust separation: if separation has not yet occurred up to period $\tau - 1$, $\theta = new$ will robustly

separate from $\theta = old$ at period τ if and only if $S_{old,\tau-1} > S_{old}^{\max}$. This happens because condition 5.14 cannot be satisfied for any policy function $r(\cdot)$ and hence $\mathbf{R}_{old}(h^\tau, c_\tau) = \emptyset$. If $S_{old,\tau-1} \leq S_{old}^{\max}$, then at $\tau+1$ the only relevant information to decide whether $h^{\tau+1}$ is $old-$rationalizable is $S_{old,\tau}$, and hence this property is Markovian. Proposition 5.2 shares this Markovian feature: the only relevant information to find the set of strong rationalizable policies $\mathbf{R}_\theta(h^\tau, c_\tau)$ is the observed sacrifice $S_{\theta,\tau-1}$.

5.5. **Characterization of Robust Implementation.** In this subsection I will use the characterization of $\mathbf{R}_\theta(h^\tau, c_\tau)$ of Proposition 5.2 to characterize the worst rationalizable payoff of trusting, $\underline{V}(h^\tau, c_\tau)$. Furthermore, to solve for the optimal robust strategy, I will derive a recursive representation of the optimal robust policy, which will allow us to solve Program 5.3 with a standard Bellman equation, using the familiar fixed point techniques of Stokey et al. [1989]. Suppose that at a given a $\theta-$rationalizable history (h^τ, c_τ), p hypothesizes he is facing type $\theta \in \{new, old\}$.

Using the characterization of Proposition 5.2, we can calculate the minimum utility he can expect from trusting as:

$$\mathcal{V}_\theta(S_{\theta,\tau-1}, c_\tau) := \min_{r(\cdot),w(\cdot)} \int r(z_\tau) U_p(z_\tau) f(z_\tau) dz_\tau \tag{5.16}$$

subject to the incentive compatibility constraint:

$$(1-\beta)[U_\theta(z_\tau) - c_\tau] r(z_\tau) + \beta w(z_\tau) \geq (1-\beta)[U_\theta(z_\tau) - c_\tau]\hat{r} + \beta \underline{\mathbb{W}}_\theta \tag{5.17}$$

$$\text{for all } \hat{r} \in \{0,1\}, z_\tau \in Z$$

the *implied promise keeping* constraint:

$$(1-\beta)\int [U_\theta(z_\tau) - c_\tau] r(z_\tau) f(z_\tau) dz_\tau + \beta \int w(z_\tau) f(z_\tau) dz_\tau \geq \frac{1-\beta}{\beta} S_{\theta,\tau-1} + \underline{\mathbb{W}}_\theta \tag{5.18}$$

and a feasibility constraint for continuation payoffs:

$$w(z_\tau) \in \left[\underline{\mathbb{W}}_\theta, \overline{\mathbb{W}}_\theta\right] \text{ for all } z_\tau \in Z \tag{5.19}$$

At a history that is both *new* and $old-$rationalizable, the worst strong rationalizable payoff of trusting is

$$\underline{V}(h^\tau, c_\tau) = \min\{\mathcal{V}_{old}(S_{old,\tau-1}, c_\tau), \mathcal{V}_{new}(S_{new,\tau-1}, c_\tau)\}$$

Note that $\underline{V}(h^\tau, c_\tau)$ depends on the observed history only through the sacrifices $(S_{old,\tau-1}, S_{new,\tau-1})$, which makes the robust implementation restriction 5.4 to be Markovian. The next proposition completely characterizes $\underline{V}(\cdot)$ for all $new-$rationalizable histories. If incentives between $\theta = old$ and $\theta = new$ satisfy an *increasing conflict* condition, then $\underline{V}(\cdot)$ will be an increasing function of the contemporaneous commitment cost.

Assumption 1 (Increasing Conflict)**.** *Distribution $f(\cdot)$ satisfies the increasing conflict condition if $f\left(U_p, U_{old}\right)$ is non-decreasing in U_{old} when $U_p < 0$ and non-increasing when $U_p > 0$*

Proposition 5.4. *Take a $new-$rationalizable history $h^\tau \in \mathcal{H}$.*

(1) *If $S_{old,k} \leq S_{old}^{\max}$ for all $k \leq \tau - 1$, then*

$$\underline{V}(h^\tau, c_\tau) \geq \underline{u}_p \iff \mathcal{V}_{old}(S_{old,\tau-1}, c_\tau) \geq \underline{u}_p \tag{5.20}$$

(2) *If $S_{old,k} > S_{old}^{\max}$ for some $k \leq \tau - 1$, then there is a unique strong rationalizable continuation strategy $\hat{\sigma}$, which corresponds to the repeated spot optimum; i.e.*

$$c^{\hat{\sigma}}(h^\tau) = 0 \text{ and } r^{\hat{\sigma}}(h^\tau, z_\tau) = \begin{cases} 0 & \text{if } U_{p,\tau} \leq 0 \\ 1 & \text{if } U_{p,\tau} > 0 \end{cases} \tag{5.21}$$

and hence, for such stories

$$\underline{V}(h^\tau, c_\tau) = \begin{cases} \mathbb{E}_z\left[\max\left(0, U_p\right)\right] & \text{if } c_\tau = 0 \\ \mathbb{E}_z\left[min\left(0, U_p\right)\right] & \text{if } c_\tau > 0 \end{cases}$$

(3) *Under assumption 1 $\mathcal{V}_{old}(\cdot)$ is increasing in c_τ.*

Assumption 1 states that when p prefers $r = 0$, then states with higher utility of $r = 1$ for $\theta = old$ are more likely. Proposition 5.4 shows that the implementation restriction can be written as a function of $S_{old,\tau-1}$ only, which makes it the relevant reputation measure. When the implied opportunity cost paid by $\theta = old$ is higher than $S_{old}^{\max}$, the maximum net present value that she could get in the continuation game, the observed history is inconsistent with strong rationalizability (i.e. it is not $old-$rationalizable). At these histories, any system of beliefs must be strongly certain that $\theta = new$ (since there are only two types), and hence robust separation is achieved. This proposition also shows that when this happens, there is a unique strong rationalizable strategy profile, which is to play the repeated spot first best, since there are no conflicts of interest between the parties, and both get their most preferred outcomes (see Lemma C.3).

When $S_{old,\tau-1} < S_{old}^{\max}$, the "promise keeping" condition 5.18 is tighter for higher values of $S_{old,\tau-1}$, since only continuation strategies with a higher net present value are consistent with

the observed history. Therefore, higher sacrifice makes $\mathcal{V}_{old}(S_{old,\tau-1}, c_\tau)$ weakly lower, which in turn relaxes the robust implementation constraint 5.4 in the sequential program 5.3. This observation reinforces the notion of sacrifice being the relevant reputation measure for robust implementation program: higher values relax the implementation constraints, which increases the value of the robust policy.

The basic assumptions made about the distribution of z_τ may allow for local non-monotonicities of $\mathcal{V}_{old}(S_{old,\tau-1}, c_\tau)$ with respect to the commitment cost c_τ. Under the increasing conflict assumption, higher commitment costs increase the minimum utility of facing $\theta = old$. Let $\underline{\mathbf{c}}(s) := \min\left\{c \in C : \mathcal{V}_{old}(s,c) \geq \underline{u}_p\right\}$ be the minimum commitment cost necessary to convince p to trust under all $old-$ rationalizable strategies, if observed past sacrifice for $\theta = old$ is s. Under Assumption 1, we have $\mathcal{V}_{old}(S_{old,\tau-1}, c_\tau) \geq \underline{u}_p$ if and only if $c_\tau \geq \underline{\mathbf{c}}(S_{old,\tau-1})$, which then characterizes the optimal commitment cost choice for the robust strategy under this assumption, since both types want to set c as small as possible.

In Appendix A I study in detail the solution $\left(\underline{r}(\cdot), \underline{w}(\cdot)\right)$ to 5.20. In Proposition A.1 I show that under assumption there exist a threshold $\hat{S} \in \left(0, S^{\max}_{old}\right)$ such that if $S_{old,\tau-1} \leq \hat{S}$, the promise keeping constraint does not bind, and hence it is identical to the solution of 5.20 when $S_{old,\tau-1} = 0$, and $\mathcal{V}_{old}(S_{old,\tau-1}, c_\tau) = \mathcal{V}_{old}(0, c)$. When $S_{old,\tau-1} \in \left(\hat{S}, S^{\max}_{old}\right)$ the promise keeping constraint starts binding, making $\mathcal{V}_{old}(S_{old,\tau-1}, c_\tau)$ strictly increasing in this interval. Figures 1 and 2 illustrate the results.

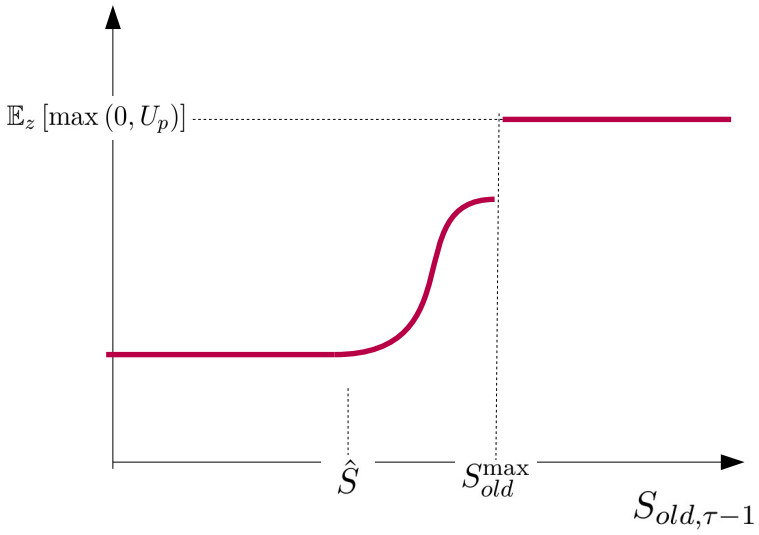

FIGURE 1. Worst Rationalizable Payoff $\mathcal{V}_{old}(s, c)$

FIGURE 2. Optimal Commitment Cost, under Assumption 1

Intuitively, for small observed sacrifices, p cannot discard the possibility that if $\theta = old$, she is expecting to behave the same as if no sacrifice was observed. Therefore, a robust choice for the commitment cost should prescribe exactly the same solution as in $\tau = 0$: the game basically "resets" and all reputation is lost at these histories. For intermediate sacrifices, p still cannot rule out that $\theta = old$, but can nevertheless impose some restrictions on the set of rationalizable strategies, which are stronger the bigger the sacrifice observed. When sacrifice is bigger than the maximum possible rationalizable net present value gain of any continuation value for $\theta = old$, the decision maker achieves strong separation, and hence she knows p is certain $\theta = new$ for all rationalizable continuation strategies; she therefore plays the first best strategy with no commitment costs.

5.6. **Recursive Representation of Optimal Robust Implementation.** Based on the recursive characterization of the implementation restriction, in this section I finally derive a recursive representation of the optimal robust strategy σ^*_{new}. To encode the state of the problem (which depends both on the past sacrifice observed and the rationalizability of the past history) we recursively define the following process: $s_0 = 0$ and for $\tau \geq 1$:

$$s_\tau = \Gamma(s_{\tau-1}, c_\tau, z_\tau, r_\tau) := \begin{cases} \max_{\hat{r}\in\{0,1\}} [U_{old}(z_\tau) - c_\tau]\hat{r} - [U_{old}(z_\tau) - c_\tau] r_\tau & \text{if } s_{\tau-1} \leq S^{\max}_{old} \\ s_{\tau-1} & \text{if } s_{\tau-1} > S^{\max}_{old} \end{cases}$$

The state variable $s_{\tau-1}$ gives the sacrifice for $\theta = old$ as long as history h^τ is $old-$rationalizable. If at some τ the observed history is no longer $old-$rationalizable, then $s_{\tau+k} = s_\tau > S^{\max}_{old}$, so it also indicates when robust separation occurs. Because of Proposition 5.4 the robust implementation restriction can be written as a function of $s_{\tau-1}$ alone: $\underline{V}(h^\tau, c_\tau) \geq \underline{u}_p$ if and only if $\mathcal{V}(s_{\tau-1}, c_\tau) \geq \underline{u}_p$, where

$$\mathcal{V}(s,c) := \begin{cases} \mathcal{V}_{old}(s,c) & \text{if } s \leq S^{\max}_{old} \\ \mathbb{E}_z\left[\max\left(0, U_p\right)\right] & \text{if } s > S^{\max}_{old} \text{and } c = 0 \\ \mathbb{E}_z\left[\min\left(0, U_p\right)\right] & \text{if } s > S^{\max}_{old} \text{and } c > 0 \end{cases} \tag{5.22}$$

With these definitions, Proposition 5.4 allows us to rewrite the optimal robust strategy program 5.3 as:

$$\underline{\mathbb{W}}_{new} = \max_{\{c(\cdot), r(\cdot), s_{\tau-1}(\cdot)\}} (1-\beta)\,\mathbb{E}\left\{\sum_{\tau=0}^{\infty} \beta^\tau \left[U_p(z_\tau) - c(h^\tau)\right] r(h^\tau, z_\tau)\right\} \tag{5.23}$$

$$\text{(5.24)} \qquad \text{s.t. : } \begin{cases} \mathcal{V}\left[s_{\tau-1}\left(h^{\tau}\right), c\left(h^{\tau}\right)\right] \geq \underline{u}_p & \text{for all } h^{\tau} \in \mathcal{H}\left(\sigma^{*}_{new}\right) \\ s_{\tau}\left(h^{\tau+1}\right) = \Gamma\left[s_{\tau-1}\left(h^{\tau}\right), c\left(h^{\tau}\right), z_{\tau}, r\left(h^{\tau}, z_{\tau}\right)\right] & \text{for all } h^{\tau+1} \in \mathcal{H}\left(\sigma^{*}_{new}\right) \end{cases}$$

To get a recursive formulation of $\underline{W}_{new}\left(h^{\tau}\right)$, let $\mathbb{B} = \left\{g : \left[\underline{U}, \overline{U}\right] \to \mathbb{R} \text{ with } g \text{bounded}\right\}$ and define the operator $T : \mathbb{B} \to \mathbb{B}$ as

$$\text{(5.25)} \qquad T\left(g\right)\left(s\right) = \max_{c \in C} \int \left\{ \max_{r(z) \in \{0,1\}} \left(1-\beta\right)\left[U_p\left(z\right) - c\right] r\left(z\right) + \beta g\left[s'\left(z\right)\right] \right\} f\left(z\right) dz$$

subject to

$$\text{(5.26)} \qquad \mathcal{V}\left(s, c\right) \geq \underline{u}_p$$

$$\text{(5.27)} \qquad s'\left(z\right) = \Gamma\left[s, c, z, r\left(z\right)\right] \text{ for all } z \in Z$$

In Lemma C.5 I show T is a contraction with modulus β. Since $\mathbb{B}$ is a complete metric space (when endowed with the sup-norm), we can use the contraction mapping theorem to show the existence of a unique function $\mathcal{W}_{new}\left(\cdot\right)$ that solves the associated Bellman equation $T\left(\mathcal{W}_{new}\right)\left(\cdot\right) = \mathcal{W}_{new}\left(\cdot\right)$; which can expressed as

$$\text{(5.28)} \qquad \mathcal{W}_{new}\left(s\right) = \max_{c \in C : \mathcal{V}(s,c) \geq \underline{u}_p} \int \left\{ \max_{r \in \{0,1\}} \left(1-\beta\right)\left[U_p\left(z\right) - c\right] r + \beta \mathcal{W}_{new}\left[s'\left(z\right)\right] \right\} f\left(z\right) dz$$

subject to 5.26. The term inside the integral is the maximization problem that $\theta = new$ faces after having chosen c and after shock z has been realized: she faces a trade-off between short run utility $\left(1-\beta\right)\left[U_p\left(z\right) - c\right] r$ and reputation gains $\beta \mathcal{W}_{new}\left[s'\left(z\right)\right]$, which depend only on the sacrificed that would be observed at the beginning of the next period. This is possible since once the commitment cost was chosen, there is no restriction linking ex-post utility in different states. The outer maximization choosing the commitment cost function corresponds to the optimal choice of the commitment cost at the beginning of the period. Because of Proposition 5.2 all past history is completely summarized by the sacrificed observed in the previous period . Notice that s only enters the right hand side problem only through restriction 5.26, which only modifies the set of feasible commitment costs.

Proposition 5.5. *Let $c\left(s\right)$ and $r\left(s, z\right)$ be the policy functions associated with the Bellman equation 5.28. Then, for all $h^{\tau} \in \mathcal{H}\left(\sigma^{*}_{new}\right)$*

(1) $\underline{W}_{new}\left(h^{\tau}\right) = \mathcal{W}_{new}\left(s_{\tau-1}\right)$, $c^{*}\left(h^{\tau}\right) = c\left(s_{\tau-1}\right)$ *and* $r^{*}\left(h^{\tau}, c_{\tau}, z_{\tau}\right) = r\left(s_{\tau-1}, z_{\tau}\right)$

(2) *If* $s_{\tau-1} > S_{old}^{\max}$ *then* $c^*(h^\tau) = 0$ *and* $r^*(s_{\tau-1}, z) = \underset{\hat{r}\in\{0,1\}}{argmax} U_p(z)\,\hat{r}$

(3) *If* $s_{\tau-1} \leq S_{old}^{\max}$ *and Assumption 1 holds,* $c^*(h^\tau) = \underline{\mathbf{c}}(s_{\tau-1})$

In the remainder of this section, I solve for the optimal robust policy $r^*(z)$ and the law of motion for the sacrifice process $s'(z)$, under the increasing conflict assumption 1. Figure 4 previews the shape of the optimal policy $r^*(z) = r^*\left(U_p, U_{old}\right)$ over the set of states $Z = \left[\underline{U}, \overline{U}\right]^2 \subset \mathbb{R}^2$. Regions where $r^*\left(U_p, U_{old}\right) = 1$ (i.e. d takes the emergency action) are depicted in red, and $r^*\left(U_p, U_{old}\right) = 0$ in green. In the bottom we include the spot optimum strategy for $\theta = new$ and for agent p. [24] In the right margin, we draw the analogous scale for $\theta = old$.

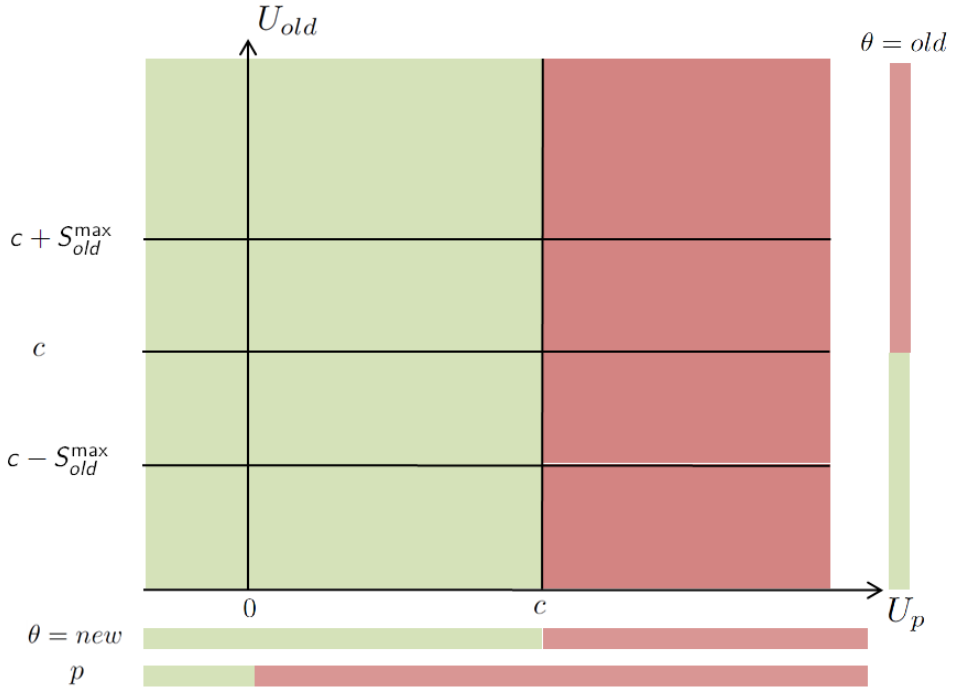

FIGURE 3. Static Best policy $r^{spot}(z)$ for $\theta = new$

[24] The spot optimal policy for p is defined as $r_p^{spot}(z) := 1 \iff U_p \geq 0$. For and for type θ we have $r_\theta^{spot}(z) := 1 \iff U_\theta \geq c$

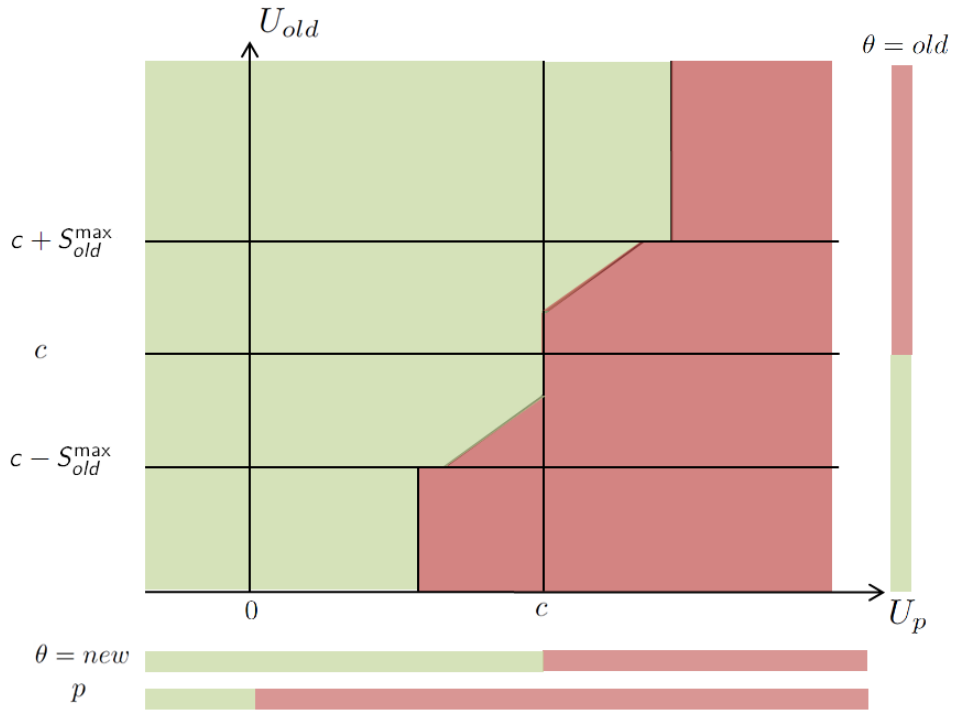

FIGURE 4. Optimal robust policy $r^*(z)$ for $\theta = new$

By comparing the optimal robust policy of Figure 4 to the spot optimal strategy for $\theta = new$ in Figure 3, we see how the optimal robust policy is distorted from the spot optimum towards actions that are spot inefficient for $\theta = old$. For example, consider the region where $U_{old} > c_\tau$ so that $r_{old}^{spot}(z) = 1$. Figure 4 shows then that the optimal policy over this region prescribes $r^*\left(U_p, U_{old}\right) = 0$ on a strictly larger set of states than $r_{new}^{spot}(z)$. The intuition for this phenomenon is simple: if d plays $r_\tau = 0$, then p will observe an opportunity cost paid by $\theta = old$ of $S_\tau = U_{old} - c_\tau > 0$ utils. This will result in a smaller commitment cost at $\tau + 1$ than the one implied by playing $r = 1$ and reseting to the time $\tau = 0$ robust policy from tomorrow on. When the relative reduction in future commitment costs is big enough (i.e. when observed implied sacrifice for $\theta = old$) then the optimal strategy will be to choose $r = 0$. In the rest of this section I will formally characterize both the robust policy $r^*(z)$ and the next period sacrifice $s'(z)$, which governs the reputation formation process.

First, take the region $R_1 = \left\{z \in Z : U_{old} > c_\tau + S_{old}^{\max}\right\}$, which corresponds to the uppermost horizontal strip of Figure 4. For any z in this region, the unique rationalizable policy for $\theta = old$ is to play $r(z) = 1$. This is because if she played $r = 0$, the implied sacrifice $S_\tau = U_{old} - c_\tau$ would be strictly greater than any potential net present value gain from switching to the best rationalizable payoff (given by $S_{old}^{\max}$). Therefore, if $\theta = new$ chooses

$r^*(z) = 0$ she would strongly separate from tomorrow on, achieving the first best payoff $\mathbb{E}_z\left[\max\left(0, U_p\right)\right]$. However, if she chooses $r^*(z) = 0$ then $s'(z) = 0$ and next period the commitment cost gets reset to c_0^*, getting a continuation value of $\underline{\mathbb{W}}_{new}$. Therefore, $r^*(z) = 0$ over region R_1 if and only if

$$\beta\mathbb{E}_z\left[\max\left(0, U_p\right)\right] \geq (1-\beta)\left(U_p - c_\tau\right) + \beta\underline{\mathbb{W}}_{new} \iff$$

$$U_p \leq c_\tau + S_{new}^{\max} \tag{5.29}$$

If $U_p \leq c_\tau$ then by playing $r^*(z) = 0$ type $\theta = new$ would maximize both her spot and her continuation values, achieving strong separation from $\tau + 1$ on. Even when $U_p > c_\tau$, $\theta = new$ could still find it optimal to sacrifice spot gains for the strong separation that would be achieved in the next period. Therefore, when the time inconsistent type has a unique rationalizable strategy, the good type would optimally *invest* in reputation, sacrificing present utility to achieve strong separation in the next period.

Second, take region $R_2 = \left\{z \in Z : U_{old} \in \left(c_\tau + \hat{S}, c_\tau + S_{old}^{\max}\right)\right\}$. In this region, $\theta = old$ preferred strategy is still $r = 1$, but now $r = 0$ is also $old-$rationalizable. By playing $r = 0$, $\theta = new$ cannot achieve separation in the next period, but she still can decrease the commitment cost in the next period to $c_{\tau+1} = c\left(U_{old} - c_\tau\right)$. Therefore, the same analysis from region R_1 applies here, with the only difference that now the continuation value will be $\mathcal{W}_{new}\left(U_{old} - c_\tau\right) < \mathbb{E}_z\left[\max\left(0, U_p\right)\right]$. Then, $r^*(z) = 0$ over region R_2 if and only if

$$U_p - c_\tau \leq \frac{\beta}{1-\beta}\left[\mathcal{W}_{new}\left(U_{old} - c_\tau\right) - \underline{\mathbb{W}}_{new}\right] := \phi\left(U_{old} - c_\tau\right)$$

where $\phi(\cdot)$ is an increasing function of the implied sacrifice $S = U_{old} - c$ of playing $r = 0$ for the time inconsistent type. As before, whenever $U_p < c_\tau$ then $r = 0$ will be optimal. When $U_p > c_\tau$ her decision will depend on two variables: the spot disutility by not choosing $r = 1$ $\left(U_p - c_\tau\right)$ and the reputation value gained by choosing $r = 0$, $\phi\left(U_{old} - c_\tau\right)$. States with very high disutility for $r = 0$ would only prescribe it as an optimal policy for states with high potential sacrifice.

Finally, I study region $R_3 = \left\{z \in Z : U_{old} \in \left(c_\tau, c_\tau + \hat{S}\right)\right\}$. See that regardless of the the action, the sacrifice potential $U_{old} - c_\tau$ is too small to make the commitment cost in the next period to be smaller than its maximum possible level, c_0^*. Therefore, regardless of the policy chosen, in the next period reputation will be lost. The optimal robust policy, then, is just the spot optimal policy: $r^*(z) = 1 \iff U_p > c_\tau$. Is easy to see that when $U_{old} < c_\tau$, then the optimal robust policy analysis will be identical, but with the role of each policy reverted, since it will be now when $r = 1$ that sacrifice may be signaled. In Figure 5 we illustrate

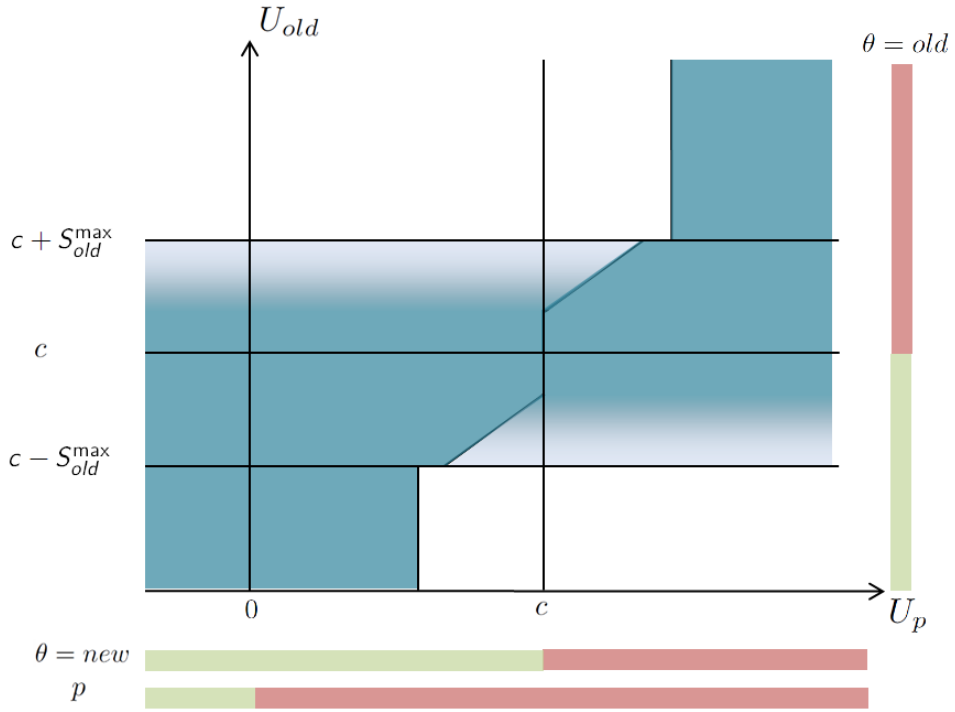

FIGURE 5. Optimal Robust Sacrifice. Dark blue areas show areas where the robust policy implies no sacrifice ($S_{\tau+1} = 0$). Shaded areas show intermediate sacrifice (without robust separation), and white areas show areas where types robustly separate at $\tau + 1$

the map $s'(z)$. The dark blue areas correspond to $s'(z) = 0$ (i.e. all reputation is lost in the next period), and white regions are those in which $\theta = new$ achieves full separation from tomorrow on. In the shaded areas, lighter color illustrate higher sacrifice levels, and hence smaller commitment costs in the next period (but not zero, as in full separation).

5.7. **Discussion.** The reason why the robust policy problem ends up being quite tractable is exactly because of the robustness condition: when having to make sure that p trusts in all histories and for all rationalizable beliefs, the *worst types* (in the sense of beliefs) that p might be facing when deciding whether to trust or not, may correspond to very different beliefs about $d's$ behavior. Because of this disconnect , we are able to separate the problems of commitment cost choice and of the optimal robust policy rule $r^*(z)$. When more restrictions are imposed (for example, a belief set $\mathcal{B} = \{\pi_{new}, \pi_{old}\}$, as in any Bayesian equilibrium), this separation will be broken.

In terms of the optimal robust policy, note that there exist regions where p and both types of decision makers would unanimously prefer certain strategy to be played, but because of reputation building motives $\theta = new$ would still want to do exactly the opposite of the

unanimous optimal decision. For example, when $U_{old} > c_\tau + S^{\max}_{old}$ and $U_p \in (c_\tau, c_\tau + S^{\max}_{new})$ all agents prefer $r = 1$, but the optimal policy prescribes the normal action $r = 0$.

In the context of the capital taxation model, this would correspond with states where the marginal utility of the public good is sufficiently high for both workers and capitalists, so that both household types would agree that the ex-post optimal strategy would be to expropriate. Through the lens of our model, we can summarize the policy maker's decision by the following argument: "Even as a pro-capitalist government, I am tempted to expropriate capitalists. However, the incentives for a benevolent, time inconsistent government to expropriate would be much higher than mine. Therefore, by not expropriating, I can show that I am in fact, not the time inconsistent type". Notice also that regardless of the beliefs that d may have, any strong rationalizable strategy of d should also achieve separation at $\tau + 1$, if she decides to play $r = 0$. This then gives a *robust prediction* about $d's$ behavior, as long it is consistent with common strong certainty of rationality.

Analogously, in the monetary policy game, the monetary authority sets partial commitments to induce low inflation expectations (or low perceived risk of devaluation in small open economies), by setting currency boards laws and other kinds of escape clause costs and trigger policies. To be able to convince markets that the monetary authority is not inflationary, it waits for situations in which a true inflationary type would find it differentially more costly to maintain a low inflation policy than a reformed monetary authority. If the sacrifice of maintaining low inflation (or not devaluating the national currency) is not so large for the reformed authority, it would be willing to "wither the storm", in order to either robustly separate from the old, time inconsistent type (and convincing the markets without a doubt, that she is not the inflationary type), or at least lower the partial commitment costs in the future.

Another application can be on the policy of a banking regulator over time. Suppose at every period, a bank is threatened to "go under", and the banking regulator needs to decide wether to bail it out ($r = 1$) or let it go bankrupt ($r = 0$). Suppose also that this regulator is concerned of being perceived as a "weak" type, that saves banks too often, because of the familiar moral hazard concerns. Such a regulator would wait until some large enough bank is on the brink of collapse and not save it, therefore signaling its "toughness", even if saving the bank was in the strong regulator interest. However, if the bank on the brink of collapse is large enough, even a tough regulator would save it, forfeiting the reputation she has gained before (as illustrated by the memoryless feature of the optimal robust policy).

A troubling feature of the robust policy, perhaps, is the impermanence of reputation gain: only the sacrifice of the previous period matters, but past sacrifices do not provide relevant

information for reputation building. In the next section I find conditions on the set of beliefs $\mathcal{B}$ so that the optimal robust implementation exhibits permanent reputation gains, and hence all past sacrifices give some information about the continuation strategies that the decision maker may be planning to follow.

6. Basic Properties of Strong Rationalizable Implementation

In this section I will study some features of the optimal robust policy. I will first show how present potential sacrifices may affect the distribution of future sacrifices, creating "momentum" for reputation formation. I will also show that the observed sacrifice process achieves almost surely the bound $S_{old}^{\max}$. Hence, by playing the robust policy d will eventually convince p that $\theta = new$, with probability one. Moreover, the speed of convergence to the absorbing complete information stage (where p is certain that $\theta = new$) is exponential, which is also the convergence rate of the best equilibrium of the game.

I also study the asymptotic behavior of the robust policy as both the time consistent and the time inconsistent type become more patient, and show that as the discount rate approaches unity, the worst rationalizable payoff $\underline{\mathbb{W}}_{new}$ converges to the first best payoff, and hence the value of the robust policy converges uniformly to the first best payoff (e.g. for all histories). This further implies that the expected value of any strong rationalizable strategy that $\theta = new$ may follow converges to the first best payoff as well, an analog result to Fudenberg and Levine [1989]

6.1. Dynamics of the optimal robust policy.

We saw in the previous section that only immediate past behavior builds reputation, and past histories are irrelevant. It seems intuitive, however, that there should be some momentum in reputation gaining. The basic idea is that gaining reputation at time τ will lower the commitment cost in the next period. The lowering of commitment cost will allow the reformed type to exploit the difference in ex-post payoffs between both types, which is the source of the difference between her and the time inconsistent decision maker, and therefore making that the commitment cost in $\tau + 1$ should also go down even more, in a probabilistic way. However, the degree of generality that I have been using so far does not allow for an easy characterization of the stochastic process followed by the commitment cost $\underline{\mathbf{c}}(S_{\tau-1})$. Therefore, in this section I will show a somewhat weak momentum result, using a plausible assumption on the primitives of the model.

Assumption 2. *In the static version of the game, we have*

$$\Pr\left(\underset{\tilde{r}\in\{0,1\}}{argmax}\,(U_{old}-\overline{c})\,\tilde{r}=0\right)>\Pr\left(\underset{\tilde{r}\in\{0,1\}}{argmax}\,(U_{old}-\overline{c})\,\tilde{r}=1\right) \tag{6.1}$$

i.e. the optimal static decision rule for $\theta = old$ induces the normal action more often than the emergency action

This assumption further reinforces our interpretation of the normal action ($r = 0$) as the status quo: it is the strategy that not only the reformed decision maker, and a trustworthy old type would play most often. As we saw in the previous section, the main driver of reputation building is the *sacrifice potential* $|U_{old,\tau} - c_\tau|$, a exogenous variable for d given the commitment cost chosen. When the sacrifice potential is high d may decide to invest in reputation building, and moreover, conditional on observing a sacrifice, higher sacrifice potential imply lower commitment cost in the next period. While I cannot provide a characterization of the commitment cost process, I can show that the expected value of the sacrifice potential goes up when the commitment cost decreases.

Proposition 6.1. *Under Assumption 1, if Assumption 2 holds, then*

$$s \geq s' \text{ implies } \mathbb{E}_z\left|U_{old}-\underline{\mathbf{c}}\,(s)\right| \geq \mathbb{E}_z\left|U_{old}-\underline{\mathbf{c}}\,(s')\right| \tag{6.2}$$

so that after higher observed sacrifices, we expect higher potential sacrifices. If $s > s' > \hat{S}$, then the inequality in 6.2 is strict.

For the second result on the dynamics of the optimal robust policy, I will need this very important lemma.

Lemma 6.1. *For all $old-$ rationalizable histories $h^\tau \in \mathcal{H}\left(\sigma_d^*\right)$, we have that:*

$$\Pr\left(U_p > c^*\left(h^\tau\right) - S_{new}^{\max}, U_{old} < c^*\left(h^\tau\right) - S_{old}^{\max}\right) > 0 \tag{6.3}$$

and

$$\Pr\left(S_\tau \geq S_{old}^{\max}\right) > \underline{q} > 0 \text{ for all } h^{\tau+1} \in \mathcal{H}\left(\sigma_d^*\right) \tag{6.4}$$

where

$$\underline{q} := \Pr\left(U_p > c_0^* - S_{new}^{\max}, U_{old} < \overline{c} - S_{new}^{\max}\right) + \Pr\left(U_p < \overline{c} + S_{new}^{\max}, U_{old} > c_0^* + S_{old}^{\max}\right) \tag{6.5}$$

Proof. See Appendix C □

Lemma 6.1 is important on its own, and states first that all the regions considered in the optimal robust policy have positive probability, and hence separation will surely occur.

Moreover, I get a uniform non-zero lower bound on the probability of separating at any history that can be easily calculated. With it, I can show its speed of convergence to strong separation

Proposition 6.2. *For all* $\tau \in \mathbb{N}$

$$\Pr(\textit{Separating before } \tau \textit{ periods}) > 1 - \left(1 - \underline{q}\right)^{\tau} \tag{6.6}$$

Proof. In every trial history (*new* and old- rationalizable) there is at least probability $\underline{q}$ of separating. Since shocks are i.i.d this implies that

$$\Pr\left(S_{old,k} < S_{old}^{\max} \text{ for all } k \leq \tau\right) < \left(1 - \underline{q}\right)^{\tau}$$

and hence $\Pr(\text{Separating before } \tau \text{ periods}) = 1 - \Pr\left(S_{old,k} < S_{old}^{\max} \text{ for all } k \leq \tau\right) = 1 - \left(1 - \underline{q}\right)^{\tau}$ □

This proposition states one of the most important results: the probability of reaching separation from the time inconsistent type is exponentially decreasing in τ. A perhaps even more important corollary is that in fact, for any belief restriction $\mathcal{B}_p$ that is consistent with strong common certainty of rationality, (i.e. $\mathcal{B}_p \subseteq \mathcal{B}_p^s$) $\theta = new$ will also achieve separation in exponential time, the probability of separation can only be higher for any smaller belief sets. The second important corollary is that eventually $S_{old,\tau} > S_{old}^{\max}$ almost surely (and states there after separation), so that d will surely separate eventually from the time inconsistent type.

6.2. **First Best Approximation by patient players.** In this subsection the assumption $\beta_{old} = \beta_{new} = \beta$ will be significant, since I will consider the limit when both discount rates limit 1. I will show that as both types become more patient, the payoff of the robust policy for $\theta = new$ converges to the payoff of the stage game after separation. Now, the probability of separation for history h^{τ} will be denoted as $q(h^{\tau}, \beta)$. I first show that these probabilities are uniformly bound away from zero for all $\delta \in (0,1)$ and all rationalizable histories

Lemma 6.2. *Let* $q(h^{\tau}, \beta)$ *be the ex-ante probability of separation under the optimal robust strategy* σ_d^* *. Then, there exist* $\hat{q} > 0$ *such that* $q(h^{\tau}, \beta) > \hat{q}$ *for all* $h^{\tau} \in \mathcal{H}\left(\sigma_d^*\right), \beta \in (0,1)$

Proof. See Appendix C □

The previous lemma shows the existence of a number $\hat{q} > 0$ such that no matter the discount rate β , the probability of reaching separation in any *new* and $old-$ rationalizable histories is greater than $\hat{q}$. Since shocks are i.i.d, even if history may exhibit time dependence, we can bound the expected time of separation by a geometric random variable with

success probability $\hat{q}$. Since once we reach separation, the unique rationalizable outcome is the First Best (i.e. no commitment, spot optimum policy for $\theta = new$) and the speed of convergence is exponential for this random variable, then for a sufficiently patient decision maker d, the expected payoff of the robust policy will be very close to the first best (i.e. the expected time for separation is very small in utility terms). This is what I show in the following proposition

Proposition 6.3. *Let $\mathbb{E}_z\left\{\max\left(0, U_p\right)\right\}$ be the first best payoff, corresponding to the case where p is certain that $\theta = new$, and let $\underline{\mathbb{W}}_{new}(\beta)$ be the ex-ante expected payoff for the optimal robust policy. Then*

$$\underline{\mathbb{W}}_{new}(\beta) \to \mathbb{E}_z\left\{\max\left(0, U_p\right)\right\} \text{ as } \beta \to 1 \tag{6.7}$$

Proof. For the robust policy, we always can bound it as

$$\underline{\mathbb{W}}_{new}(\beta) \geq \mathbb{E}_\tau\left\{\beta^\tau \mathbb{E}_z\left[\max\left(0, U_p\right)\right]\right\}$$

where $\tau \sim Geom(\hat{q})$. This is true since $\mathbb{E}_z\left\{r\left(h^\tau, z\right)\left(U_p(z) - c\left(h^\tau\right)\right)\right\} \geq 0$ by Lemma C.3. Since we always have that the contemporaneous utility greater than zero and separation is achieved with a probability greater than $\hat{q}$ in any period, we have that this is a lower bound for the robust policy payoff. See also that

$$\mathbb{E}_\tau\left(\beta^\tau\right) = \sum_{\tau=1}^{\infty} \beta^\tau (1-\hat{q})^{\tau-1} \hat{q} = \frac{\hat{q}}{1-\hat{q}} \frac{\beta(1-\hat{q})}{1-\beta(1-\hat{q})} = \frac{\beta\hat{q}}{1-\beta(1-\hat{q})}$$

Therefore

$$\mathbb{E}\left[\max\left(0, U_p\right)\right] - \underline{\mathbb{W}}_{new}(\beta) \leq \mathbb{E}_z\left[\max\left(0, U_p\right)\right]\left(1 - \mathbb{E}_\tau\left(\beta^\tau\right)\right) =$$

$$= \mathbb{E}_z\left[\max\left(0, U_p\right)\right]\left(1 - \frac{\beta\hat{q}}{1-\beta(1-\hat{q})}\right) = \mathbb{E}_z\left[\max\left(0, U_q\right)\right]\left(\frac{1-\beta}{1-\beta(1-\hat{q})}\right) \to 0$$

as $\beta \to 1$. □

If d is patient enough, because of discounting and the exponential speed of convergence to separation, the payoff of the robust policy will be very close to the first best payoff. Therefore, if events of distrust are sufficiently bad (as in our infinite cost interpretation of $p's$ distrust), the risk of using a weaker solution concept may be substantial, if we are not quite sure about the restrictions implied by it, while the potential increase in payoffs would be almost irrelevant if d is patient enough.

6.3. **Discrete Shocks.** One of the simplifying properties of our model is that shocks are absolutely continuous, which makes any particular realization of a shock to have zero probability. As I showed, this implied that the worst continuation value for any $\theta-$rationalizable strategy at a rationalizable history (h^τ, c_τ) is $W_\theta(h^\tau, c_\tau) = \frac{1-\beta}{\beta} S_{\theta,\tau-1} + \underline{\mathbb{W}}_\theta$, which makes observed sacrifice the relevant state variable. However, a general approach in more general settings (e.g. if z_τ is not absolutely continuous) is to define the worst rationalizable payoff $W_\theta(h^\tau, c_\tau)$ as:

$$W_\theta(h^\tau, c_\tau) := \min_{\sigma_d \in \Sigma_\theta^{SR}(h^\tau, c_\tau), \pi \in \mathcal{B}_\theta^{SR}} W_\theta^{\pi_d}(\sigma_d \mid h) \tag{6.8}$$

Lemma C.1 can be adapted here for the general case (for general repeated games), to show that if a history h^τ is $\theta-$rationalizable, then a pair $(\sigma_\theta, \pi_\theta)$ is $\theta-$ rationalizable if and only if $W_\theta^{\pi_\theta}(\sigma_\theta \mid h^\tau, c_\tau) \geq W_\theta(h^\tau, c_\tau)$, which is the equivalent state variable in this setting. In the rest of the section, I characterize the rule of motion for $W_\theta(h^\tau)$ in closed form, for the case where z_τ follows a discrete random process. In this case there is room for longer memory in the reputation formation process (unlike the continuous case, where only the previous period matter), whenever high probability shocks are realized, which may relate the promise keeping and incentive constraints across periods. This result is analogous to the result in Passadore and Xandri [2016].

Formally, suppose $Z = \{z_1, z_2, \ldots, z_k\}$ with probability distribution $p(z_k)$. To illustrate the main ideas, I will consider only $\theta-$rationalizable histories at $\tau = 0$ and $\tau = 1$ (i.e. we have observed (c_0, z_0, r_0) and (c_1, z_1, r_1)). The worst rationalizable payoff for θ at histories h^1 and h^2 must, then, satisfy the incentive constraints in both periods:

$$W_\theta(h^\tau, c_\tau) \geq \frac{1-\beta}{\beta} S_{\theta,\tau-1} + \underline{\mathbb{W}}_\theta \text{ for } \tau = 1, 2 \tag{6.9}$$

Now, unlike the continuous case, we need to make sure that there exist a rationalizable pair $(\sigma_\theta, \pi_\theta)$ that could generate both $W_\theta(h^\tau)$ at $\tau = 1, 2$, which translates to the well known promise keeping condition:

$$\sum_{j=1}^{k} \left[(1-\beta) \left(U_\theta\left(z_j\right) - c_1 \right) r^{\sigma_\theta}\left(z_j\right) + \beta W_\theta^{\pi_d}\left(\sigma_d \mid h^1\right) \right] p\left(z_j\right) \geq W_\theta\left(h^1, c_1\right) \tag{6.10}$$

since $W_\theta\left(h^1, c_1\right)$ is the minimum across rationalizable strategies. In the absolutely continuous case, the fact that any particular shock had zero probability implied that we had almost complete freedom in finding a suitable pair $(\sigma_\theta, \pi_\theta)$: simply define it as the best continuation rationalizable payoff at all $z \neq z_1$, and constraint 6.10 would be trivially true. With a discrete distribution, however, this might not be enough: if the realized shock is such that

$p(z_1)$ is large enough, as is the observed sacrifice at $\tau = 0$, then constraint 6.10 might be violated, and the one period memory result would not hold. I show, however, that the worst rationalizable payoff $W_\theta(h^\tau, c_\tau)$ follows a deterministic, closed form difference equation, which the decision maker can use as a state, to substitute the sacrifice $S_{old,\tau-1}$ in the absolutely continuous case. Here, reputation in the form of higher continuation values might be long-lived, but always has the possibility of being reset; i.e. if a shock is realized with small enough probability, then the worst continuation value is the same as in the continuous distribution case.

Proposition 6.4. *Let (h^τ, c_τ) be a $\theta-$rationalizable history, and $W_\theta(h^\tau, c_\tau)$ be defined as in 6.8. Also, let $\overline{\mathbb{W}}_\theta(c) := \mathbb{E}_z\{\max(U_\theta(z) - c, 0)\}$. Then, $W_\theta(\cdot)$ follows the recursion (for $k \leq \tau$)*

$$W_\theta\left(h^k, c_k\right) = \frac{1-\beta}{\beta} S_{\theta,k-1} + \max\left\{\underline{\mathbb{W}}_\theta, \overline{\mathbb{W}}_\theta - \frac{\overline{\mathbb{W}}_\theta(c_{k-1}) - W_\theta\left(h^{k-1}, c_{k-1}\right)}{\beta p(z_\tau)}\right\} \tag{6.11}$$

Proposition 6.4 gives the relevant state variable at period τ for the robust policy of $\theta = new$: namely, $\mathbf{W}_\tau := W_{old}(h^\tau, c_\tau)$, following the difference equation given by 6.11. See that whenever $p(z_\tau) \leq \left[\overline{\mathbb{W}}_{old}(c_\tau) - \mathbf{W}_\tau\right] / (1-\beta) S^{\max}_{old}$, the state "resets" and the next period minimum continuation value is $\mathbf{W}_{\tau+1} = (1-\beta) S_{old,\tau-1}/\beta + \underline{\mathbb{W}}_{old}$, as in the case with absolutely continuous shocks, which effectively has $p(z_\tau) = 0$ in every period. We then can replicate the same analysis from above with this new state variable: define the optimal commitment cost $c(\mathbf{W}_\tau)$, and the value function $\mathcal{W}_{new}(\mathbf{W}_\tau)$.

7. Alternative Model: Transfers as Commitment Cost

In this section I study a slightly different model, where a decision maker (d) needs to convince an agent (p) to delegate the right to make a decision for both of their sakes, in exchange for an upfront fee $c \geq 0$ offered by the decision maker. This model allows us to consider larger strategy spaces (instead of the binary action model studied above), and is hence useful to highlight the main theoretical mechanisms working behind the optimal robust implementing policy.

Formally, the stage game is as follows: the decision maker offers a transfer of $c \geq 0$ in exchange for the delegation to d of a decision to be made, contingent on an exogenous state to be realized ex post; after that the agent (p) decides whether to trust ($a = 1$) d and accept the transfer, or reject the offer. If p trusts, nature draws a shock $z \in Z$ according to some

absolutely continuous distribution $f(z)$ and the decision maker, takes an action $r \in R$, a compact set from from a compact set R. Preferences are given by $U_\theta := u_\theta(r,z) - c$ for the decision maker, where $\theta \in \{old, new\}$, and $U_p := u_p(r,z) + c$ for the p, where u_{old}, u_{new} and u_p are continuous functions of r and z. This model allows for more general action spaces (I assume only that Z and R are compact subsets of euclidean vector spaces). Shocks are independent and identically distributed (not a crucial assumption) according to a density $f(z)$. For simplicity, I assume that the spot optimal actions $r_i^*(z) := \text{argmax}_{r\in R} u_i(r,z)$ are singleton for all almost all z, with $i \in \{old, new, p\}$, and let $u_i^*(z)$ be its maximum value.

As in the commitment cost model, I assume that there is enough difference between the types of decision makers, such that in the complete information case, the reformed decision maker type $\theta = new$ would not need to pay a fee to convince p to trust her, but $\theta = old$ would; i.e.

$$\int u_p\left(r_{old}^*(z), z\right) f(z)\, dz < \underline{u}_p < \int u_p\left(r_{new}^*(z), z\right) f(z)\, dz \tag{7.1}$$

The second important assumption is that both the old and reformed types of decision maker stand to gain from gaining $p's$ trust; i.e.

$$\mathbb{E}_z\left[u_{old}^*(z)\right] - \overline{c} = \mathbb{E}_z\left[u_{new}^*(z)\right] - \overline{c} > \underline{u}_d \tag{7.2}$$

where $\overline{c} := \underline{u}_p - \mathbb{E}_z\left[\min_{r\in R} u_p(r,z)\right]$. This is a natural upper bound for the ex ante transfer to p to make her trust d, even if he believed she would play as an adversary (minimizing his payoffs ex post). Condition 7.2 also ensures that no robust separation can occur through choices in the transfer choices alone (analogous to the result of Proposition 5.3), since both types would be willing to pay the transfer to induce p to trust her with the decision.

For some results, I will assume $u_{new}(\cdot) = u_p(\cdot)$ as in the commitment cost model, and that that the gain from trusting the new decision maker with same ex post incentives is sufficiently large, even if the agent had to pay the cost herself;

$$\mathbb{E}_z\left\{u_p^*(z)\right\} - \overline{c} > \underline{u}_p \tag{7.3}$$

Condition 7.3 guarantees that, under complete information, there is a unique strong rationalizable outcome when $\theta = new$, which involves choosing $c = 0$ in every period[25]. A sufficient condition for this is that the reformed decision maker has the same preferences than p, even when delegation does not occur; i.e. if $\underline{u}_d = \underline{u}_p$. The analog assumption in the commitment cost model was that $\underline{u}_p < 0$, so that even if commitment cost were arbitrarily high, and the

[25]This is equivalent to assuming $\frac{1}{2}\mathbb{E}\left\{\max_{r\in R} u_p(r,z)\right\} + \frac{1}{2}\mathbb{E}\left\{\min_{r\in R} u_p(r,z)\right\} > \underline{u}_p$

decision maker would not choose $r = 1$ for any realization of z, it would still be optimal for p to trust the decision maker.

I will study an infinitely repeated version of this game, where the decision maker has a permanent payoff type and preferences are given by $U_\theta = (1-\beta)\,\mathbb{E}\{\sum_\tau \beta^\tau [u_\theta(r_\tau, z_\tau) - c_\tau]\}$. She faces a string of myopic delegating agents, which decide to trust ($a(h^\tau, c_\tau) = 1$) if and only if $\mathbb{E}\left[u_p(r_\tau, z_\tau)\right] + c_\tau \geq \underline{u}_p$. As before, we focus on the characterization of the optimal robust implementation of trust for type $\theta = new$; i.e. we want to find a strategy $\sigma = (c(h^\tau), r(h^\tau, z_\tau))$ such that $a^{\sigma_p}(h^\tau, c_\tau) = 1$ for all $\sigma_p \in SBR_p\left(\mathcal{B}_p\right)$. Is easy to see that robust implementation can always be achieved, by choosing $c = \overline{c}$, in every period. This provides an upper bound on the payment for achieving delegation, which shows that the problem of finding the optimal robust policy is well defined. Analogously to the previous model, finding the optimal robust policy boils down to finding a strategy $\sigma^* = \{c^*(\cdot), r^*(\cdot)\}$ to solve the following maximization problem:

$$W^*_{new} := \max_{\{c^*(\cdot), r^*(\cdot)\}} \mathbb{E}\left\{(1-\beta)\sum_{\tau=0}^{\infty}\beta^\tau\left[u_p(r_\tau, z_\tau) - c^*(h^\tau)\right]\right\} \tag{7.4}$$

subject to

$$\underline{V}(h^\tau, c^*(h^\tau)) := \min_{r(\cdot)\in\mathbf{R}(h^\tau, c_\tau)} \int u_p(r(z), z)\, f(z)\, dz + c^*(h^\tau) \geq \underline{u}_p \text{ for all } h^\tau \in \mathcal{H}\left(\sigma^*_\theta\right) \tag{7.5}$$

where $\mathbf{R}(h^\tau, c_\tau) \subseteq R^Z$ is the set of rationalizable policy rules at history (h^τ, c_τ), as before. Analogous to the representation of $\underline{V}(h^\tau, c(h^\tau))$ in 5.16, to find the worst $\theta-$rationalizable belief for each type, at a history where $s = S_{\theta,\tau-1} \leq S^{\max}_\theta$, one needs to solve the following programming problem:

$$\mathcal{V}_\theta(s, c) := \min_{r(\cdot), w(\cdot)} \int u_p(r(z), z) + c$$

subject to

$$(1-\beta)\, u_\theta(r(z), z) + \beta w(z) \geq (1-\beta)\, u_\theta(\hat{r}, z) + \beta\underline{\mathbb{W}}_\theta \text{ for all } z \in Z, \hat{r} \in R \tag{7.6}$$

$$\int \{(1-\beta)[u_\theta(r(z), z) - c] + \beta w(z)\}\, f(z)\, dz \geq \frac{1-\beta}{\beta} s + \underline{\mathbb{W}}_\theta \tag{7.7}$$

The worst rationalizable payoff can then be expressed as $\underline{V}(h^\tau, c_\tau) = \min_\theta \mathcal{V}_\theta(S_{\theta,\tau-1}, c_\tau)$, minimizing over the types for which history (h^τ, c_τ) is $\theta-$rationalizable. As in the commitment cost model, $\mathcal{V}_\theta(s, c)$ is weakly increasing in s, for all types. Moreover, V_θ is *always strictly increasing* in c, without imposing any conditions (as we had to do in the credible

policy model above). The reason behind this is simple: firstly, agent p benefits directly from increases in c, (this was not true in the commitment cost model, where p would only benefit from distortions in the payoffs of the decision maker, and hence distorting her decision making). Secondly, it does not affect the incentive constraint 7.6, since payments have to be made regardless of the decision taken. Moreover, it affects the implied promise keeping constraint 7.7 in almost the exact same way than sacrifice, so agent p not only trusts a decision maker from the direct effect of the payment, but also potentially from the strong certainty of rationality effect embedded in 7.7.

It is also easy to show that if history (h^τ, c_τ) is $old-$ rationalizable, then $\underline{V}(h^\tau, c_\tau) = \mathcal{V}_{old}(S_{old,\tau-1}, c_\tau)$ just as in the commitment cost model. Since any decision maker wants to make the participation constraint of the p binding, she will always choose c_τ to make p indifferent between trusting or not, under her worst strong rationalizable beliefs; i.e she will offer a payment of

Let $\underline{\mathbf{c}}_\theta(S_{\theta,\tau-1})$ be the minimum transfer required, under complete information, to robustly implement $a = 1$ if the implied sacrifice in the last period is $S_{\theta,\tau-1}$:

$$\underline{\mathbf{c}}_\theta(s_\theta) := \begin{cases} \min\left\{c : \mathcal{V}_\theta(s_\theta, c) \geq \underline{u}_p\right\} & \text{if } s_\theta \leq S_\theta^{\max} \\ 0 & \text{if } s_\theta > S_\theta^{\max} \end{cases} \tag{7.8}$$

The next Lemma characterizes the choice of c in the optimal robust policy, for any history that is rationalizable for both types of decision maker.

Lemma 7.1. *Let h^τ be a $\{old, new\}$ rationalizable history. In the optimal robust policy for both types, the optimal payment is*

$$c\left(h^t\right) = \underline{\mathbf{c}}(S_{old,\tau-1}, S_{new,\tau-1}) := \max\left\{\underline{\mathbf{c}}_{old}(S_{old,\tau-1}), \underline{\mathbf{c}}_{new}(S_{new,\tau-1})\right\} \tag{7.9}$$

A history $(h^\tau, c_\tau) \in \mathcal{H}(\sigma^*)$ is $\theta-$rationalizable if and only if $S_{\theta,k} \leq S_\theta^{\max}$ for all $k \leq \tau - 1$. If $\mathcal{V}_{new}(0,0) > \underline{u}_p$, we have $\underline{\mathbf{c}}_{new}(s) = 0$ for all $s \in [0, S_{new}^{\max}]$, and hence $c(h^t) = \underline{\mathbf{c}}_{old}(S_{old,\tau-1})$. If $u_{new}(\cdot) = u_p(\cdot)$ and condition 7.3 is satisfied, then $\mathcal{V}_{new}(0,0) > \underline{u}_p$

Proof. See Appendix C □

See that sacrifice does not incorporate information about past payments, because, from the point of view of the decision maker, it is a sunk cost. When deciding wether to trust or not, however, p internalizes the fact that d must be expecting to benefit from this investment (i.e. the payment of c), which is the same mechanism underlying the models with "money burning" (Kolberg and Mertens [1986], Mailath et al. [1993]). Lemma 7.1 allows one to write the optimal robust policy problem recursively, using $s = (S_{old,\tau-1}, S_{new,\tau-1})$ as the

relevant state variable:

(7.10)
$$\mathcal{W}_\theta(s_{old}, s_{new}) = \max_{r(\cdot)\in R^Z} \mathbb{E}_z\left\{(1-\beta)\left[u_\theta(z, r(z)) - \underline{\mathbf{c}}\left(s'_{old}, s'_{new}\right)\right] + \beta\mathcal{W}_\theta\left(s'_{old}, s'_{new}\right)\right\}$$

where

$$s'_\theta = \Gamma(s_\theta, z, r) := \begin{cases} \max_{\hat{r}\in R} u_\theta(z, \hat{r}) - u_\theta(z, r(z)) & \text{if } s \le S^{\max}_\theta \\ s & \text{if } s > S^{\max}_\theta \end{cases} \quad \text{for } \theta \in \{old, new\} \tag{7.11}$$

The dynamics of 7.11 are the same as in the commitment cost model. The main advantage of this model is that the optimal policy is static, in that the decision maker, playing the robust optimal policy, follows a stationary policy (when choosing r) while robust separation is not achieved. This is a consequence of the commitment cost c being a sunk cost at the moment of choosing the policy r, and hence the optimal robust policy will be forwards looking. The characterization of the robust optimal policy is presented in the following proposition, which has a quasi closed form solution:

Proposition 7.1. *Suppose h^τ is $\{old, new\}-$ rationalizable, and let $(c^*(h^\tau), r^*(h^\tau, z_\tau))$ be the optimal robust policy for $\theta = new$. Then $c^*(h^\tau) = \underline{\mathbf{c}}(S_{old,\tau-1})$ and $r^*(h^\tau, z_\tau)$ solves the maximization program:*

$$W(z_\tau) := \max_{r\in R} u_{new}(r, z_\tau) - \beta\underline{\mathbf{c}}\left(s'_{old}, s'_{new}\right) \text{ subject to } 7.11 \tag{7.12}$$

and the value function of the optimal robust policy is $\mathcal{W}_{new}(s_{old}, s_{new}) = \mathbb{E}_z[W(z_\tau)] - (1-\beta)\underline{\mathbf{c}}(s_{old}, s_{new})$. If $\mathcal{V}_{new}(0,0) > \underline{u}_p$ then $\mathcal{W}_{new}(s_{old}, s_{new}) = \mathcal{W}_{new}(s_{old})$ and $\underline{\mathbf{c}}(s_{old}, s_{new}) = \underline{\mathbf{c}}_{old}(s_{old})$.

Proof. Is easy to show that the functional equation in 7.10 has a unique solution, using the contraction mapping theorem. By guessing and verifying the functional form $\mathcal{W}_{new}(s_{old}, s_{new}) = A - (1-\beta)\underline{\mathbf{c}}(s_{old}, s_{new})$ we get $A = \mathbb{E}_z\{W(z_t)\}$. If $u_{new}(\cdot) = u_p(\cdot)$ and 7.3 is satisfied, we know from Lemma 7.1 that $\underline{\mathbf{c}}(s_{old}, s_{new}) = \underline{\mathbf{c}}(s_{old})$, which then implies that $\mathcal{W}_{new}(s_{old}, s_{new}) = \mathcal{W}_{new}(s_{old})$, finishing the proof. □

Proposition 7.1 shows the tradeoff faced by the new decision when choosing how to respond to shocks. When $u_p(\cdot) = u_{new}(\cdot)$ and $\underline{\mathcal{V}}_{new}(0,0) > \underline{u}_p$, the new decision maker trades off her contemporaneous incentives, $u_p(r, z)$, with the effect it has on the next period cost it needs to pay in order to get $p's$ trust, which is solely a function of the implied sacrifice $\theta = old$ would have made if she was the true type of decision maker playing. Whenever robust separation is achieved, we have $\underline{\mathbf{c}} = 0$ and hence the new decision maker optimizes statically in every period, analogous to the commitment cost model above.

8. Extensions And Further ResearcH

8.1. **Legislative Approach.** I now address several extensions of the model and strategies for future research. First, a natural alternative is to take a *legislative approach*, as in Athey et al. [2005] and Persson and Tabellini [1993]. The policy maker may have delegated the commitment choice to the public. The idea here is that if one delegates the commitment cost to the public then certainly one will have robust implementation. The relevant source of uncertainty in the problem is that the public mistrusts the government. The intuition comes from contract theory: we should give control rights precisely to the party who has the first-order inability to trust. However, this will come at a cost in terms of efficiency. Specifically, the public would always put a higher commitment cost, to make the optimal policy for $\theta = old$ not drive him to indifference between trusting or not. As such, the public would increase commitment costs relative to the levels chosen by the new regime government. Formally, given beliefs $\pi_p \in \Delta(\Sigma)$Therefore, it is easy to show that, if the government has the same robustness concerns, then the executive approach is superior for her in terms of welfare, given their information.

8.2. **Multiple Types.** Second, we may consider robustness to not just a single time inconsistent "old type" but a multitude of time inconsistent types. Is straightforward to see that Proposition (5.2) would still be true for any type space Θ_d and hence the characterization of $\underline{V}(h^\tau, c_\tau)$ would now be:

$$\underline{V}(h^\tau, c_\tau) = \min_{\theta \in \Theta_d} \mathcal{V}_\theta(S_{\theta,\tau-1}, c_\tau) \tag{8.1}$$

where the function $\mathcal{V}_\theta(c, s)$ is the minimum problem in (5.20) for a given payoff type θ. Therefore, this will be equivalent to our dynamic contracting characterization of the problem above but with multiple types. In the case of a finite type set $\Theta_d = \{\theta_1, \theta_2, ..., \theta_k\}$ where we now have the vector of observed sacrifices $S_{\tau-1} = (S_{\theta_1,\tau-1}, S_{\theta_1,\tau-1}, ..., S_{\theta_k,\tau-1})$ as the state variables for the implied promise keeping constraints. The solution would exhibit separation from certain types across time, and if the other types satisfy the same assumptions made about $\theta = old$, then it will also eventually convince p about her being the time consistent type.

A third extension is to an environment in which d has an imperfect signal about $p's$ perceived incentives of the time inconsistent type. If signals are bounded and its support may be affected by some signal that d observes, then robust policy would be qualitatively identical.

Finally, looking forward, I would like to extend our analysis to situations in which there are a continuum of strategies and policies available. This will allow researchers to apply

this robust modeling approach to various macroeconomic applications of interest, as the inflation setting model of subsection (2.2). Is easy to see how Proposition (5.2) would remain valid on more general models, so that the Markovian nature of reputation formation would be a very general characteristic of this type of robustness.

9. CONCLUSIONS

I have studied the problem of a government with low credibility. A government faces ex-post time inconsistent incentives due to lack of commitment, such as an incentive to tax capital or an incentive to allow for undesirably high levels of inflation. The government undergoes a reform in order to remove these incentives; however, the reform is successful only if the public actually believes that the government has truly reformed its ways. As such, the crux of the problem relies on the government building reputation in the eyes of the public.

After arguing that the typical approach to this problem relies on equilibrium concepts, which are highly sensitive to small perturbations about the public's beliefs, I turned to studying the problem through the lens of optimal robust policy that will implement the public's trust over any rationalizable belief that any party can hold. Focusing on robustness to all extensive-form rationalizable beliefs, I characterize the solution as well as the speed of reputation acquisition.

This is a particularly desirable property from the point of view of macroeconomic mechanism design. Equilibrium type solution concepts rely on every party knowing every higher order belief of every other party involved in the interaction. This is an extremely high dimensional object and in all likelihood it may be very difficult to believe that such an assumption really holds in settings in which one agent is trying to convince the other agent that he is not adversarial. Furthermore, equilibrium concepts rely on high dimensional belief functions off the path of play – that is, nodes or histories that may never be reached. This sort of sensitivity is problematic when advising a policy maker as small deviations in how a party truly conjectures some off the path of play belief may severely affect the policy maker's ability to obtain trust. This sort of analysis, studying optimal robust policy, can be a very powerful tool within macroeconomic policy making.

APPENDIX A. TYPE SPACES

See online version at Credible Reforms: A Robust Implementation Approach

APPENDIX B. UNIVERSAL TYPE SPACE AND STRONG RATIONALIZABLE STRATEGIES

See online version at Credible Reforms: A Robust Implementation Approach

APPENDIX C. PROOFS AND SUPPLEMENTARY RESULTS

I will need some extra notation for the proofs in this section. Given an appended history $h^s = \left(h^\tau, h^k\right)$, I write $h^s \sim h^\tau = h^k$ for the tail of the history. Also, whenever we can decompose h^s in this manner, I will say that h^τ *precedes* h^s and write $h^\tau \prec h^s$.

Proof of Proposition 4.1. Any equilibrium must induce p to trust since d can always choose $c > \overline{U}$ (so it will never be optimal to take the emergency action) and get a payoff of $0 > \underline{u}_d$. There cannot be any pooling equilibrium with $c < \underline{c}(\pi)$, since the definition of $\underline{c}(\pi)$ implies it would give p less than his reservation utility $\underline{u}_p$. It cannot happen either if $c > \overline{c}$, since either type would deviate and choose $\overline{c}$ and induce p to trust, regardless of his updated beliefs $\pi_p(\overline{c})$. This follows from

$$\pi_p(\overline{c}) \int_{U_p > \overline{c}} U_p f(z)\, dz + \left[1 - \pi_p(\overline{c})\right] \int_{U_{old} > \overline{c}} U_p f(z)\, dz \underset{(1)}{\geq}$$

$$\pi_p(\overline{c})\, \mathbb{E}\left[\max\left(U_p - \overline{c}, 0\right)\right] + \left[1 - \pi_p(\overline{c})\right] \underline{u}_p \underset{(2)}{\geq} \pi_p(\overline{c})\, \underline{u}_p + \left[1 - \pi_p(\overline{c})\right] \underline{u}_p = \underline{u}_p \tag{C.1}$$

where (1) follows from definition 4.3 and (2) from the fact that $0 > \underline{u}_p$. I will now show that for any $\hat{c} \in \left[\underline{c}(\pi), \overline{c}\right]$ there exist a pooling equilibrium in which both $\theta = new$ and $\theta = old$ find it optimal to choose $c = \hat{c}$. Conjecture the following belief updating rule:

$$\pi_p^{\hat{c}}(c) := \begin{cases} 0 & \text{if } c < \hat{c} \\ \pi & \text{if } c \geq \hat{c}. \end{cases} \tag{C.2}$$

Under a pooling equilibrium, since $\hat{c} \geq \underline{c}(\pi)$, p will trust d. Neither type will deviate from $\hat{c}$ since the optimal deviation that would make p trust would be to choose $\hat{c} = \overline{c}$. The non-existence of other PBE is left to the appendix.To finish the proof of this proposition, we need to show there is no separating or semi-separating equilibria. Suppose there exist a separating equilibrium with (c_{new}, c_{old}) with $c_{new} \neq c_{old}$. Since $\underline{u}_d \approx -\infty$, we know that both types induce p to trust, on the equilibrium path. If $c_{old} > c_{new}$, $\theta = old$ can imitate $\theta = new$ by choosing $c = c_{new}$, inducing $p's$ trust, and getting strictly higher utility ex post. Same reasoning follows if $c_{old} < c_{new}$, which implies that in any equilibrium, we must have $c_{new} = c_{old}$. We cannot have mixing in equilibrium, since both types have monotonically decreasing preferences over commitment cost choices, and hence cannot be ex-post indifferent among different cost choices.

The first part is a consequence of Lemma (C.3) in Appendix C. For the second result, take a robust and strong rationalizable strategy σ_d and suppose there exist a history h and a strong rationalizable pair $(\hat{\sigma}_d, \hat{\pi}_d)$ that deliver an expected payoff that is less than the payoff

of the robust policy:

$$W_\theta^{\hat{\pi}_d}(\hat{\sigma}_d \mid h) < W_\theta(\sigma_d \mid h).$$

However, if $\hat{\pi}_d$ has common strong certainty of rationality, then she is also certain that p plays strong rationalizable strategies (Proposition 3.10 in Battigalli and Bonanno [1999]), and hence she should be also certain that by following the robust strategy σ_d from history h on she will get a higher expected payoff. Since this is true for any rationalizable belief, $\hat{\sigma}_d$ cannot be the sequential best response for beliefs $\hat{\pi}_d$ (since it is conditionally dominated by σ_d at h), reaching a contradiction □

Lemma C.1. *Take a history h^τ and $\theta-$rationalization (σ_d, π_d). Also, let $\nu = (\hat{\sigma}_d, \hat{\pi}_d)$ be another $\theta-$rationalizable pair that satisfies:*

$$W_\theta^{\hat{\pi}_d}\left(\hat{\sigma}_d \mid h^0\right) \geq \frac{1-\beta}{\beta} S_{\theta,\tau-1} + \underline{\mathbb{W}}_\theta \tag{C.3}$$

Then, there exists another pair $\left(\sigma_d^\nu, \pi_d^\nu\right)$ that also $\theta-$rationalizes h^τ and is such that

$$\sigma_d^\nu(h^s) = \hat{\sigma}_d(h^s \sim h^\tau), \pi_d^\nu(\cdot \mid h^s) = \hat{\pi}_d(\cdot \mid h^s \sim h^\tau) \tag{C.4}$$

for all histories $h^s \succ h^\tau$

Proof. Define the pair $\left(\sigma_d^\nu, \pi_d^\nu\right)$ for any history $\tilde{h}^s$ as

$$\sigma_d^\nu\left(\tilde{h}^s\right) := \begin{cases} \sigma_d\left(\tilde{h}^s\right) & \text{if } s < \tau \text{ or } \tilde{h}^s = h^\tau \\ \sigma_\theta^*\left(\tilde{h}^s \sim \tilde{h}^\tau\right) & \text{if } s \geq \tau \text{ and } h^\tau \not\prec \tilde{h}^s \\ \hat{\sigma}_d\left(\tilde{h}^s \sim h^\tau\right) & \text{if } s \geq \tau \text{ and } h^\tau \prec \tilde{h}^s \end{cases} \tag{C.5}$$

and for any measurable set $A \subset \Sigma_p$

$$\pi_d^k\left(A \mid \tilde{h}^s\right) := \begin{cases} \pi_d\left(A \mid \tilde{h}^s\right) & \text{if } s < \tau \\ \underline{\pi}_\theta\left(A \mid \tilde{h}^s \sim \tilde{h}^\tau\right) & \text{if } s \geq \tau \text{ and } h^\tau \not\prec \tilde{h}^s \\ \hat{\pi}_d\left(A \mid \tilde{h}^s \sim h^\tau\right) & \text{if } s \geq \tau \text{ and } h^\tau \preceq \tilde{h}^s \end{cases} \tag{C.6}$$

so the pair $\left(\sigma_d^\nu, \pi_d^\nu\right)$ coincides with (σ_d, π_d) for any histories of length less than $\tau - 1$, and strategies also do it up to time τ. If at history $\left(h^{\tau-1}, c_{\tau-1}, a_{\tau-1}, z_{\tau-1}\right)$ d deviates from $r = r^{\sigma_d}\left(h^{\tau-1}, z_{\tau-1}\right)$ going to h'^τ, then type θ believes that she will switch to the optimal strong rationalizable strategy from then on, to which the best response is σ_θ^* and the expected payoff is

$$W_\theta^{\pi_d^\nu}\left(\sigma_d^\nu \mid h'^\tau\right) = W_\theta^{\underline{\pi}_\theta}\left(\sigma_d^* \mid h^0\right) = \underline{\mathbb{W}}_\theta$$

which is a rationalizable continuation pair. Same is true for the continuations at all histories after h^τ, and so the pair $\left(\sigma_d^\nu, \pi_d^\nu\right)$ is rationalizable. Then, to finish our proof, we need to show that it is consistent with h^τ only at $r_{\tau-1}$. Consider first the case where $r_{\tau-1} = 0$ and $S_{\theta,\tau-1} = U_{\theta,\tau-1} - c_{\tau-1} > 0$. Then, the optimal choice under $\left(\sigma_d^k, \pi_d^k\right)$ is

$$\beta W_\theta^{\pi_d^\nu}\left(\sigma_d^\nu \mid h^\tau\right) \geq (1-\beta)\left(U_{\theta,\tau-1} - c_\tau\right) + \beta \underline{\mathbb{W}}_\theta \iff$$

$$W_\theta^{\hat{\pi}_d}\left(\hat{\sigma}_d \mid h^0\right) \geq \frac{1-\beta}{\beta} S_{\theta,\tau-1} + \underline{\mathbb{W}}_\theta$$

which is the assumption made in C.3. The other cases are shown in a similar fashion. □

Proof of Proposition 5.2. Given the functions $(r(\cdot), w(\cdot))$ that satisfy conditions 5.13 and 5.14, I need to construct a $\theta-$rationalizable pair (σ_d, π_d) such that $r^{\sigma_d}(h^\tau, z_\tau) = r(z_\tau)$ for all $z \in Z$. Because the set or rationalizable payoffs is convex, we know that for any $w \in \left[\underline{\mathbb{W}}_\theta, \overline{\mathbb{W}}_\theta\right]$ there exist some rationalizable pair (σ_w, π_w) such that

$$W_\theta^{\pi_w}\left(\sigma_w \mid h^0\right) = w$$

then, for all $z \in Z$ we can find a rationalizable pair $(\hat{\sigma}_z, \hat{\pi}_z)$ such that

$$W_\theta^{\hat{\pi}_z}\left(\hat{\sigma}_z \mid h^0\right) = w(z) \tag{C.7}$$

which are rationalizable continuations from time 0 perspective. Moreover, see that that $r(z)$ solves the IC constraint 5.13 for this continuations, which means that it would be the best response at $\tau = 0$ if θ expected the continuation values $w(z)$ starting from $\tau = 1$. Formally, let $h^1(z) = (c_0, a_0, z_0 = z, r_0 = r(z))$ and define the strategy σ_0 as

$$\sigma_0(h^\tau) = \begin{cases} (c_\tau, r(\cdot)) & \text{if } h = h^0 \\ \sigma_z\left(h^s \sim h^1(z)\right) & \text{if } h^1(z) \prec h^s \\ \sigma_\theta^*\left(h^s \sim h^1\right) & \text{otherwise} \end{cases}$$

i.e. upon deviations in the first period, goes to the optimal robust strategy, and by following the proposed policy $r(z)$ it continues prescribing strategy σ_z after that history, which gives an expected payoff of $w(z)$. This then implies that the policy function is $\theta-$rationalizable at h^0, and that it's expected payoff is

$$W_\theta^{\pi_0}\left(\sigma_0 \mid h^0\right) = \mathbb{E}_z\left[(1-\beta) r(z)\left(U_\theta - c_\tau\right) + \beta w(z)\right] \geq \frac{1-\beta}{\beta} S_{\theta,\tau-1} + \underline{\mathbb{W}}_\theta$$

But then we can use Lemma C.1 for the pair $(\hat{\sigma}_d, \hat{\pi}_d) = (\sigma_0, \pi_0)$, finishing the proof. □

To show Proposition 5.3 we will need the following Lemma

Lemma C.2 (**No strong separation by commitment costs**). *Take a history h^τ that is strong rationalizable for both types, and a commitment cost $\hat{c}$ such that $(h^\tau, \hat{c})$ is new−rationalizable. Then, $(h^\tau, \hat{c})$ is old−rationalizable as well.*

Proof. Suppose not. Then, at history $(h^\tau, \hat{c})$ type $\theta = new$ would achieve robust separation. I will now construct a system of beliefs $\pi \in \mathcal{B}_d^s$, for any continuation history h :

$$\pi(A \mid h) = \begin{cases} 1 & \text{if } h \succ (h^\tau, \hat{c}) \text{ and } \sigma_p^{FB} \in A \\ 1 & \text{if } h \not\succ (h^\tau, \hat{c}) \text{ and } \underline{\sigma}_p \in A \\ 0 & \text{otherwise} \end{cases} \tag{C.8}$$

where σ_p^{FB} is the first best strategy for p if he faces $\theta = new$, and $\underline{\sigma}_p(h) = 0$ for all histories (i.e. not trust for all continuation histories). See that because of robust separation, for any continuation history h that is new−rationalizable, this will be a rationalizable strategy if p puts measure 1 on $\theta = new$. If a continuation history h is not new−rationalizable, then because we assumed it is not old−rationalizable either, then strong rationalizability puts no restrictions on beliefs after such histories, and hence $\underline{\sigma}_p$ is a strong rationalizable continuation strategy at these histories. Define $\hat{\sigma}_d$ as

$$\hat{\sigma}_d(h) = \begin{cases} \left(\hat{c}, r_{old}^{spot}(\cdot \mid \hat{c})\right) & \text{if } h = h^\tau \\ \left(0, r_{old}^{spot}(\cdot \mid c = 0)\right) & \text{if } h \succ (h^\tau, \hat{c}) \\ \left(\infty, r_g(\cdot)\right) & \text{if } h \not\succ (h^\tau, \hat{c}) \end{cases} \tag{C.9}$$

where $r_\theta^{spot}(z \mid c) = \underset{r \in (0,1)}{\operatorname{argmax}} (U_\theta - c) r$ and $r_g(z) = 0$ for all $z \in Z$. Is easy to see that $\hat{\sigma} \in SBR_{old}(\pi)$ since if $c \neq \hat{c}$ then utility will be $\underline{u}_{old}$, and

$$\underline{u}_{old} < 0 < (1-\beta)\,\mathbb{E}\{\max(0, U_{old} - \hat{c})\} + \beta \mathbb{E}\{\max(0, U_{old})\} = W_{old}^{\pi}(\hat{\sigma}_d \mid h^\tau)$$

and clearly it is the best response for the continuation histories. But then choosing $c = \hat{c}$ is a strong rationalizable strategy for $\theta = old$, a contradiction. □

Proof of Proposition 5.3. We will do it by induction: suppose $k = 0$. Since $h^0 = \emptyset$ is rationalizable for both types, Lemma C.2 implies that if c_0 is new−rationalizable, history $\left(h^0, c_0\right)$ is old−rationalizable as well. For $k > 1$, suppose that history $\left(h^{k-1}, c_{k-1}\right)$ has been both new and old−rationalizable, and we know that $\left(h^k, c_k\right)$ is also new−rationalizable. Because of Lemma C.2 history $\left(h^k, c_k\right)$ can be old−rationalizable as well if and only if

$h^k = \left(h^{k-1}, c_{k-1}, a_{k-1}, z_{k-1}, r_{k-1}\right)$ is also $old-$rationalizable. Since by the induction step we assumed $\left(h^{k-1}, c_{k-1}\right)$ is $old-$rationalizable, we need to rationalize only the choice of r_{k-1} after shock z_{k-1}. But here we can apply directly Proposition 5.2, getting that h^k is $old-$rationalizable if and only if $S_{old,k-1} = \max_{\tilde{r}\in\{0,1\}} \left(U_{old,k-1} - c_{k-1}\right)\tilde{r} - \left(U_{old,k-1} - c_{k-1}\right) r_{k-1} \leq S_{old}^{\max}$. This concludes the proof. □

To prove Proposition 5.4, we will need two lemmas first:

Lemma C.3. *For any strong rationalizable strategy $\sigma_d \in \Sigma_{new}^{SR}$, and any new−rationalizable history, we have*

$$\mathbb{E}_{z_\tau}\left\{r^{\sigma_d}\left(h^\tau, z_\tau\right)\left[U_p\left(z_\tau\right) - c^{\sigma_d}\left(h^\tau\right)\right]\right\} \geq 0 \tag{C.10}$$

Proof. The proof will follow from 2 steps:

Step 1: Let $\underline{\mathbb{W}}_{new} \geq S_{new}^{\max}$. This is equivalent to showing

$$\underline{\mathbb{W}}_{new} \geq \frac{\beta}{1-\beta}\left\{\mathbb{E}_{z_\tau}\left[\max\left(0, U_p\left(z_\tau\right)\right)\right] - \underline{\mathbb{W}}_{new}\right\} \iff$$

$$\underline{\mathbb{W}}_{new} \geq \beta\mathbb{E}_{z_\tau}\left[\max\left(0, U_p\left(z_\tau\right)\right)\right]$$

Suppose $\underline{\mathbb{W}}_{new} < \beta\mathbb{E}_{z_\tau}\left[\max\left(0, U_p\left(z_\tau\right)\right)\right]$. Then the following strategy would be strongly rationalizable: prohibit $r = 1$ at h^τ and in $\tau + 1$ d separates completely. See that since type $\theta = old$ never prohibits r in any rationalizable strategy, then strong certainty of rationality would imply that $\theta = new$ from then on. Therefore, this strategy would then be a robust one, and therefore $\underline{\mathbb{W}}_{new} \geq \beta\mathbb{E}_{z_\tau}\left[\max\left(0, U_p\left(z_\tau\right)\right)\right]$ from the fact that $\underline{\mathbb{W}}_{new}$ is the maximum utility over robust strategies, and thus reaching a contradiction.

Step 2: $\mathbb{E}_{z_\tau}\left[r^{\sigma_d}\left(h^\tau, z_\tau\right)\left(U_p\left(z_\tau\right) - c^{\sigma_d}\left(h^\tau\right)\right)\right] \geq 0$ for all $\sigma_d \in \Sigma_{new}^{SR}$ and all rationalizable histories h^τ.

For any rationalizable strategy σ_d we have

$$(1-\beta)\,\mathbb{E}_{z_\tau}\left[r^{\sigma_d}\left(h^\tau, z_\tau\right)\left(U_p\left(z_\tau\right) - c^{\sigma_d}\left(h^\tau\right)\right)\right] + \beta\mathbb{E}_{z_\tau}\left[\max\left(0, U_p\left(z_\tau\right)\right)\right] \geq W_{new}^{\sigma_d}\left(h^\tau\right) \geq \underline{\mathbb{W}}_{new}$$

This also implies then that

$$(1-\beta)\,\mathbb{E}_{z_\tau}\left[r^{\sigma_d}\left(h^\tau, z_\tau\right)\left(U_p\left(z_\tau\right) - c^{\sigma_d}\left(h^\tau\right)\right)\right] \geq \beta\underline{\mathbb{W}}_{new} - \beta\mathbb{E}_z\left[\max\left(0, U_p\left(z_\tau\right)\right)\right] + (1-\beta)\,\underline{\mathbb{W}}_{new} \iff$$

$$\mathbb{E}_{z_\tau}\left[r^{\sigma_d}\left(h^\tau, z_\tau\right)\left(U_p\left(z_\tau\right) - c^{\sigma_d}\left(h^\tau\right)\right)\right] \geq \underline{\mathbb{W}}_{new} - S_{new}^{\max} \geq 0$$

using Step 1 in the last inequality.

□

Proof of Proposition 5.4. For $S_{old,k-1} > S_{old}^{\max}$ for some $k \leq \tau - 1$, Proposition 5.3 implies that p should have strong certainty that $\theta = new$. Lemma C.3 also implies that,

$$\mathbb{E}_{z_\tau}\left\{r^{\sigma_d}\left(h^\tau, z_\tau\right) U_p\left(z_\tau\right)\right\} \geq \mathbb{E}_z\left\{r^{\sigma_d}\left(h^\tau, z_\tau\right)\left[U_p\left(z_\tau\right) - c^{\sigma_d}\left(h^\tau\right)\right]\right\} \geq 0 > \underline{u}_p$$

Therefore, in any strong rationalizable history where p is strongly certain that $\theta = new$, p strictly prefers to trust. Since the repeated first best is a strong rationalizable continuation strategy (since it maximizes both d and p's utilities), and p will trust regardless of what rationalizable commitment cost is chosen, $\theta = new$ will optimally choose $c_\tau = 0$ and play her first best afterwards, regardless of her beliefs, as long as they are also consistent with common strong certainty of rationality.

When $S_{old,k-1} \leq S_{old}^{\max}$ for all $k \leq \tau - 1$, Lemma C.3 also implies that $\underline{\mathcal{V}}_{new}\left(S_{new,\tau-1}, c_\tau\right) \geq 0 > \underline{u}_p$. Therefore, the implementation restriction

$$\underline{V}\left(h^\tau, c_\tau\right) = \min\left\{\underline{\mathcal{V}}_{old}\left(S_{old,\tau-1}, c_\tau\right), \underline{\mathcal{V}}_{new}\left(S_{new,\tau-1}, c_\tau\right)\right\} \geq \underline{u}_p$$

is satisfied if and only if $\underline{\mathcal{V}}_{old}\left(S_{old,\tau-1}, c_\tau\right) \geq \underline{u}_p$, proving the desired result. □

Lemma C.4. *Under the increasing misalignment assumption 1, given $\epsilon, \delta > 0$, the functions:*

$$G\left(a, b \mid \epsilon, \delta\right) := \int_{a-\epsilon}^{a}\left[\int_{b-\delta}^{b} u_p f\left(u_p, u_o\right) du_p\right] du_o$$

and

$$H\left(a, b \mid \epsilon, \delta\right) = \int_{a-\epsilon}^{a+\epsilon}\left[\int_{b-\delta}^{b+\delta} u_p f\left(u_p, u_o\right) du_p\right] du_o$$

satisfies $\frac{\partial G}{\partial a}.\frac{\partial H}{\partial a} \leq 0$. If $u_p \frac{\partial f}{\partial u_p} \geq 0$ for all z, then we also have $\frac{\partial G}{\partial b} > 0$

Proof. Using Leibniz rule:

$$\frac{\partial G}{\partial a} = \int_{b-\delta}^{b} u_p f\left(u_p, a\right) du_p - \int_{b-\delta}^{b} u_p f\left(u_p, a-\epsilon\right) du_p = \int_{b-\delta}^{b} u_p\left[f\left(u_p, a\right) - f\left(u_p, a-\epsilon\right)\right] du_p$$

which is negative given our assumption. Moreover,

$$\frac{\partial H}{\partial a} = \int_{b-\delta}^{b+\delta} u_p\left[f\left(u_p, a+\epsilon\right) - f\left(u_p, a-\epsilon\right)\right] du_p < 0$$

If $u_p \frac{\partial f}{\partial u_p} \geq 0$ for all z, then

$$\frac{\partial G}{\partial b} = \frac{\partial}{\partial b}\left\{\int_{b-\delta}^{b}\left[\int_{a-\epsilon}^{a} u_p f\left(u_p, u_o\right) du_o\right] du_p\right\} = \int_{a-\epsilon}^{a} b f\left(b, u_o\right) du_o - \int_{a-\epsilon}^{a} (b-\delta) f\left(b-\delta, u_o\right) du_o =$$

$$b\int_{a-\epsilon}^{a}\left[f\left(b, u_o\right) - f\left(b-\delta, u_o\right)\right] du_o + \delta\int_{a-\epsilon}^{a} f\left(b-\delta, u_o\right) du_o > 0$$

as we wanted to show. □

Proposition C.1. *Under assumption 1 , $\underline{\mathcal{V}}_{old}(s,c)$ is decreasing in c_τ*

Proof. To prove the monotonicity of $\underline{\mathcal{V}}_{old}(S_{old,\tau-1},c_\tau)$ with respect to c_τ we use the characterization of the solution to program 5.16 in Proposition A.1. When $S_{old,\tau-1}\leq\hat{s}$

$$\underline{\mathcal{V}}_{old}(S_{old,\tau-1},c_\tau)=\int_{U_o>c_\tau+S_{old,\tau-1}}U_p(z_\tau)\,f(z_\tau)\,dz+=$$

$$+\int_{U_o\in(c_\tau-S_{old,\tau-1},c_\tau+S_{old,\tau-1})}\min\left[0,U_p(z_\tau)\right]f(z)\,dz=G\left(c+\overline{U},\underline{U}\mid\overline{U}-s,\overline{U}\right)+H\left(c,\frac{\underline{U}+\overline{U}}{2}\mid\frac{\overline{U}-\underline{U}}{2},s\right)$$

using the definitions in Lemma C.4, and hence it is decreasing in c, as we wanted to show. □

Lemma C.5. *T as defined in 5.25 is a contraction mapping with modulus β*

Proof. I use Blackwell's conditions to show the result (see Theorem 3.3 in Stokey et al. [1989]). We only need to check monotonicity and discount. See that if $g\leq h$ then $T(g)(s)\leq T(h)(s)$ for all s, since the integrand is an increasing operator. Moreover, $T(g+a)(s)=T(g)(s)+\beta a$ for all s, and hence T is a contraction mapping of module β, as we wanted to show. □

Proof of Proposition 6.1 . Define $P(c)=\mathbb{E}_z[|U_{old}-c|]$. It can be expressed as

$$P(c)=\int_{z\in Z}|U_{old}-c|\,f(z)\,dz=\int_{\underline{U}}^{c}(c-u)\,f_o(u)\,du+\int_{c}^{\overline{U}}(u-c)\,f_o(u)\,du$$

where $f_o(u):=\int_{\underline{U}}^{\overline{U}}f\left(U_p,u\right)dU_p$ denotes the partial of U_{old}. Using Leibniz rule

$$P'(c):=\frac{\partial P(c)}{\partial c}=\int_{\underline{U}}^{c}f_o(u)\,du-\int_{c}^{\overline{U}}f_o(u)\,du=\Pr(U_{old}<c)-\Pr(U_{old}>c)$$

so $\frac{\partial P(c)}{\partial c}>0\iff\Pr(U_{old}<c)\geq\Pr(U_{old}>c)$ or equivalently $\Pr(U_{old}<c)\leq\frac{1}{2}$. Then, is easy to see that if condition 2 holds, then for all $c\geq\overline{c}$ we get $P'(c)>0$ and hence P is increasing in c. Because $c(\cdot)\in\left[\overline{c},c_0^*\right]$ for all $s\in\left[0,S_{old}^{\max}\right]$ and is weakly decreasing in s, the result holds. □

Proof of Lemma 6.1 . I present the proof for the case with $s = 0$, which corresponds to the greatest commitment cost $c_0^* \geq \mathbf{c}(s)$ for all s. For smaller commitment costs the proof will be analogous. It follows from various steps:

Step 1: $\max \left| U_{old} - c_0^* \right| > S_{old}^{\max}$.

If this was not the case, then for all z, $c_0^* - S_{old}^{\max} \leq U_{old} \leq c_0^* + S_{old}^{\max}$. If this was the case, using Proposition A.1 we have that

$$\underline{V}\left(h^\tau, c_0^*\right) = \int_{z \in Z} \min\left(0, U_p\right) f(z)\, dz \leq \int_{z: U_{old} > 0} U_p dF(z) < \underline{u}_P$$

which violates the definition of c_0^*

Step 2: $\min(U_{old}) = \underline{U} < c_0^* - S_{old}^{\max} < c_0^* < \overline{U} = \max\left(\overline{U}\right)$

The right hand side inequality follows from the fact that if $\overline{U} \leq c_0^* - S_{old}^{\max}$ then

$$\underline{V}\left(h^\tau, c_0^*\right) = 0 > \underline{u}_p$$

which will never hold for c_0^* (since $\theta = old$ can drive them to indifference by decreasing the commitment cost enough). From step 1, we either must have that $c_0^* - S_{old}^{\max} > \underline{U}$ or $\overline{U} > c_0^* + S_{old}^{\max}$ (or both). Suppose that the result is not true, so that$\underline{U} \geq c_0^* - S_{old}^{\max}$. Suppose first that $c_0^* - S_{old}^{\max} < \overline{c}$. Then

$$\underline{V}\left(h^\tau, c_0^*\right) = \int_{U_{old} > c_0^* - S_{old}^{\max}} \min\left(0, U_p\right) f(z)\, dz \tag{C.11}$$

$$= \int_{U_{old} \in \left(c_0^* - S_{old}^{\max}, \overline{c}\right)} \min\left(0, U_p\right) f(z)\, dz + \int_{U_p > \overline{c}} \min\left(0, U_p\right) f(z)\, dz \leq$$

$$\int_{U_p > \overline{c}} \min\left(0, U_p\right) f(z)\, dz < \int_{U_p > \overline{c}} U_p f(z)\, dz = \underline{u}_p$$

violating the definition of c_0^*. If $\overline{c} \leq c_0^* - S_{old}^{\max}$ then

$$\underline{V}\left(h^\tau, c_0^*\right) = \int_{U_p > c_0^* - S_{old}^{\max}} \min\left(0, U_p\right) f(z)\, dz < \int_{U_p > c_0^* - S_{old}^{\max}} U_p f(z)\, dz <$$

$$< \int_{U_p > \overline{c}} U_p f(z)\, dz = \underline{u}_P$$

from the definition of $\overline{c}$ (since it's the minimum cost that achieves $\underline{u}_P$ in the spot game). Therefore, we have shown that if $\underline{U} \leq c_0^* - S_{old}^{\max}$ then we have $\underline{V}\left(h^\tau, c_0^*\right) < \underline{u}_p$, violating the definition of c_0^*. Finally, to show $c_0^* > \overline{U}$, suppose that $\overline{U} \leq c_0^*$. Then any strategy consistent with this choice would give the $\theta = old$ an utility of 0, while we know we will make the reservation utility to be binding (i.e. choose the commitment cost a little smaller so that the contrarian behavior is enough to reach the reservation utility).

Step 3 : $\Pr\left(U_p > c_0^* - S_{new}^{\max}, U_{old} < c_0^* - S_{old}^{\max}\right) > 0$

Follows from the fact that $\overline{U} > c_0^* > c_0^* - S_{new}^{\max}$, Step 2 and the full support assumption.

Step 4: $\overline{U} > c_0^* + S_{old}^{\max}$

Suppose that this is not the case: then

$$\underline{V}\left(h^\tau, c_0^*\right) = \int_{U_{old} \in \left(c_0^* - S_{old}^{\max}, c_0^* + S_{old}^{\max}\right)} \min\left(0, U_p\right) f\left(z\right) dz$$

but see that this is identical to expression C.11. Therefore, replicating the same proof as in Step 2, we conclude the result.

Step 5: $\Pr\left(U_p < c_0^* + S_{new}^{\max}, U_{old} > c_0^* + S_{old}^{\max}\right) > 0$

Since $\underline{U} < 0$ we clearly have that $\underline{U} < c_0^* + S_{new}^{\max}$. This, together with the Step 5 and the full support assumption proves the result. □

Proof of Lemma 6.2 . I first show that for any $old-$rationalizable history h^τ we have $\inf_{\beta \in (0,1)} q\left(h^\tau, \beta\right) > 0$. I present the proof for when $c^*\left(h^\tau\right) = c_0^*$. Suppose not: then there exists an increasing sequence $\beta_n \in (0,1)$ such that $q\left(h^\tau, \beta_n\right) > 0 \forall n \in \mathbb{N}$ and $q\left(h^\tau, \beta_n\right) \searrow 0$. For all δ define the expected utility for the people $\underline{v}\left(\beta_n\right) := \underline{V}\left(h^\tau, c_0^*\right) = \underline{u}_p$. For all n we have:

$$\underline{v}\left(\beta_n\right) < \int_{U_{old} \in \left(c_0^*(\beta_n) - S_{old}^{\max}(\beta_n), c_0^*(\beta_n) + S_{old}^{\max}(\beta_n)\right)} \min\left(0, U_p\right) f\left(z\right) dz + q\left(h^\tau, \beta_n\right) \left[\max_{U_P \in \left[\underline{U}, \overline{U}\right]} \left(0, U_p\right)\right]$$

where the first term is the utility in the middle region, and the second term is the natural bound on all regions (particularly in separation regions). Taking limits as $n \to \infty$:

$$\underline{u}_p = \lim_{n \to \infty} \underline{v}\left(\beta_n\right) \leq \mathbb{E}\left[\min\left(0, U_p\right)\right] < \underline{u}_P$$

reaching a contradiction. □

Proof of Lemma 7.1 . The results on $\theta-$rationalizability follows from the same arguments as in the proofs of Propositions 5.2 and 5.3. In the case where $u_{new} = u_p$, we need to show that under condition 7.3 we have $\mathcal{V}_{new}\left(0,0\right) > \underline{u}_p$ and hence $c\left(h^t\right) = \underline{\mathbf{c}}_{old}\left(S_{old,\tau-1}\right)$. This follows from the fact that, in any rationalizable outcome for type θ, she must get at least $\mathbb{E}_z\left\{\max_{r \in R} u_p\left(z\right)\right\} - \overline{c}$, and hence at $c = s = 0$ we must have $\mathcal{V}_{new}\left(0,0\right) \geq \mathbb{E}_z\left\{\max_{r \in R} u_p\left(z, r\right)\right\} - \overline{c} > \underline{u}_p$. Since $\mathcal{V}_{new}$ is decreasing in both c and $s \in [0, S_{new}^{\max}]$, we have that $\mathcal{V}_{new}\left(c_\tau, S_{new,\tau-1}\right) > \underline{u}_p$ for all $\theta-$rationalizable histories, and hence $\underline{\mathbf{c}}_{new}\left(S_{new,\tau-1}\right) = 0$. □

APPENDIX A. CHARACTERIZATION OF $\mathcal{V}(s,c)$

In this section I solve and analyze the solution to the programming problem in subsection (5.20)

Proposition A.1 (**Rationalizable Contrarian Strategy**). *Consider the programming problem 5.20. Then*

(1) *We can rewrite it as*

$$\mathcal{V}(s,c) = \max_{r(.),n(.)} \mathbb{E}_z\left[U_p r(z)\right] \tag{A.1}$$

$$s.t: \begin{cases} \mathbb{E}_z\left[(U_{old}-c)\, r(z) + n(z)\right] \geq \frac{1}{\beta}s + \underline{\mathbb{W}}_{old} & (PK \text{ for sacrifice}) \\ r(z)\left[U_{old} - c + n(z)\right] \geq 0 \text{ for all } z \in Z & (IC \text{ for } r=1) \\ \left[1-r(z)\right]\left[n(z) - U_{old} + c\right] \geq 0 \text{ for all } z \in Z & (IC \text{ for } r=0) \\ n(z) \in \left[0, S_{old}^{\max}\right] \text{ for all } z \in Z & (Feasibility) \end{cases} \tag{A.2}$$

(2) *There exist* $\hat{S} \in \left(0, S_{old}^{\max}\right)$ such that if for $s < \hat{S}$ *then the solution policy* $\underline{r}(z)$ *is*

$$\underline{r}(z) = \begin{cases} 1 & \text{if } U_{old} - c > S_{old}^{\max} \\ 1 & \text{if } U_{old} - c \in \left(-S_{old}^{\max}, S_{old}^{\max}\right) \text{ and } U_{old} < 0 \\ 0 & \text{if } U_{old} - c \in \left(-S_{old}^{\max}, S_{old}^{\max}\right) \text{and } U_{old} > 0 \\ 0 & \text{if } U_{old} - c < -S_{old}^{\max} \end{cases} \tag{A.3}$$

(3) *If* $s \in \left[\hat{S}, S_{old}^{\max}\right]$, *there exist a positive constant* $\alpha(s) \in (0,1)$ *such that*

$$\hat{r}(z) = \begin{cases} 1 & \text{if } U_{old} - c > S_{old}^{\max} \\ 1 & \text{if } U_{old} - c \in \left(-S_{old}^{\max}, S_{old}^{\max}\right) \text{ and } U_p < \gamma(s)\left(U_{old} - c\right) \\ 0 & \text{if } U_{old} - c \in \left(-S_{old}^{\max}, S_{old}^{\max}\right) \text{and } U_p > \gamma(s)\left(U_{old} - c\right) \\ 0 & \text{if } U_{old} - c < S_{old}^{\max} \end{cases} \tag{A.4}$$

(4) *For all* $s \in (0, S^{\max})$ *we have* $\mathbf{c}(s) \in (\overline{c}, S^{\max})$

Proof. Define $n(z) = \frac{\beta}{1-\beta}\left[w(z) - \underline{\mathbb{W}}_{old}\right]$. If $r(z) = 1$ then we can rewrite the enforceability constraint in 5.13 as $(1-\beta)(U_{old} - c) + \beta w(z) \geq \beta \underline{\mathbb{W}}_{old} \iff U_{old} - c + n(z) \geq 0$. Likewise, if $r(z) = 0$ the IC constraint is $\beta w(z) \geq (1-\beta)(U_{old} - c) + \beta \mathbb{W}_{old} \iff n(z) - U_{old} + c \geq 0$. Finally, rewrite (PK) as

$$\mathbb{E}_z\left[(1-\beta)(U_{old} - c)\, r(z) + \beta\left(w(z) - \underline{\mathbb{W}}_{old}\right)\right] \geq \left(\frac{1-\beta}{\beta}\right)s + (1-\beta)\,\underline{\mathbb{W}}_{old} \iff$$

$$\mathbb{E}_z\left[(U_{old}-c)\,r(z)+n(z)\right]\geq\frac{1}{\beta}s+\underline{\mathbb{W}}_{old}$$

See that for any $z:|U_{old}-c|<S_{old}^{\max}$then any $r\in\{0,1\}$ is implementable. However, if $U_{old}>c+S_{old}^{\max}$ then only $r=1$ is implementable, and if $c-U_{old}<-S_{old}^{\max}$ then only $r=0$ is implementable. Then, without the promise keeping constraint (PK) the solution to A.1 is simple:

$$\underline{r}(z):=\begin{cases}1 & \text{if } U_{old}>c+S_{old}^{\max}\\ 1 & \text{if } |U_{old}-c|<S_{old}^{\max},U_p<0\\ 0 & \text{otherwise}\end{cases}$$

i.e. whenever both policies $r\in\{0,1\}$ are rationalizable, $\theta=old$ picks the worst policy for p. We will refer to this policy as the *rationalizable contrarian policy.* It will be also the solution when $s=0$ when the policy $\underline{r}$ satisfies (PK) with strict inequality. Define $\underline{n}(z)$ as the implementing continuation for $\underline{r}(z)$ that maximizes $\mathbb{E}\left\{\left((U_{old}-c)\,\underline{r}(z)+\underline{n}(z)\right)\right\}$. Then, it will be also the solution of A.1 if and only if

$$s\leq\beta\mathbb{E}_z\left[(U_{old}-c)\,\underline{r}(z)+\underline{n}(z)\right]-\underline{\mathbb{W}}_{old}\equiv\hat{s}$$

showing (2). For (3), ignoring for now the IC constraints, use the Lagrangian method (Luenberger [1997])

$$\mathcal{L}=\int U_p r(z)\,f(z)\,dz-\gamma\left\{\int\left[(U_{old}-c)\,r(z)+n(z)\right]-\frac{1}{\beta}s-\underline{\mathbb{W}}_{old}\right\}$$

where $\gamma\geq0$ is the Lagrange multipliers of the problem.

$$\frac{\partial\mathcal{L}}{\partial r(z)}=U_p-\gamma\,(U_{old}-c)$$

then, if $r(z)=1$ is implementable, the optimum will be $r(z)=1\iff U_p\leq\gamma\,(U_{old}-c)$. If we want to implement $r=1$ we then set $n(z)=\min\left\{0,c-U_p\right\}$. Then, given γ we solve for $r(z\mid\gamma)$ and $n(z\mid\gamma)$, and we solve for γ using the promise keeping constraint

$$\int\left[r(z\mid\gamma)\,(U_{old}-c)+n(z\mid\gamma)\right]f(z)=\frac{1}{\beta}s+\underline{\mathbb{W}}_{old}$$

which determines γ as a function of s, showing (3). □

REFERENCES

Dilip Abreu. On the theory of infinitely repeated games with discounting. *Econometrica*, 56(2):pp. 383–396, 1988.

Dilip Abreu, David Pearce, and Ennio Stacchetti. Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica*, 58(5):pp. 1041–1063, 1990.

Fernando Alvarez and Urban J. Jermann. Efficiency, equilibrium, and asset pricing with risk of default. *Econometrica*, 68(4):775–797, 2000.

Susan Athey, Andrew Atkeson, and Patrick J. Kehoe. The optimal degree of discretion in monetary policy. *Econometrica*, 73(5):pp. 1431–1475, 2005.

Robert Aumann and Adam Brandenburger. Epistemic conditions for nash equilibrium. *Econometrica*, 63(5):pp. 1161–1180, 1995.

David Backus and John Driffil. Rational expectations and policy credibility following a change in regime. *Review of Economic Studies*, 52:211–221, 1984.

Gibbons Baker and Murphy. Relational contracts and the theory of the firm. *The Quarterly Journal of Economics*, 117:39–84, 2002.

Robert J. Barro. Reputation in a model of monetary policy with incomplete information. *Journal of Monetary Economics*, 17(1):3 – 20, 1986.

Robert J. Barro and David B. Gordon. Rules, discretion and reputation in a model of monetary policy. *Journal of Monetary Economics*, 12(1):101 – 121, 1983.

Pierpaolo Battigalli and Giacomo Bonanno. Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics*, 53(2):149 – 225, 1999.

Pierpaolo Battigalli and Marciano Siniscalchi. Hierarchies of conditional beliefs and interactive epistemology in dynamic games. *Journal of Economic Theory*, 88(1):188 – 230, 1999.

Pierpaolo Battigalli and Marciano Siniscalchi. Strong belief and forward induction reasoning. *Journal of Economic Theory*, 106(2):356 – 391, 2002.

Pierpaolo Battigalli and Marciano Siniscalchi. Rationalization and incomplete information. *Advances in Theoretical Economics*, 3, 2003.

Elchanan Ben-Porath. Rationality, nash equilibrium and backwards induction in perfect-information games. *The Review of Economic Studies*, 64(1):pp. 23–46, 1997.

Dirk Bergemann and Stephen Morris. Robust mechanism design. *Econometrica*, 73(6): 1771–1813, 2005.

Dirk Bergemann and Stephen Morris. Robust implementation in direct mechanisms. *The Review of Economic Studies*, 76(4):pp. 1175–1204, 2009.

Clive Bull. The existence of self-enforcing implicit contracts. *The Quarterly Journal of Economics*, 102:147–159, 1987.

Matthew B. Canzoneri. Monetary policy games and the role of private information. *The American Economic Review*, 75(5):1056–1070, 1985.

Marco Celentani and Wolfgang Pesendorfer. Reputation in dynamic games. *Journal of Economic Theory*, 70(1):109 – 132, 1996.

V. V. Chari and Patrick J. Kehoe. Sustainable plans. *Journal of Political Economy*, 98(4): pp. 783–802, 1990.

V. V. Chari and Patrick J. Kehoe. Sustainable plans and debt. *Journal of Economic Theory*, 61(2):230–261, 1993a.

V. V. Chari and Patrick J. Kehoe. Sustainable plans and mutual default. *The Review of Economic Studies*, 60(1):175–195, 1993b.

In-Koo Cho and David M. Kreps. Signaling games and stable equilibria. *The Quarterly Journal of Economics*, 102(2):179–221, 1987.

Davide Debortoli and Ricardo Nunes. Fiscal policy under loose commitment. *Journal of Economic Theory*, 145(3):1005 – 1032, 2010.

Drew Fudenberg and David K. Levine. Reputation and equilibrium selection in games with a patient player. *Econometrica*, 57(4):pp. 759–778, 1989.

Stephen Hansen and Michael McMahon. First impressions matter: Signalling as a source of policy dynamics. *Review of Economic Studies*, pages 1–28, 2016.

Robert G. King, Y. K Lu, and Ernesto S. Pasten. Policy design with private sector skepticism in the textbook new keynesian model. *Working P*, 2012.

Narayana R Kocherlakota. Implications of efficient risk sharing without commitment. *Review of Economic Studies*, 63(4):595–609, October 1996.

Elon Kolberg and Jean François Mertens. On the strategic stability of equilibria. *Econometrica*, 54(5):1003–1037, September 1986.

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

David M Kreps, Paul Milgrom, John Roberts, and Robert Wilson. Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, 27(2):245 – 252, 1982.

Finn E. Kydland and Edward C. Prescott. Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy*, 85(3):pp. 473–492, 1977.

Levin. Relational incentives contracts. *The American Economic Review*, 93:835–857, 2003.

Ethan Ligon. Risk sharing and information in village economics. *Review of Economic Studies*, 65(4):847–64, 1998.

Ethan Ligon, Jonathan P. Thomas, and Tim Worrall. Mutual insurance, individual savings, and limited commitment. *Review of Economic Dynamics*, 3(2):216 – 246, 2000.

Susanne Lohmann. Optimal commitment in monetary policy: Credibility versus flexibility. *The American Economic Review*, 82(1):pp. 273–286, 1992.

Yang K. Lu. Optimal policy with credibility concerns. *HK UST Working Paper*, 2012.

D.G. Luenberger. *Optimization by vector space methods*. Wiley-Interscience, 1997.

George Mailath, Masahiro Okuno-Fujiwara, and Andrew Postlewaite. Belief-based refinements in signalling games. *Journal of Economic Theory*, 60(2):241–276, 1993.

George J. Mailath and Larry Samuelson. *Repeated Games and Reputations: Long Run Relationships*. Oxford, 2006.

Paul Milgrom and John Roberts. Predation, reputation, and entry deterrence. *Journal of Economic Theory*, 27(2):280 – 312, 1982.

Maurice Obstfeld and Kenneth Rogoff. *Foundations of International Macroeconomics*. MIT Press, 1996.

Juan Passadore and Juan Pablo Xandri. Robust conditional predictions in dyanmic games: An application to robust conditional predictions in dynamic games: An aplication to sovereign debt. *unpublished*, 2016.

Alessandro Pavan and George-Marios Angeletos. Selection-free predictions in global games with endogenous information and multiple equilibria. *Theoretical Economics, Forthcoming*, 2012. URL `http://ideas.repec.org/a/the/publsh/1156.html`.

David G. Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52(4):pp. 1029–1050, 1984.

Antonio Penta. Robust dynamic mechanism design. *Working Paper*, 2011.

Antonio Penta. Higher order uncertainty and information: Static and dynamic games. *Econometrica*, 80(2):631–660, March 2012.

Torsten Persson and Guido Tabellini. *Macroeconomic Policy, Credibility and Politics*. Harwood Academic Publishers GmbH, 1990.

Torsten Persson and Guido Tabellini. Designing institutions for monetary stability. *Carnegie-Rochester Conference Series on Public Policy*, 39(0):53 – 84, 1993.

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

Kenneth Rogoff. The optimal degree of commitment to an intermediate monetary target. *The Quarterly Journal of Economics*, 100(4):pp. 1169–1189, 1985.

Thomas J. Sargent and Lars Ljungqvist. *Recursive Macroeconomic Theory*. MIT Press, 2004.

Anne Sibert. Monetary policy committees: Individual and collective reputations. *The Review of Economic Studies*, 70(3):649–665, 2003.

Nancy L. Stokey. Reputation and time consistency. *The American Economic Review*, 79 (2):pp. 134–139, 1989.

Nancy L. Stokey. Credible public policy. *Journal of Economic Dynamics and Control*, 15 (4):627 – 656, 1991.

Nancy L. Stokey, Robert E. Lucas, and Edward C. Prescott. *Recursive methods in economics dynamics*. Presidents and Fellows of Harvard College, 1989.

Lars EO Svensson and Michael Woodford. *Implementing optimal policy through inflation-forecast targeting*, chapter 2, pages 19–92. University of Chicago Press, 2004.

Jonathan Thomas and Tim Worrall. Self-enforcing wage contracts. *Review of Economic Studies (1988)*, 55(4):541–554, 1988.

John Vickers. Signalling in a model of monetary policy with incomplete information. *Oxford Economic Papers*, 38(3):443–455, November 1986.

Carl Walsh. Optimal contracts for central bankers. *The American Economic Review*, 85: 150–167, 1995.

Jonathan Weinstein and Muhamet Yildiz. A structure theorem for rationalizability with application to robust predictions of refinements. *Econometrica*, 75(2):365–400, 2007.

Jonathan Weinstein and Muhamet Yildiz. Robust predictions in infinite-horizon games - an unrefinable folk theorem. *Review of Economic Studies, forthcoming*, 2012.

Alexander Wolitzky. Reputational bargaining with minimal knowledge of rationality. *Econometrica*, 80(5):2047–2087, 2012.