# CREDIBLE REFORMS: A ROBUST IMPLEMENTATION APPROACH

JUAN PABLO XANDRI*

Abstract. I study the problem of a government with low credibility, who 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 the 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 the policy maker eventually convinces the public 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. I finally 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 situation. 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 on this literature to model reputation in a way that reflects the uncertainty faced by policy makers.

---

*Date*: April 2013.

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 Gabriel Carroll, Arun Chandrasekhar, Sebastian Di Tella, Juan Dubra, Glenn Ellison, Alvaro Forteza, Ben Golub, Horacio Larreguy, Anton Kolotilin, Pablo Kurlat, Plamen Nenov, Juan Passadore, Juuso Toikka, Xiao Yu Wang, Alex Wolitzky, Luis Zermeño and Jan Zilinsky .

†Massachusetts Institute of Technology. Email: jxandri@mit.edu.

Following the seminal papers of **?** and **?**, 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. In the canonical example of monetary policy, a policy maker wants the public to trust her announcements of inflation policy. The fundamental problem resides in the fact that even if the decision 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, since 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. First, some authors have argued that the government should *forfeit some flexibility* through a formal arrangement (e.g., inflation targeting/caps and tax restrictions). Second, others have argued that the government should *modify its incentives*, either by having reputational concerns through repeated interactions (**??** and **???**) or by delegating the decision to an agency with different incentives that will limit its time inconsistency bias. Examples of such delegation include appointing a conservative central banker (**?**) or making the monetary authorities subject to a formal or informal incentive contract (**?**, **?**, **?**). 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 approaches is that policies that are 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 enough 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 **??**, but also allowing the public to have uncertainty about the governments expectations for the continuation game. The main goal of our paper will be to investigate if, through repeated interactions, the government can convince the public about its reformed incentives.

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. 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. 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 (**??**).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, a smaller set makes it trivial; the feasible set I consider is discussed below. I study the case where the only constraints on the sets of beliefs are that they are consistent with 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 not to question the government's rationality, unless proven otherwise, which is given by the solution concept of *strong* or *extensive-form rationalizability* of **??**.

I show that this policy exhibits endogenous transitory gains and losses of reputation. Moreover, the policy achieves permanent separation (i.e. public is convinced about the success of the reform from then on) almost surely and 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 *max-min* 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 that he faces a time inconsistent policy maker and has observed her taking an action that did not maximized her spot utility. To fix ideas, suppose she took an action such that, if she was the time inconsistent type, gave her 10 utils. Meanwhile, she could have reacted in a manner that gave her 25 in spot utility instead. The implied opportunity cost paid by her 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 the government beliefs and planned course of action from tomorrow onward must deliver (from the government's point of view) more than 15 utils, which is a constraint that rationality imposes on the goverment's expected future behavior. If however, 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 policy maker being time consistent (e.g. she had an opportunity cost of 2) then the public should be fully convinced from then on that they are facing the reformed government. Therefore, the implied spot opportunity cost paid by the time inconsistent type 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 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 promise keeping constraint*, analogous to **?**, **?** and **?** 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. I also study how further restrictions on the set of feasible beliefs can help accelerate the reputation formation process and in particular find a characterization of restrictions that generate monotone robust policies (i.e. policies that exhibit only permanent gains of reputation on its path). In Section 7 I study some extensions to our model and discuss avenues for future research. Finally, Section 8 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 questions in the 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 **?** and **?**, where the time inconsistent type is a benevolent goverment, instead of just opportunistic. Consider an economy with two type of households: workers $(w)$ and capitalists $(k)$. There is a continuum of measure one of identical households, for each type. 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 households utilities are given by

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

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

where $\tau^e = \mathbb{E}\left[\tau(z)\right]$. 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:

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

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

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

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:

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

where $\alpha \ge 0$ 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 $I > \Pr(z_k + \alpha z_w \le 1)$ , 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

$$\max_{\tilde{r},\tilde{\tau}} \tilde{r}\,(z_k + \alpha z_w) - 1\{\tilde{\tau} > 0\}\, c \text{ s.t. } \tilde{r} \le \tilde{\tau} q.$$

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

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 can now induce households to invest, by credibly distorting its tax policy.

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

$$\max_{\tilde{r},\tilde{\tau}} \tilde{r}\left(z_k + \alpha_{new} z_w\right) + \left(1 - \tilde{\tau}\right) q \text{ s.t. } \tilde{r} \leq \tilde{\tau} q.$$

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 be socially desirable for the benevolent government (without taken into account the commitment cost payed) 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 **?**, the analog to a pro-capitalist type is a commitment type (as in **?**) that always pick the same tax rate. In **?**, the government can make announcements, and can be either a committed type (i.e. one that is 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

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

Condition 2.6 implies that for sufficiently low $\pi$, 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 $\pi$), 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 in **?**[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

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

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

[3] Section 9.5, pp 634-657.

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 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*, in which 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 $\pi^e$ and $z$, she chooses $\pi$ 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.

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 the monetary

[4]See **?** and **?** for a discussion of such potential inefficiencies.

authority can choose from. **?** 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 **?**, 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

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

**?** suggests placing a "conservative central banker", that has $k^{new} = k$ but $\chi^{new} > \chi$, so that it places a greater importance on inflation stabilization than society does. From 2.12 we see that increasing the weight $\chi$ 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:

$$\mathcal{L}_{new} = \mathcal{L} + \alpha\left[\eta\pi\right]$$

**?** and **?** argue that this can be done by offering a contract 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 **?**). 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.

[5]In their paper, **?** 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:

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

[6]Suppose the monetary authority minimizes $\hat{\mathcal{L}}_{old} = \mathcal{L}_{old} - \alpha u\left[\phi(\pi)\right]$ 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.

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. The literature has studied the case of

## 3. Literature Review

The literature on time inconsistency of government policy is extensive, beginning with the seminal papers by **?** and **?**, where the idea of the commitment solution (i.e. choosing policy first) was first introduced. The reputation channel was first explored by **?** and **?**, 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 type that plays a constant strategy at all histories. They show (following the arguments in **???**) that for 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, **?** show that a long lived agent facing a sequence 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. **?** 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 **?** results to get bounds on equilibrium payoffs. **?** 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 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 are policies that are the outcome of a subgame perfect equilibrium of the policy game, starting with **??** and **???**.

In such environments, governments may have incentives to behave well under the threat of punishment by switching to a bad equilibrium afterwards. See ?[7] for a tractable unified framework to study these issues.[8]

This paper studies reputation formation on both dimensions: in terms of payoff heterogeneity and in terms of equilibrium punishments. The main point of departure is that instead of designing the optimal policy for a time inconsistent policy maker (that wants to behave as if she was time 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. Therefore, its goal is essentially to separate itself from the time inconsistent, untrustworthy type, if possible. The most related papers in spirit to mine are ?,? and particularly ?. ? 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. ? 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. ? apply these ideas in the context of the standard New Keynesian model, similar to our setup in subsection 2.2. Ultimately these papers study equilibrium in an environment where all players involved know that the government is ex-ante either a type that can commit or not (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., ?) select a different equilibrium. In macroeconomics, the most related paper to mine that studies robustness to specific refinements is ?. 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 ?, and robust implementation in ?. 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, like normal form Interim Correlated Rationalizability (as in ?) and Interim Sequential Rationalizability (??), among others. This paper focuses on the stronger assumption of common

[7] Chapter 16, pp 485-526

[8] This principle is also exploited in the relational contract literature (?,? ,?) 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 (?, ?, ? and ?) 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.

strong certainty of rationality (**???**) which is also equivalent to **?** notion of "Extensive-form rationalizability". In a similar vein, the paper most related in spirit to mine is **?**. He studies reputational bargaining in a continuous time setting in which agents announce bargaining postures that they may become committed 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 they think they are facing a rational opponent or not. 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. Section 4.1 describes the stage game and shows the multiplicity of equilibria. I then setup the repeated game in Section 4.2 and develop the concept of system of beliefs in Section 4.3. Section 4.4 introduces weak and strong rationalizability and Section 4.5 argues 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). The agent represents the public. In the no-commitment benchmark, $p$ is asked to trust a state-contingent decision to $d$, who after a state of nature $z \in Z$ is realized, has to choose a policy that affects both parties payoffs. For simplicity, I will assume that $d$ has only two options: a "normal" policy that is optimal most of the time and an "emergency" policy that needs to be taken in certain instances. As a mnemonic device, we will write $g$ for the normal policy (pushing a "*g*reen button") and $r$ for the emergency policy (pushing a "*r*ed button"). The extensive form game is described in Figure 1.

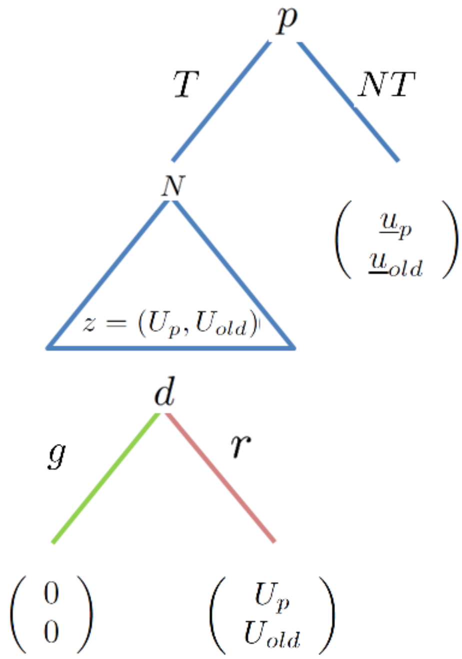


FIGURE 1. Stage game with time inconsistent type

The random shock $z = (U_p, U_{old})$ is the profile of relative utilities of the emergency action $r$ with respect to $g$, 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 $f(z)$. The subscript "old" serves to remind the reader that these are the preferences of the policy maker before a reform is undertaken, which will be described below.

To make this concrete, consider the capital taxation example described in 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, $T$ corresponds to "invest $\overline{q}$". The shock $z$ can be written as

$$z = (U_p, U_{old}) := \overline{q}\,(z_k - 1, z_b - 1) \tag{4.1}$$

Returning to the general model, I will additionally assume that

$$\mathbb{E}_z\left[\max\left(0, U_p\right)\right] > \underline{u}_p > \int_{U_{old}>0} U_p f(z)\, dz \tag{4.2}$$

This implies that if $p$ was guaranteed his most desirable policy, he would place trust in $d$.[9] I will further assume that

$$\underline{u}_p, \underline{u}_{old} < 0 \tag{4.3}$$

so that both parties would benefit from ensuring that the decision maker plays the normal action $g$ for all shocks (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.

As I discussed in the introduction, I will explore two (potentially complementary) ways to solve this *time inconsistency problem:* by credibly loosing *some* flexibility to react to the economic shock, or by reforming the incentives of the decision maker, in order to alleviate the time inconsistency problem.

First, I introduce a commitment cost technology, under which the decision maker can choose, before the game starts, a cost $c \geq 0$ of taking the emergency action $r$. This can be interpreted as a *partial commitment* to the normal policy $g$, that includes an escape clause to break the commitment, forcing the decision maker to suffer a cost of $c \geq 0$ utils (as in **?**). 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. The modified stage game is illustrated in Figure 2.

[9] In the context of the capital taxation problem, by setting $\underline{u}_p = -(1-I)\,\mathbb{E}\left(z_k \mid z_b \geq 1\right)(\overline{q}-1)$, the left hand side inequality of 4.2 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).

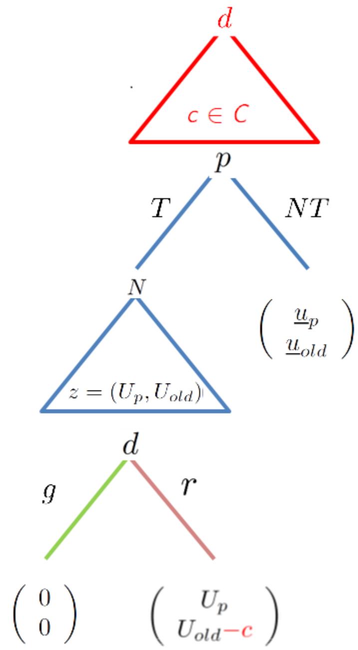


FIGURE 2. Stage game with commitment cost choice

I will allow $c = \infty$, the case where $d$ decides to shut down the emergency action $r$, as I previously discussed, so that the commitment cost set is $C = [0, \max_{z \in Z} \max(|z_p|, |z_{old}|)] \cup \{\infty\}$. In this game, the decision maker would then choose the commitment cost $c = \bar{c}$ that makes $p$ indifferent between trusting and not trusting:

$$\bar{c} := \min\left\{c \geq 0 : \int_{U_{old} > c} U_p f(z)\, dz \geq \underline{u}_p\right\} < \infty \tag{4.4}$$

While this technology is available, I will consider the case in which the policy maker makes a reform, by effectively changing the ex-post incentives that the decision maker faces, by either delegating the decision to a different agent (like the conservative central banker of **?**) or by designing a contract for the decision maker (as in **?** and **?**). I model this reform by creating a new policy maker type, $\theta = new$, with ex-post payoffs given by

$$U_{new}(z) = U_p \tag{4.5}$$

i.e. the reformed decision maker has the same ex-post incentives as the public. Condition 4.2 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.

The key problem I will study in this paper is the government's lack of credibility. 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 $\theta = old$. This results on an incomplete information game, described in Figure 3.

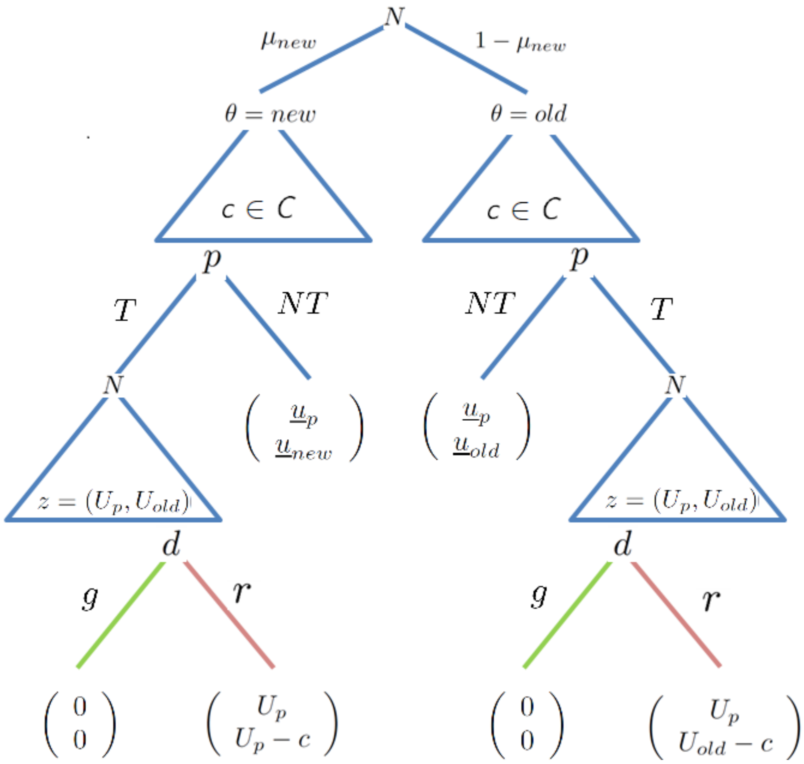


FIGURE 3. Incomplete information game

This model follows an *executive approach* to optimal policy, where it is the decision maker herself who decides the commitment cost and the policy rule. This contrasts with the *legislative approach* studied in **?** and **?** who instead solve for the optimal mechanism design problem from the point of view of $p$. In Section (7) I briefly explore this route and find that in our setting, it will be detrimental to welfare, conditional on the government being of type $\theta = new$.

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.

Let $\underline{c}(\pi)$ be the minimum cost that induces $p$ to trust $d$ in any pooling equilibrium:

$$\underline{c}(\pi) := \min\left\{c \geq 0 : \pi \int_{U_p > c} U_p f(z)\, dz + (1-\pi) \int_{U_{old} > c} U_p f(z)\, dz \geq \underline{u}_p\right\}. \tag{4.6}$$

It is easy to see that $\underline{c}(\pi)$ is decreasing in $\pi$ and that $\underline{c}(\pi) < \underline{c}(0) = \overline{c}$ according to 4.4. 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 $\hat{c} \in [\underline{c}(\pi), \overline{c}]$ there exists a PBE in which both types choose $\hat{c}$ as the commitment cost.*

**Proof.** 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

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

$$\pi_p(\overline{c})\, \mathbb{E}[\max(U_p - \overline{c}, 0)] + [1 - \pi_p(\overline{c})]\, \underline{u}_p \underset{(2)}{>} \pi_p(\overline{c})\, \underline{u}_p + [1 - \pi_p(\overline{c})]\, \underline{u}_p = \underline{u}_p \tag{4.7}$$

where (1) follows from definition 4.4 and (2) from the fact that $0 > \underline{u}_p$. I will now show that for any $\hat{c} \in [\underline{c}(\pi), \overline{c}]$ 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:

$$\pi_p^{\hat{c}}(c) := \begin{cases} 0 & \text{if } c < \hat{c} \\ \pi & \text{if } c \geq \hat{c}. \end{cases} \tag{4.8}$$

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. $\square$

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. 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 prescription was the commitment cost supported by the *best equilibrium* of the game, Proposition 4.1 shows that $c = \underline{c}(\pi)$ is selected for either $\theta = new$ or $\theta = old$. We will argue that this will not be a particularly "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 $\pi$ were commonly known, after the commitment cost has been chosen, $p$ updates beliefs to $\pi_p(c) \in (0, 1)$. As illustrated by the proof of Propostion 4.1 (in particular the belief updating rule in 4.8), 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 (for example, by changing $\hat{c}$ in 4.8) generates potentially very different behavior for $p$.[11]

Our main question becomes whether we can choose a policy that is robust to mis-specifications of both the prior $\pi$ and the updating rule $\pi_p(c)$. It is clear that by choosing $c = \infty$ and effectively removing all flexibility, a rational $p$ would trust $d$ independently of his beliefs. However, we can do better. Inequality 4.7 implies that if $c = \overline{c}$ and $p$ will find it optimal to trust irrespective of the updating rule $\pi_p(c)$ if he still believes he is facing a rational decision maker. I will show 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: types cannot separate in any equilibrium of the stage game, by their choice of $c$. 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 $\pi_{new}$. $d$ 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 iid across periods: $z_\tau := (U_{p,\tau}, U_{old,\tau}) \sim_{i.i.d} f(z_\tau)$. I assume that there is a sequence of myopic short run players $p_\tau$ (or equivalently $\beta_p = 0$) which is a standard assumption in the reputation literature (**?**, **?**). 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 policy 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

[11]More generally, we apply **?** 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 $\pi_p(.)$ is also common knowledge

[12]**?** 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 $\delta_k \in (0,1)$.

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 $\tau$ *outcome* is a 4-tuple $h_\tau = (c_\tau, a_\tau, z_\tau, r_\tau)$ where $c_\tau$ 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 $\tau$ 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 $(h^\tau, c_\tau)$, 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 $\tau$, a commitment cost $c_\tau$ and the contingent choice provided $p$ trusts. Then, we can always write a strategy as a pair:

$$\sigma_d(h^\tau) = \left(c^{\sigma_d}(h^\tau), r^{\sigma_d}(h^\tau, \cdot) : C \times Z \to \{0,1\}\right) \tag{4.9}$$

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 $c_\tau$ The superscript $\sigma_d$ serves to remind the reader these objects are part of a single strategy $\sigma_d$. Likewise, a strategy $\sigma_p$ for $p$ is a function $\sigma_p : \mathcal{H}_p \to \{0,1\}$ that assigns to every observed history, his trust decision

$$\sigma_p(h^\tau, c_\tau) = a^{\sigma_p}(h^\tau, c_\tau) = \begin{cases} 1 & \text{if } p \text{ trusts} \\ 0 & \text{if } p \text{ does not trust.} \end{cases}$$

Write the set of strategies of each agent as $\Sigma_i$ for $i \in \{d, p\}$ Also let $\Sigma = \Sigma_d \times \Sigma_p$ be the set of strategy profiles $\sigma = (\sigma_d, \sigma_p)$. If player $i \in \{d, p\}$ plays strategy $\sigma_i$, the set of histories that will be consistent with $\sigma_i$ is denoted $\mathcal{H}(\sigma_i) \subset \mathcal{H}$. For a history $h \in \mathcal{H}$ we say a strategy $\sigma_i$ *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 $\sigma = (\sigma_p, \sigma_d)$ let $W_\theta(\sigma \mid h)$ be the expected continuation utility for $d's$ type $\theta \in \{old, new\}$ given history $h$

$$W_\theta(\sigma \mid h) := (1-\beta)\,\mathbb{E}\left\{\sum_{s=\tau}^{\infty} \beta^{s-\tau}\left[a_s r_s (U_{\theta,s} - c_s) + (1-a_s)\underline{u}_p\right] \mid h\right\} \tag{4.10}$$

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, denote $V(\sigma \mid h)$ for the spot utility for agent $p$ at history $h$

$$V(\sigma \mid h) := a_\tau \mathbb{E}_z(r_\tau U_{p,\tau}) + (1-a_\tau)\underline{u}_p. \tag{4.11}$$

4.3. **Systems of Beliefs.** Agents form beliefs both about the payoff types of the other player, as well as the strategies that they may be planning to play. In static games, such

[13]These are the basic assumptions in **?**.

beliefs are characterized by some distribution $\pi \in \Delta(\Theta_{-i} \times S_{-i})$ where $\Theta_{-i}$ 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*, that 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 infomation sets for $i$. A *system of beliefs* $\pi_i$ 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 $\pi_d \in \Delta^{\mathcal{H}}(\Theta_p \times \Sigma_p) = \Delta^{\mathcal{H}}(\Sigma_p)$ 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 $\pi_d$:

$$W_\theta^{\pi_d}(\sigma_d \mid h) := \int W_\theta(\sigma_d, \hat{\sigma}_p \mid h)\, d\pi_d(\hat{\sigma}_p \mid h) \tag{4.12}$$

Analogously, given a system of beliefs $\pi_p$ the expected utility of strategy $\sigma_p$ conditional on history $h$ is

$$V^{\pi_p}(\sigma_p \mid h) := \int V(\hat{\sigma}_d, \sigma_p \mid h)\, d\pi_p(\hat{\sigma}_d \mid h) \tag{4.13}$$

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

4.4. **Weak and Strong Rationalizability.** In this subsection I introduce the notions of weak and strong rationalizability. Our 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*, 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 **?**. 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.

[14]A sequence $\{\sigma_{i,n}\}_{n\in\mathbb{N}}$ converges to $\sigma_i$ in the product topology in $\Sigma_i$ 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 $\sigma_p$ is a *sequential best response to* $\pi_p$ for all $(h^\tau, c_\tau) \in \mathcal{H}_p$ and all other strategies $\hat{\sigma}_p \in \Sigma_p$, we have $V^{\pi_p}(\sigma_p \mid h^\tau, c_\tau) \geq V^{\pi_d}(\hat{\sigma}_p \mid h^\tau, c_\tau)$. Likewise, $\sigma_d$ is a *sequential best response to belief system* $\pi_d$ *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 $W_\theta^{\pi_d}(\sigma_d \mid h) \geq W_\theta^{\pi_d}(\hat{\sigma}_d \mid h)$.

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 $\pi_i(E \mid h^0) = 1$

**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 $\pi_p$, 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, $\pi_p$ 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 $\pi_p$ 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 $\pi_p$ 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" (**?**)

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

$$R_i = \left\{(\theta_i, \sigma_i) : \ \sigma_i \in SBR_{\theta_i}(\pi_i) \text{ for some } \pi_i \in \Delta^{\mathcal{H}}(\Theta_{-i} \times \Sigma_{-i})\right\}. \tag{4.14}$$

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

I will now formally follow the iterative procedure of **?**. 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 $\pi_i$ 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

[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 (**?**, **?**, **??**) instead of knowledge, for dynamic games.

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:

$$WCR_i^k := \left\{ (\theta_i, \sigma_i) : \begin{cases} (1): & (\theta_i, \sigma_i) \in WCR_i^{k-1} \\ (2): & \exists \pi_i \in \mathbf{W}_i \left( WCR_{-i}^{k-1} \right) : \sigma_i \in SBR_{\theta_i}(\pi_i) \end{cases} \right\} \tag{4.15}$$

$$SCR_i^k := \left\{ (\theta_i, \sigma_i) : \begin{cases} (1): & (\theta_i, \sigma_i) \in SCR_i^{k-1} \\ (2): & \exists \pi_i \in \mathbf{S}_i \left( SCR_{-i}^{k-1} \right) : \sigma_i \in SBR_{\theta_i}(\pi_i) \end{cases} \right\}. \tag{4.16}$$

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

$$WCR_i^\infty = \bigcap_{k \in \mathbb{N}} WCR_i^k \tag{4.17}$$

$$SCR_i^\infty = \bigcap_{k \in \mathbb{N}} SCR_i^k \tag{4.18}$$

The sets $WCR_i^\infty$,$SCR_i^\infty \subset \Sigma_i$ 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

$$\mathcal{B}_i^{WR} := \Delta^{\mathcal{H}} \left( WCR_{-i}^\infty \right) \text{ and } \mathcal{B}_i^{SR} := \Delta^{\mathcal{H}} \left( SCR_{-i}^\infty \right). \tag{4.19}$$

I will say that a strategy-belief pair $(\sigma_i, \pi_i)$ is $\theta_i-$*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 $\theta_i-$strong rationalizable whenever $h \in \mathcal{H}(\sigma_i)$ for some weak rationalizable pair $(\sigma_i, \pi_i)$. I will refer to such pairs as a $\theta_i-$*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. We briefly connect the concepts now, and argue that equilibrium refinements are not robust to a variety of perturbations we might think of.

First, and most importantly for our applications is the one considered in this paper which is robustness to *strategic uncertainty*. Recall that the static game had multiple equilibria.

The dynamic game only exacerbates this problem via classical folk theorem like arguments.[18] This suggests that in order for either to form predictions or to make policy recommendations, some equilibrium refinement is needed, as selecting optimal or efficient equilibria. 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. 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, [19]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., **???**). 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. [20]However, the richness assumption may be too a stringent condition for our 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 7 how to create policies that are robust to 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.[21] When this happens, we will say that $\theta = new$ has achieved *full or strong separation* from $\theta = old$. This is one of

[18]See **?** for an exhaustive review on these topics.

[19]Formally, for every strategy $\sigma_i$ there exist a type $\hat{\theta}_i(\sigma_i) \in \hat{\Theta}_i$ such that $\sigma_i$ 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 $\hat{\sigma}_i \in \Sigma_i, \sigma_{-i} \in \Sigma_{-i}, h \in \mathcal{H}_i(\sigma_i)$.

[20]**?** 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 (**?**).

[21]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 **?**. 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.

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 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 in some subset $\mathcal{B}_p \subset \Delta^{\mathcal{H}}(\Theta_d \times \Sigma_d)$. Write $SBR_p(\mathcal{B}_p) = \bigcup_{\pi_p \in \mathcal{B}_p} SBR_p(\pi_p) \subset \Sigma_p$ as the set of all sequential best responses to beliefs in $\mathcal{B}_p$. We will say that a strategy $\sigma_d$ 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

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

Under the assumptions on the stage game, for a given belief system $\pi_p$ , its sequential best response will be generically unique. Therefore, if $d$ knows both that $p$ is rational and that he has beliefs $\pi_p$, 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 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.[22]

**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 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 *min-max* 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 (see the construction of belief system in 4.8), 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.

[22]This argument extends to any game of *private values* with multiple weak rationalizable outcomes. A Bayesian game is of if utility for each agent depends only on their own payoff parameter, and not about the other agents payoffs. Formally, is of private values if for all mathnorma$\theta \in \Theta \equiv \times_{i=1}^I \Theta_i$

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 payed, the only relevant object to determine his expected payoff is the way he expects $d$ to react to shocks at time $\tau$. Define then the set of all strong rationalizable policy functions

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

Define also $\mathbf{R}_{\theta}\left(h^{\tau}, c_{\tau}\right) \subset \mathbf{R}\left(h^{\tau}, c_{\tau}\right)$ as those policy functions that are $\theta-$rationalizable. Is easy to show that $a^{\sigma_p}\left(h^{\tau}, c_{\tau}\right)$ for all strong rationalizable strategies if and only if [23]

$$\int r\left(z_{\tau}\right) U_p\left(z_{\tau}\right) f\left(z_{\tau}\right) dz_{\tau} \geq \underline{u}_p \text{ for all } r\left(\cdot\right) \in \mathbf{R}\left(h^{\tau}, c_{\tau}\right)$$

which can be rewriten in a single condition as:

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

i.e. the worst rationalizable payoff for $p$ must be higher than the reservation utility. In Appendix B.2 we show that $\mathbf{R}\left(h^{\tau}, c_{\tau}\right)$ and $\mathbf{R}_{\theta}\left(h^{\tau}, c_{\tau}\right)$ are compact sets and the objective function in the minimization problem of 5.2 is continuous in the product topology, which makes $\underline{V}\left(h^{\tau}, c_{\tau}\right)$ a well defined object.

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

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

subject to

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

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 our programming problem. Theorem B.3 shows that $\Sigma_{\theta}^{SR}$ is a compact set, and so are the subsets $\Sigma_{\theta}^{SR}\left(h\right) \subset \Sigma_{\theta}^{SR}$ of history consistent strategies, for all $\theta$. 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 $\theta \in \{new, old\}$

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

I will refer to $\underline{W}_{\theta}\left(\cdot\right)$ and $\overline{W}_{\theta}\left(\cdot\right)$ as the *best* and *worst strong rationalizable payoffs* for type $\theta$. 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

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

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 $\sigma_{\theta}^{*}$ be the optimal robust strategy for type $\theta$. Then $\sigma_{\theta}^{*} \in \Sigma_{\theta}^{SR}$, with rationalizing belief $\underline{\pi}_{\theta} \in \mathcal{B}_{\theta}^{SR}$. Moreover, for all histories $h \in \mathcal{H}_d$*

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

*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 min-max strategies for type $\theta \in \{old, new\}$ (as in **?**) 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 $\theta$, the most pessimistic beliefs that type $\theta$ can have about the strategy that $p$ may be playing. Therefore, at history $h^{\tau}$, the optimal robust policy gives the *best payoff that type $\theta$ 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.[24].

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 $\mathbf{R}_{\theta}\left(h^{\tau}, c_{\tau}\right)$. Since $\mathbf{R}\left(h^{\tau}, c_{\tau}\right) = \mathbf{R}_{new}\left(h^{\tau}, c_{\tau}\right) \cup \mathbf{R}_{old}\left(h^{\tau}, c_{\tau}\right)$, characterizing these sets will determine the shape of $\underline{V}\left(h^{\tau}, c_{\tau}\right)$. 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 payed by type $\theta$, to be able to characterize the set $\mathbf{R}_{\theta}\left(h^{\tau}, c_{\tau}\right)$. In this sense, the set of strong rationalizable policies $\mathbf{R}_{\theta}\left(h^{\tau}, c_{\tau}\right)$ is Markovian, with a state variable that is observable by all agents in the game.

Consider a history $\left(h^{\tau}, c_{\tau}\right) \in \mathcal{H}_p$ observed by agent $p$. Suppose first that $p$ hypothesizes that $d$ is of payoff type $\theta$, and that history $h^{\tau}$ is such that $r_{\tau-1} = 0$ and $U_{\theta,\tau-1} - c_{\tau-1} > 0$,

[24]Applying **?** 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 $\pi = \Pr\left(\theta = new\right)$) 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.

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 $\theta$. Let $h^{\tau}(r=1)$ be the continuation history had $d$ chosen $r_{\tau-1}=1$ instead. Then, a $\theta-$rationalizable pair $(\sigma_d, \pi_d)$ is consistent with the observed $h^{\tau}$ if and only if

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

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 $\theta$ of playing $r_{\tau-1}=0$ instead of $r_{\tau-1}=1$. Also, let

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

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

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

i.e. it would have been optimal for type $\theta$ 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

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

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

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

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

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

5.11 puts restrictions on $\theta-$rationalizing pairs $(\sigma_d, \pi_d)$ (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*

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

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

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

Condition 5.13 is analogous to the **?** 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 in **?**. 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 $\sigma_d$ 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 $(\sigma_d, \pi_d)$. Its proof resembles closely the well known "optimal penal codes" argument in **?**: 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 $\theta$ thinks that if she deviated, she will have to play the optimal robust policy from then on.

Proposition 5.2 requires the history $(h^\tau, c_\tau)$ to be $\theta-$rationalizable. In order to be able to use this characterization, we need to determine whether $(h^\tau, c_\tau)$ 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 $\theta$, 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 we infer 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 occured up to period $\tau-1$, $\theta = new$ will robustly separate from $\theta = old$

at period $\tau$ 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 **?**. Suppose that at a given a $\theta-$rationalizable history $(h^\tau, c_\tau)$, $p$ 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:

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

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 "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

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

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 propostition 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 $f(U_p, U_{old})$ 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*

$$\underline{V}\left(h^{\tau}, c_{\tau}\right) \geq \underline{u}_{p} \iff V_{old}\left(S_{old,\tau-1}, c_{\tau}\right) \geq \underline{u}_{p} \tag{5.20}$$

(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.*

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

*and hence, for such stories*

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

(3) *Under assumption 1* $V_{old}\left(\cdot\right)$ *is increasing in* $c_{\tau}$*.*

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 payed 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 is only two types), and hence robust separation is achieved. This proposition also shows that when this happens, there is an unique strong rationalizable strategy profile, which is to play the repeated spot first best, since there are no confilct of interest between the parties, and both get their most prefered 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 $V_{old}\left(S_{old,\tau-1}, c_{\tau}\right)$ 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 $z_{\tau}$ may allow for local non-monotonicities of $V_{old}\left(S_{old,\tau-1}, c_{\tau}\right)$ with respect to the commitment cost $c_{\tau}$. Under the increasing conflict assumption, higher commitment costs increase the minimum utility of facing $\theta = old$. Defining $\mathbf{c}\left(s\right) = \min\left\{c \in C : V_{old}\left(s, c\right) \geq \underline{u}_{p}\right\}$, under this assumption we have $V_{old}\left(S_{old,\tau-1}, c_{\tau}\right) \geq \underline{u}_{p} \iff c_{\tau} \geq \mathbf{c}\left(S_{old,\tau-1}\right)$

In Appendix A I study in detail the solution $\left(\underline{r}\left(\cdot\right), \underline{w}\left(\cdot\right)\right)$ to 5.20. In Proposition A.1 I show that under assumption there exist a threshold $\hat{S} \in (0, S_{old}^{\max})$ 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 $V_{old}\left(S_{old,\tau-1}, c_\tau\right) = V_{old}\left(0, c\right)$. When $S_{old,\tau-1} \in \left(\hat{S}, S_{old}^{\max}\right)$ the promise keeping constraint starts binding, making $V_{old}\left(S_{old,\tau-1}, c_\tau\right)$ strictly increasing in this interval. Figure 4 illustrates the results.

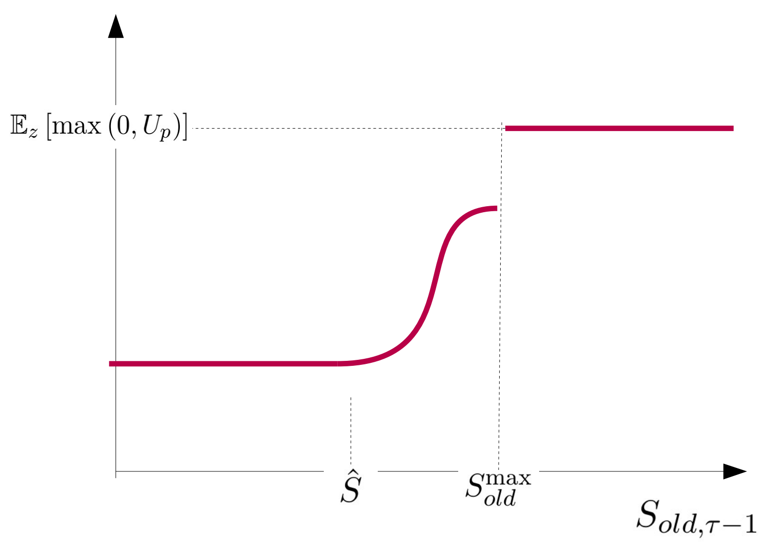


Figure 4. Worst Rationalizable Payoff $\underline{\mathcal{V}}_{old}\left(s, c\right)$

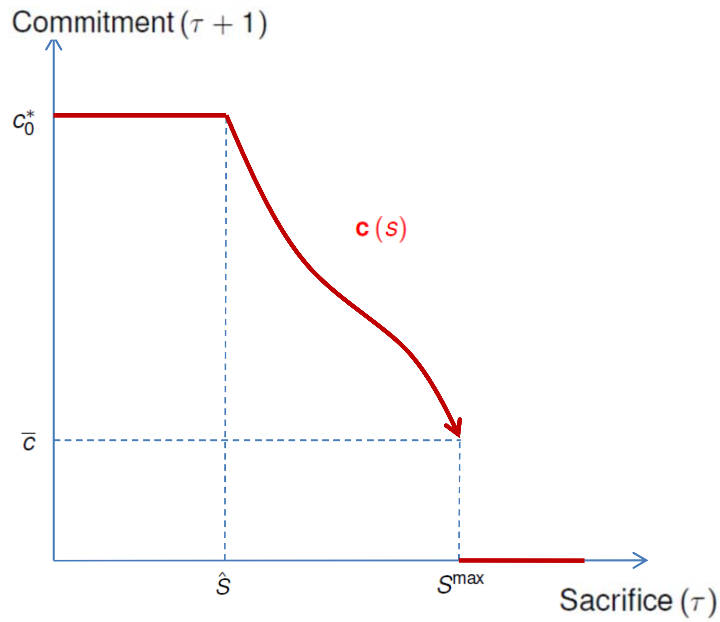


FIGURE 5. Minimum commitment cost function

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, and therefore play 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 $\sigma^*_{new}$. 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 $\tau \geq 1$:

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

The state variable $s_{\tau-1}$ gives the sacrifice for $\theta = old$ as long as history $h^\tau$ is $old-$rationalizable. If at some $\tau$ 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

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

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

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

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

To get a recursive formulation of $\underline{W}_{new}(h^\tau)$, 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

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

subject to

$$\mathcal{V}(s,c) \geq \underline{u}_p \tag{5.26}$$

and

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

In Lemma C.7 I show $T$ is a contraction with modulus $\beta$. 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 $\mathcal{W}_{new}(\cdot)$ that solves the associated Bellman equation $T(\mathcal{W}_{new})(\cdot) = \mathcal{W}_{new}(\cdot)$; which can expressed as

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

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 $(1-\beta)\left[U_p(z) - c\right] r$ and reputation gains $\beta \mathcal{W}_{new}\left[s'(z)\right]$, which depend only on the sacrificed that would be observed at the begining 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 begining 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 $c(s)$ and $r(s,z)$ be the policy functions associated with the Bellman equation 5.28. Then, for all $h^\tau \in \mathcal{H}(\sigma^*_{new})$*

(1) $\underline{W}_{new}(h^\tau) = \mathcal{W}_{new}(s_{\tau-1})$, $c^*(h^\tau) = c(s_{\tau-1})$ *and* $r^*(h^\tau, c_\tau, z_\tau) = r(s_{\tau-1}, z_\tau)$
(2) *If* $s_{\tau-1} > S^{\max}_{old}$ *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* $s_{\tau-1} \leq S^{\max}_{old}$ *and Assumption 1 holds,* $c^*(h^\tau) = \mathbf{c}(s_{\tau-1})$

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 7 previews the shape of the optimal policy $r^*(z) = r^*(U_p, U_{old})$ over the set of states $Z = \left[\underline{U}, \overline{U}\right]^2 \subset \mathbb{R}^2$. Regions where $r^*(U_p, U_{old}) = 1$ (i.e. $d$ takes the emergency action) are depicted in red, and $r^*(U_p, U_{old}) = 0$ in green. In the bottom we include the spot optimum strategy for $\theta = new$ and for agent $p$. [25] In the right margin, we draw the analogous scale for $\theta = old$.

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

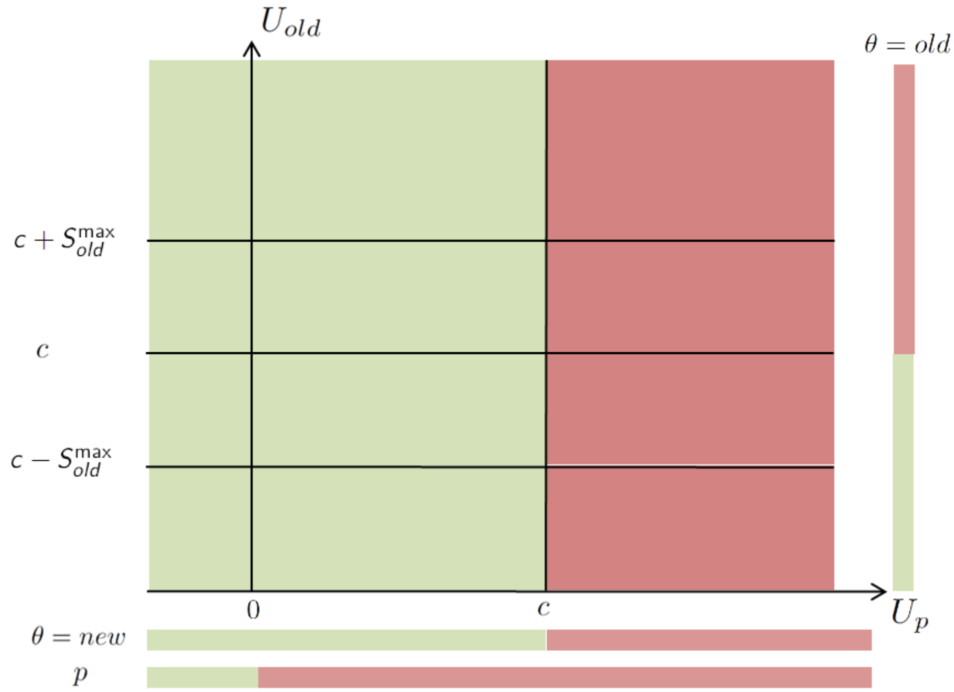


FIGURE 6. Static Best policy $r^{spot}(z)$ for $\theta = new$

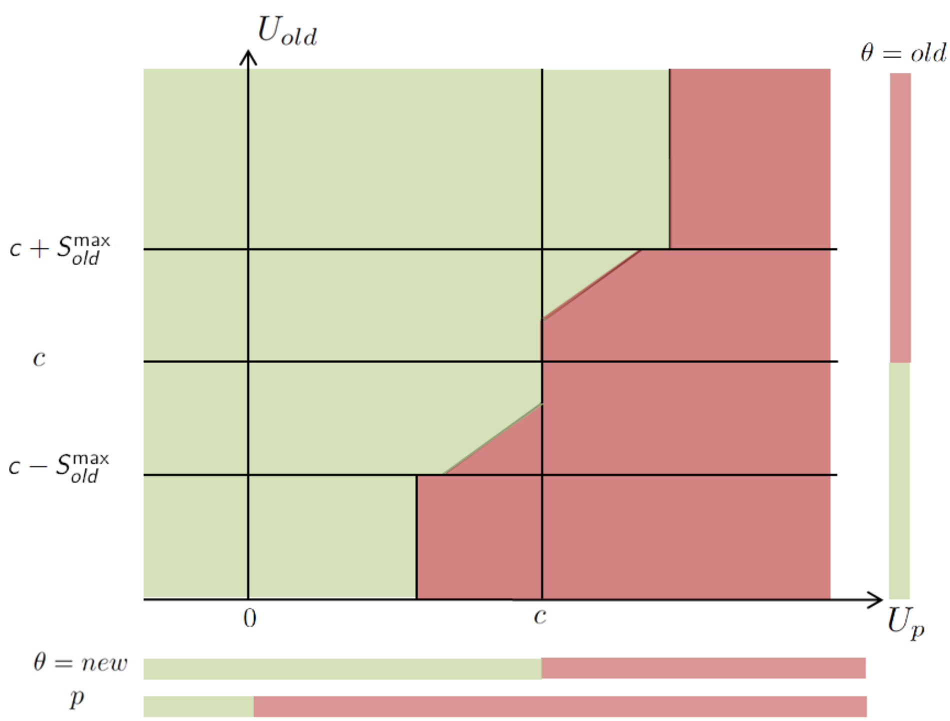


FIGURE 7. Optimal robust policy $r^{*}(z)$ for $\theta = new$

By comparing the optimal robust policy of Figure 7 to the spot optimal strategy for $\theta = new$ in Figure 6, 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 $U_{old} > c$ so that $r^{spot}_{old}(z) = 1$. Figure 7 shows then that the optimal policy over this region prescribes $r^*(U_p, U_{old}) = 0$ on a strictly larger set of states than $r^{spot}_{new}(z)$. The intuition for this phenomenon is simple: if $d$ plays $r_\tau = 0$, then $p$ will observe an opportunity cost payed by $\theta = old$ of $S_\tau = U_{old} - c > 0$ 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 $R_1 = \{z \in Z : U_{old} > c + S^{\max}_{old}\}$, which corresponds to the uppermost horizontal strip of Figure 7. 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 $S = U_{old} - c$ would be strictly greater than any potential net present value gain from switching to the best rationalizable payoff (given by $S^{\max}_{old}$). Therefore, if $\theta = new$ chooses $r^*(z) = 0$ she would strongly separate from tomorrow on, achieving the first best payoff $\mathbb{E}_z[\max(0, U_p)]$. However, if she chose $r^*(z) = 0$ 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

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

$$U_p \leq c + S^{\max}_{new} \tag{5.29}$$

If $U_p \leq c$ 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 $U_p > c$, $\theta = new$ 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 $R_2 = \left\{z \in Z : U_{old} \in \left(c + \hat{S}, c + S^{\max}_{old}\right)\right\}$. 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 $c_{\tau+1} = \mathbf{c}(U_{old} - c)$. Therefore, the same analysis from region $R_1$ applies here, with the only difference that now the continuation value will be $\mathcal{W}(U_{old} - c) < \mathbb{E}_z[\max(0, U_p)]$. Then, $r^*(z) = 0$ over region $R_2$ if and only if

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

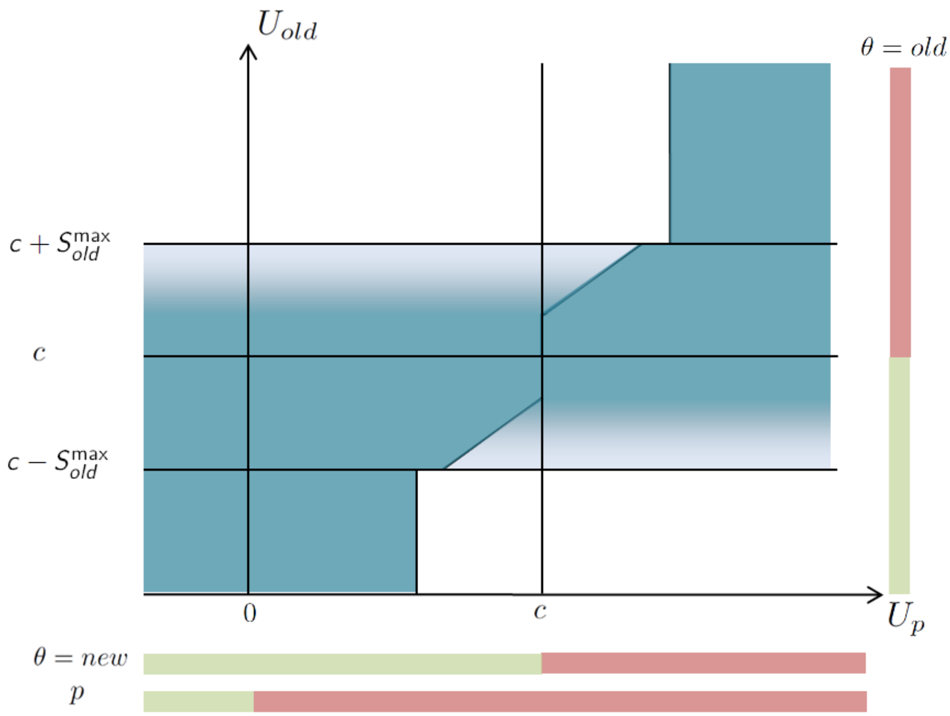


FIGURE 8. Optimal Robust Sacrifice

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

Finally, study region $R_3 = \left\{z \in Z : U_{old} \in \left(c, c + \hat{S}\right)\right\}$. See that regardless of the the action, the sacrifice potential $U_{old} - c$ 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. Therefore the optimal robust policy is just the spot optimal policy: $r^*(z) = 1 \iff U_p > c$ .

Is easy to see that when $U_{old} < c$, 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 8 we illustrate 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 maker 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 $U_{old} > c + S_{old}^{\max}$ and $U_p \in (c, c + S_{new}^{\max})$ 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.

A perhaps troubling feature of the robust policy 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 **?**

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. However, it seems intuitive that there should be some momentum in reputation gaining. The basic idea is that gaining reputation at time $\tau$ 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 I have been using so far does not allow for an easy characterization of the stochastic process followed by the commitment cost $\mathbf{c}\left(S_{\tau-1}\right)$. 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*

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

*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 green button strategy $(r = 0)$ as the status quo: it is the strategy that both the good guy, and a trustworthy bad 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 is when $d$ 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.** *If Assumption 2 holds, then*

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

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

To show our 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 it's 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* $\tau \in \mathbb{N}$

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

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

This proposition states one of the most important results: the probability of reaching separation from the time inconsistent type is exponentially decreasing in $\tau$. 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. $\mathcal{B}_p \subseteq \mathcal{B}_p^s$) we will also achieve separation in exponential time, the probability of separation can only be higher for any smaller belief sets. In the next section we will explore some "reasonable" restrictions we could impose, and see how the solution would be improved.

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 we will be increasing both discount rates. 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 $h^\tau$ 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 $\sigma_d^*$. Then, there exist $\hat{q} > 0$ such that $q(h^\tau, \beta) > \hat{q}$ for all $h^\tau \in \mathcal{H}(\sigma_d^*), \beta \in (0,1)$*

**Proof.** See Appendix C □

The previous lemma shows the existence of a number $\hat{q} > 0$ such that no matter the discount rate $\beta$ , 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 $\mathbb{E}_z\{\max(0, U_p)\}$ 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*

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

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

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

where $\tau \sim Geom(\hat{q})$ . This is true since $\mathbb{E}_z\{r(h^\tau, z)(U_p(z) - c(h^\tau))\} \geq 0$ 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

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

Therefore

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

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

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. **Restrictions on Beliefs.** In certain situations, the policy maker may have more information about the public $p$'s beliefs. I describe how this may be incorporated into the problem. Gains of reputation are not permanent, so a natural question to ask is: what restrictions on beliefs make reputation gains permanent? That is, when is $c_\tau \geq c_{\tau+k}$ for every $k$?

Formally, say a strategy $\sigma_d \in \Sigma_d$ exhibits *permanent reputation gains* if and only if $c^{\sigma_d}\left(h^{\tau+1}\right) \leq c^{\sigma_d}\left(h^\tau\right)$ for all histories $h^\tau, h^{\tau+1} \in \mathcal{H}\left(\sigma_d\right)$. We already know that the optimal robust strategy does not satisfy this property. The goal then is to find what type of restrictions on beliefs should we impose to get a robust implementing strategy that exhibits permanent reputation gains. Say a belief system $\pi_d$ is $\hat{\sigma}_d$ - *nondecreasing* if and only if, for all histories $h^\tau, h^{\tau+1} \in \mathcal{H}\left(\hat{\sigma}_d\right)$,

$$\mathbf{NPV}_{old}^{\pi_d}\left(\hat{\sigma}_d \mid h^{\tau+1}\right) \geq \mathbf{NPV}_{old}^{\pi_d}\left(\hat{\sigma}_d \mid h^\tau\right).$$

This means that under belief $\pi$, $\theta = old$ cannot get less than what she expected in the previous period, by playing strategy $\hat{\sigma}_d$. Denote also $\mathcal{ND}\left(\hat{\sigma}_d\right) \subset \Sigma_d$ be the set of best responses to $\hat{\sigma}_d-$nondecreasing beliefs.

**Proposition 6.4.** *Take a belief restriction set $\mathcal{B}_p \subseteq \mathcal{B}_p^s$ and $\sigma_d^*$ the optimal robust policy in $\mathcal{B}_p$. Then*

$$\sigma_d^* \text{ exhibits permanent reputation gains} \Longleftrightarrow \mathcal{B}_p \subseteq \mathbf{S}_p\left[\{\theta = old\} \times \mathcal{ND}\left(\sigma_d^*\right)\right].$$

*That is, $p$ is strongly certain that $\theta = old$ has $\sigma_d^* -$ nondecreasing beliefs.*

**Proof.** I show necessity. Take a $old-$rationalizable pair $\left(\sigma_d, \pi_d\right)$ and a history $h^\tau \in \mathcal{H}\left(\sigma_d^*\right)$ such that $c_k = c_0^*$ for all $k < \tau - 1$ and $c_\tau < c_0^*$. This is a history where there has been only one gain in reputation so far, and which has been realized only in the last period. The fact that the commitment cost decreased has, as a necessary condition, that the observed sacrifice should be higher than certain threshold level $\hat{S}$; i.e.

$$\mathbf{NPV}_{old}^{\pi_d}\left(\sigma_d \mid h^\tau\right) \geq S_\tau \tag{6.8}$$

Also, because $h^\tau \in \mathcal{H}\left(\sigma_d^*\right)$ condition (6.8) also holds for $\sigma_d^*$.Then, the only way the commitment cost could go up in some other history $h^{\tau+s} \in \mathcal{H}\left(\sigma_d^*\right)$, is that $\mathbf{NPV}_{old}^{\pi_d}\left(\sigma_d \mid h^{\tau+s}\right) < S_\tau$. But since $\pi_d$ is $\sigma_d^*-$nondecreasing, we have

$$\mathbf{NPV}_{old}^{\pi_d}\left(\sigma_d^* \mid h^{\tau+s}\right) \geq \mathbf{NPV}_{old}^{\pi_d}\left(\sigma_d^* \mid h^\tau\right) \geq S_\tau > \mathbf{NPV}_{old}^{\pi_d}\left(\sigma_d \mid h^{\tau+s}\right)$$

implying that $\sigma_d$ is dominated by $\sigma_d^*$ at $h^{\tau+s}$ . Then, the fact that the net present value is always increasing will imply that the resulting commitment cost will be always non-increasing. □

Intuitively, to get a strategy with permanent reputation gains, the assumption we need to make on $p's$ beliefs about the time inconsistent type are the following: if $\theta = old$ and we have observed that $S_{\tau-1} = S$, then the fact that she was willing to sacrifice $S$ utils will "stick", and $p$ will always think that $\theta = old$ will not settle for any smaller net present value. This feature of beliefs are actually pretty common in dynamic adverse selection and signaling problems.

Note that the important restriction is about $p's$ higher order beliefs: they are not about what $p$ thinks $d$ will do, but rather what $p$ believes $d$ believes about the continuation game. While working directly over system of beliefs can always be implemented, assumptions about higher order beliefs are not very transparent in this framework. In Appendix (A) I explore a different approach, by modeling restrictions on beliefs as type spaces, which allow the modeling of assumptions on higher order beliefs more tractable.

## 7. **Extensions And Further Research**

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 **?** and **?**. 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 indiference between trusting or not. As such, the public would increase commitment costs relative to the levels chosen by the new regime government.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.

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 $\Theta_d$ and hence the characterization of $\underline{V}(h^\tau, c_\tau)$ would now be:

$$\underline{V}(h^\tau, c_\tau) = \min_{\theta \in \Theta_d} \mathcal{V}_\theta (S_{\theta,\tau-1}, c_\tau) \tag{7.1}$$

where the function $\mathcal{V}_\theta(c, s)$ is the minimum problem in (5.20) for a given payoff type $\theta$. 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.

## 8. **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

As mentioned before, the decision maker $d$ acting as a "policy maker", may have some information about people's beliefs about what strategy may be played by $d$ , as well as as beliefs $d$ may hold, which may involve assumptions about higher order beliefs (what agent $i$ believes about $j$, what $i$ believes about $j$'s beliefs about her beliefs, and so on. As reviewed in **?** the literature on epistemic game theory distinguishes between two approaches: an explicit and an implicit approach. In the *explicit approach*, beliefs are modeled as subjective probability measures over (1) the other players strategy, and (2) probability measures over the beliefs of the other agents (which are themselves probability measures over the player's own strategy), etc. In the *implicit approach*, beliefs are formed over other player strategies and "types", where a type is directly mapped to a belief over strategies and types played by the other agent. We will argue that for most applications, the implicit approach will be a more tractable modeling assumption.

Formally, the explicit approach consists on modeling people's beliefs as a hierarchy of beliefs over the types and strategies that $d$ could play, the beliefs that $d$ may have about $p$'s beliefs, and so on, ad infinitum. This means that a *belief hierarchy* is a sequence of measures $(\pi_0, \pi_1, \pi_2, ...)$ where $\pi_0$ is a CPS over $\{old, new\} \times \Sigma_d$ (the payoff types of $d$ and with the strategies that she can choose from), $\pi_1$ is a measure over the space of such probability systems (i.e. $\pi_1 \in \Delta^{\mathcal{H}}\left(\Delta^{\mathcal{H}}\left(\{old, new\} \times \Sigma_d\right)\right)$ ), $\pi_2 \in \Delta^{\mathcal{H}}\left(\Delta^{\mathcal{H}}\left(\Delta^{\mathcal{H}}\left(\{old, new\} \times \Sigma_d\right)\right)\right)$ and so on. The set of all possible coherent hierarchies of beliefs[26] is denoted by $H_p^*$ , and can be shown to have desirable topological properties[27] Therefore, information about people's beliefs can be then represented as restrictions over the set of all coherent hierarchies; i.e. we can represent our information on beliefs as a subset $I \subset H_p^*$

While this approach has the advantage of being explicit about higher order beliefs, it is cumbersome to work with. Note that we have to consider beliefs over exponentially larger spaces. As an alternative, Harsanyi proposed an implicit approach (**?**). He suggested that one could bundle all relevant information about beliefs and payoff parameters into different "(epistemic) types" of agents, in the same way that we think about payoff types. As such, we can model the system as a Bayesian game, with a larger type space. We formalize this idea for our context:

**Definition A.1** (**Type Space**)**.** A type space $\mathcal{T}$ is a 5-tuple $\left(T_p, T_d, \hat{\theta}\left(\cdot\right), \hat{\pi}_p\left(\cdot\right), \hat{\pi}_d\left(\cdot\right)\right)$ where $T_i$ are sets of types for each agent, $\hat{\theta} : T_d \to \{old, new\}$ is a function that assigns to each type $t_d \in T_d$ to a payoff type $\hat{\theta}\left(t_d\right)$ , and $\hat{\pi}_i : T_i :\to \Delta^{\mathcal{H}}\left(T_{-i} \times \Sigma_{-i}\right)$ assigns a CPS $\hat{\pi}_i\left(t_i\right)$ over strategy and type pairs $(t_{-i}, \sigma_{-i})$

[26]See Definition B.1 in Appendix B

[27]See Appendix B for a formal treatment of the topological properties of the set of all coherent hierarchies.

I intersect the approaches of **?** and **?**, a type encodes it's payoff type (since $d$ knows its type) together with his beliefs about the other agent. The set of "states of the world" $\Omega_{-i}$ that agent $i$ form beliefs over is then the set of pairs $\omega_{-i} = (t_{-i}, \sigma_{-i})$ of strategies and types of the other agent. Unlike **?** however, agents hold beliefs over other player's types and strategies as well, since in a dynamic environment these are not perfectly observed.[28]This method of representing restrictions on beliefs has the advantage of being more compact (since only first order beliefs have to be specified) and also being a natural generalization of Bayesian games. For example, suppose the information we have about $p$ is that $p$ thinks that $\theta = old$ could have two possible beliefs about the future play of the game: optimistic (expecting her best equilibrium to be played) or pessimistic (expecting her worst equilibrium). If this was the case, we could model this situation by simply augmenting the type space by creating two copies of the old type: an "optimistic old type" $t_O$ and a "pessimistic old type" $t_W$ with the same payoff parameter ($\hat{\theta}(t_O) = \hat{\theta}(t_W) = old$ ) but with different beliefs $\hat{\pi}_d(t_O) \neq \hat{\pi}_d(t_W)$. The type sets assigned to $d$ would then be $T_d = \{new, t_O, t_P\}$.

In compact static games, **?** and **?** showed that these two approaches are equivalent: for any subset of possible hierarchies of beliefs $I \subset H^*$ there exist a type space $\mathcal{T}$ that generates the exact same belief hierarchies[29] and vice versa.

In particular, if no restrictions are imposed on hierarchies (i.e. $I = H^*$) there exist a *universal type space* $\mathcal{T}^*$ which is capable of generating all possible hierarchies of beliefs. In another paper, (**?**) I extend **?** to (non-metrizable)[30]topological spaces, to show that this is also true in a relevant class of extensive form games. In Appendix B I provide an application to our particular setting, and also a formal description of how we can make the mapping between these two approaches.

Because of this equivalence result, I will use the implicit approach throughout this paper, when modeling $d$'s information and assumptions about $p's$ beliefs. I will further consider only *compact type spaces,* where the sets $T_i$ are compact and Hausdorff topological spaces (with some topology) and the belief functions $\pi_i(t_i)$ are continuous in the weak convergence sense: if a sequence $t_{i,n} \to t_i$ then $\pi_i(t_{i,n})(. \mid h)$ converges in distribution to $\pi_i(t_i)(. \mid h)$ for all $h \in \mathcal{H}$. For most applications this will not be restrictive, since any type space that is "closed" is homeomorph to a subset of the universal type space $\mathcal{T}^*$[31] which is itself compact (see Theorem B.1 in Appendix B), making $\mathcal{T}$ itself compact.

For an epistemic type $t_d \in T_d$ and a strategy $\sigma_d \in \Sigma_d$ define the expected continuation value for type $t_d$ as

---

[28]This is the *interactive epistemic characterization* of **?** and **?**. This definition also corresponds to **?** notion of "conjectures"

[29]Formally, there is always a *belief morphism* between both types of spaces, as studied in Appendix B

[30]This is a relevant extension, since a very large class of relevant games in economics cannot be modeled with metrizable type spaces. For example, any infinitely repeated dynamic game with a continuous strategy space (such as Cournot duopoly, or most macro applications) are not metrizable.

[31]This is the main lesson from the Universal Type Space Theorem of **?**,**?** and **?**. See Theorem B.1 in Appendix B

$$W^{t_d}(\sigma_d \mid h) = W^{\hat{\pi}}_{\hat{\theta}}(\sigma_d \mid h) \text{ with } \hat{\theta} = \hat{\theta}(t_d), \hat{\pi} = \hat{\pi}_d(t_d) \tag{A.1}$$

Likewise, given a type $t_p \in T_p$ and a strategy $\sigma_p \in \Sigma_p$ define $p$'s expected value as

$$V^{t_p}(\sigma_p \mid h) = V^{\hat{\pi}}(\sigma_p \mid h) \text{ where } \hat{\pi} = \hat{\pi}(t_p)$$

Also, we write $SBR(t_i) = SBR_{\hat{\theta}(t_i)}[\hat{\pi}(t_i)]$. An agent is then *sequentially rational* if the strategy she chooses is a sequential best response to her beliefs: i.e. $\sigma_i \in SBR(t_i)$ The interactive epistemic representation of types allows us to easily write this assumption; as the subset of *sequentially rational states* $R_i \subset T_i \times \Sigma_i$ defined as:

$$R_i := \{(t_i, \sigma_i) \in T_i \times \Sigma_i : \sigma_i \in SBR_i(t_i)\} \tag{A.2}$$

for $i \in \{d, p\}$. We write $\Sigma_i^*(\mathcal{T}) \subseteq \Sigma_i$ as the set of all sequentially rational strategies.

**Definition A.2** (**Robust Implementation**). Given a type space $\mathcal{T} = \left(T_d, T_p, \hat{\theta}, \hat{\pi}_p, \hat{\pi}_d\right)$ we say that a strategy $\sigma_d$ is a *robust implementation of trust* if and only if for all histories $(h^\tau, c_\tau) \in \mathcal{H}_p(\sigma_d)$, all $t_p \in T_p$ and all $\sigma_p \in SBR(t_p)$ we have $a^{\sigma_p}(h^\tau, c_\tau) = 1$

Besides the information we have about beliefs (modeled by a type space $\mathcal{T}$ ) we might also know (or be willing to assume) some common certainty restrictions on agents beliefs. Following the construction of subsection 4.4 we can extend the definitions of weak and strong rationalizable sets to type spaces, where the sets $WCR_i^k(\mathcal{T})$ and $SCR_i^k(\mathcal{T}) \subset T_i \times \Sigma_i$ correspond to all the weak and strong rationalizable pairs. In Proposition B.3 we adapt the result of **?** to show that for compact type spaces, these sets are compact, which also implies that the set of weak and strong rationalizable sets $WCR_i^\infty(\mathcal{T})$ and $SCR_i^\infty(\mathcal{T})$ are non-empty, compact subsets of $T_i$. Together with the upper hemicontinuity of the sequential best response correspondence, this implies that the sets of weakly and strongly rationalizable strategies

$$\Sigma_i^w := SBR_i\{\mathbf{W}[WCR_i^\infty(\mathcal{T})]\}, \Sigma_i^s := SBR_i\{\mathbf{S}[SCR_i^\infty(\mathcal{T})]\} \tag{A.3}$$

are compact with respect to the product topology. With this formulation, we can work with common certainty assumptions (of rationality or other assumptions about beliefs) and still retain the type space representation we have been considering. When besides the restrictions on beliefs modeled by the type space $\mathcal{T}$, $d$ is also willing to make assumptions about common certainty of rationality, this can be thought of as refining the type space by getting a subspace $\hat{\mathcal{T}} \subset \mathcal{T}$ , formed only by types that have survived the iterative deletion procedure just described. That is $t_i \in \hat{T}_i = WCR_i^\infty(\mathcal{T})$ if we use common weak certainty, and analogously with $SCR_i^\infty(\mathcal{T})$ for strong certainty.

With some abuse of notation, we will denote the type space resulting of this refinement as $\hat{\mathcal{T}} = \mathbf{WCR}^\infty(\mathcal{T})$ and $\hat{\mathcal{T}} = \mathbf{SCR}^\infty(\mathcal{T})$ for the case of strong certainty. When a type space $\mathcal{T}$ can be written as it's own weak rationalizable refinement (i.e. $\mathcal{T} = \mathbf{WCR}^{\infty}(\mathcal{T})$) we will say that $\mathcal{T}$ is consistent with weak common certainty of rationality, and analogously for strong certainty

Armed with this concepts, we can give a definition of robust implementation that relates to **?**.

**Definition A.3** (**Weak Robust Implementation**)**.** A strategy $\sigma_d$ is a weak robust implementation of trust if if it implements it for all type spaces $\mathcal{T}$ such that $\mathcal{T} = \mathbf{WCR}^{\infty}(\mathcal{T})$

In Appendix B I show how this is actually equivalent to doing robustness with the belief space $\mathcal{B}_p^w$, and likewise for $\mathcal{B}_p^s$

## Appendix B. Universal Type Space and Strong Rationalizable Strategies

In this section I adapt the results in **?** on the characterization of the Universal Type Space theorem and the study of the topological properties of the sets of weak and strong rationalizable strategies to my setup, for any compact type space we might consider. It generalizes **???** to general topological spaces, which is necessary because their results do not apply to my paper. Their results require either finite strategies or finite periods, because their are obtained by extending **?** to dynamic settings, which works with complete metrizable strategy spaces. However, infinitely repeated games typically involve using the weak convergence topology on the set of strategies, which is non-metrizable if, for example, agents have a continuum of actions in each period. This section is organized as follows: in subsection B.1 I introduce and give some results on the topology of strategy spaces. In subsection B.2 I introduce the notion of hierarchies of beliefs, and study their topological properties. In subsection B.3 I provide a version of the Universal Type Space Theorem (as in **??**) and finally, in subsection B.4 we apply the results we found in the previous sections to characterize the compactness of the set of weak and strong rationalizable strategies, a crucial result for the model studied in this paper.

### B.1. Topological Properties of Strategy Spaces.

We have that the set from which $p$ chooses is clearly Hausdorff, regular and compact $S_p := \{0, 1\}$.

We will now show that the set from which $d$ chooses is also Hausdorff, compact and regular since it is the product of two compact, regular and Hausdorff spaces:

$$S_d := C \times G \tag{B.1}$$

where $G = \mathcal{M}(Z, \{0, 1\})$, the set of measurable functions $g : Z \to \{0, 1\}$. $C \subset \mathbb{R}$ is compact by assumption, (and Hausdorff and regular because $\mathbb{R}$ is). We will show that $G$ is also a Hausdorff, compact and regular space with the product topology:i.e. point-wise convergence

$$g_n(.) \to g(.) \iff g_n(z) \to g(z) \text{ for all } z \in Z \tag{B.2}$$

The compactness follows from 3 reasons:

(1) $G \subset \{0,1\}^Z = \prod_{z \in Z} \{0,1\} = \hat{G}$ which is a compact space with the product topology described B.2, because of Tychonoff's Theorem. It is also Hausdorff and regular (Theorem 31.2 in **?**).
(2) $G$ is a closed subset of $\hat{G}$, because of the Dominated convergence theorem (Theorem 2.24 in **?**)
(3) $G$ is therefore compact (since it is a closed subset of a Hausdorff compact space, Theorem 26.2 in **?**)

The fact that $G$ is regular and Hausdorff is simply because it is a subspace of $\hat{G}$, which is a regular and Hausdorff space itself (Theorem 31.2 in **?**). Because $C$ and $G$ and both Hausdorff, compact and regular, $S_d$ is also Hausdorff, compact and regular. Strategies for $d$ are functions

$$\sigma_d : \mathcal{H}_d \to C \times G$$

which can be written as

$$\Sigma_d \equiv (C \times G)^{\mathcal{H}_d} = S_d^{\mathcal{H}_d}$$

which by Tychonoff's Theorem and Theorem 31.2 in **?**, is also Hausdorff, compact and regular, with the product topology; i.e.

$$\sigma_d^{(n)} \to \sigma_d \text{ if and only if } \begin{cases} c_n(h^\tau) \to c(h^\tau) & \text{for all histories } h^\tau \in \mathcal{H}_d \\ g_n(h^\tau, z) \to g(h^\tau, z_\tau) & \text{for all } (h^\tau, c_\tau, 1, z_\tau) \in \mathcal{H}_d \end{cases} \tag{B.3}$$

and

$$\sigma_p^{(n)} \to \sigma_p \text{ if and only if } a_n(h^\tau, c_\tau) \to a(h^\tau, c_\tau) \tag{B.4}$$

Note that, because of the boundedness of both $c$ and $g(\cdot)$ we can apply the Dominated convergence theorem (Theorem 2.24 in **?**) to show that the function $V(\cdot \mid h)$ defined in 4.11 is a continuous function of $\sigma \in \Sigma$, and using Theorem Theorem 2.25 in **?**, we can also show that the continuation value function $W_\theta(\cdot \mid h)$ defined in 4.10 is also continuous function of $\sigma \in \Sigma$ with the product topology. We summarize the results of this subsection in the following Lemma.

**Lemma B.1** (**Topology of $\Sigma_i$**). *The strategy spaces $\Sigma_i$ for $i \in \{p, d\}$ are Hausdorff, compact and regular topological spaces, with the topology of point-wise convergence (as in B.3 and B.4). Moreover, for all histories $h \in \mathcal{H}_i$, the conditional expected utility functions $V(\sigma \mid h)$ and $W_\theta(\sigma \mid h)$ as defined in 4.11 and 4.10 are continuous*

B.2. **Hierarchies of Beliefs.** Given a topological space $(X, \tau)$, define $\Delta(X)$ as the set of all Borel probability measures on $X$. If $X$ is a compact, Hausdorff space, then $\Delta(X)$ is also a Hausdorff and compact topological space (Theorem 3 in **?**) with the *weak-* topology*. This is the topology of the convergence in distribution: a sequence $\{\lambda_n\}_{n \in \mathbb{N}}$ converges in distribution to $\lambda$ (written as $\lambda_n \rightsquigarrow \lambda$) if and only if

$$\int f(x)\, d\lambda_n(x) \to \int f(x)\, d\lambda(x) \text{ for all } f \in \mathcal{M}(X, \mathbb{R}) \tag{B.5}$$

where $\mathcal{M}(X, \mathbb{R})$ is the set of all measurable functions with respect to the Borel $\sigma-$algebra. Moreover, all Borel probability measures on $X$ are also **regular** (Theorem 5, **?**). Therefore, using Tychonoff's theorem, the set $[\Delta(X)]^{\mathcal{H}}$ is also a compact, Hausdorff space with the product topology (having point-wise the weak topology). The set of conditional probability systems on $X$, which we write $\Delta^{\mathcal{H}}(X)$ is a closed subset of $[\Delta(X)]^{\mathcal{H}}$ (Lemma 1 in **?**), and therefore inherits compactness and Hausdorff property. We will say a conditional probability system $\pi$ is **regular** if and only if $\pi(\cdot \mid h)$ is a regular measure over $X$ for all $h \in \mathcal{H}$. These results are summarized in the following Lemma

**Lemma B.2** (**Topology of $\boldsymbol{\Delta^{\mathcal{H}}(X)}$**). *Given a Hausdorff and compact topological space $X$ and a family $\mathcal{H}$ of histories, the space $\Delta^{\mathcal{H}}(X)$ of conditional probability systems on $X$ is also a Hausdorff, compact space with the product topology of convergence in distribution: i.e. given a sequence $\{\pi_n\} \in \Delta^{\mathcal{H}}(X)$ and $\pi \in \Delta^{\mathcal{H}}(X)$, we say*

$$\pi_n \to \pi \text{ in } \Delta^{\mathcal{H}}(X) \iff \pi_n(\cdot \mid h) \rightsquigarrow \pi(\cdot \mid h) \text{ for all } h \in \mathcal{H} \tag{B.6}$$

*Moreover, every $\pi \in \Delta^{\mathcal{H}}(X)$ is regular.*

A useful corollary of Lemma B.2 will be needed for characterizing the best reply correspondence. Given a type space $\mathcal{T} = \left(T_p, T_d, \hat{\theta}(\cdot), \hat{\pi}_p(\cdot), \hat{\pi}_d(\cdot)\right)$, define now the functions $\mathcal{V}(\cdot \mid h) : \Sigma_p \times \Delta^{\mathcal{H}}(T_d \times \Sigma_d) \to \mathbb{R}$ and $\mathcal{W}_\theta(\cdot \mid p) : \Sigma_d \times \Delta^{\mathcal{H}}(T_p \times \Sigma_p)$ as

$$\mathcal{V}_p(\sigma_p, \pi_p \mid h) := V^{\pi_p}(\sigma_p \mid h) \tag{B.7}$$

and

$$\mathcal{W}_\theta(\sigma_d, \pi_d \mid h) := W_\theta^{\pi_d}(\sigma_d \mid h) \tag{B.8}$$

**Corollary B.1** (**Continuity of Expected Utility over types**). *If $T_i$ are compact topological spaces for $i = 1, 2$, then the functions $\mathcal{V}(\cdot \mid h)$ and $\mathcal{W}_\theta(\cdot \mid h)$ are continuous functions (in the weak topology).*

**Proof.** Since both functions are linear functionals in the space $\Delta(T_{-i} \times \Sigma_{-i})$, for continuity I only need to show boundedness of both functions. This follows from directly from Lemma B.2 and the Dominated convergence theorem (which makes the convergence the weak convergence). The continuity of $\mathcal{V}(. \mid h)$ with respect to $\sigma_d$ has already been established in Lemma B.1 □

I will now replicate here the inductive construction of the set of hierarchies of beliefs, as in **?**: Define first

$$X_p^0 := \Theta_d \times \Sigma_d \text{ and } X_d^0 = \Sigma_p \tag{B.9}$$

$$X_i^1 := X_i^0 \times \Delta^{\mathcal{H}_i}\left(X_j^0\right) \text{ for } i \in \{d,p\}$$

and in general

$$X_i^k := X_i^{k-1} \times \Delta^{\mathcal{H}_i}\left(X_j^{k-1}\right) \tag{B.10}$$

**Proposition B.1.** *$X_i^k$ is a Hausdorff and compact topological space for all $k = 0,1,2,....$ and $i \in \{d,p\}$. Moreover, $x \in X_i^k \iff x = (x_{k-1}, \pi_1, \pi_2, ..., \pi_{k-1})$ where $x_{k-1} \in X_i^{k-1}$ and $\pi_s$ is a regular CPS on $X_j^s$ for all $s = 1,2,...,k-1$.*

**Proof.** By induction, I will show that $X_i^k$ is compact, Hausdorff, and it consists of regular measures on its previous "level". Clearly is true for $k = 0$, since from B.9 and Lemma B.1, we know that $X_i^0$ is a Hausdorff and compact topological space. Now, assuming $\left\{X_i^{k-1}\right\}_{i\in\{p,d\}}$ is Hausdorff and compact, I need to show that $\left\{X_i^k\right\}_{i\in\{p,d\}}$ is also Hausdorff and compact. Using Lemma B.2 we then know that $\Delta^{\mathcal{H}}\left(X_j^{k-1}\right)$ is compact and Hausdorff, and consists of regular measures. This together with definition B.10 gives the desired result. The second result follows from **?** which show that we can write $X_i^k$ simply as

$$X_i^k = \Sigma_j \times \prod_{s=0}^{k-1} \Delta^{\mathcal{H}_i}\left(X_j^s\right) \tag{B.11}$$

□

Define the set of hierarchies of beliefs for agent $i \in \{p,d\}$ to be the set $H_i = \lim_{k\to\infty} X_i^k$, which can be written (according to B.11) as

$$H_i = \prod_{k=1}^{\infty} \Delta^{\mathcal{H}_i}\left(X_j^k\right) \tag{B.12}$$

So, an element $\mathbf{h} = (\pi_0, \pi_1, ...) \in H_i$ consists on a CPS $\pi_0$ on $\Sigma_j$ (the strategies of the other agent), a CPS $\pi_1$ on $\Delta^{\mathcal{H}_j}(\Sigma_i)$ (the CPS's of $j$ about $i's$ strategies), a CPS $\pi_2$ on $\Delta^{\mathcal{H}_j}\left(\Delta^{\mathcal{H}_i}\left(\Delta^{\mathcal{H}_j}(\Sigma_d)\right)\right)$, and so on. Clearly the space $H_i$ is compact and Hausdorff, because of Proposition B.1 and Tychonoff's theorem. We summarize these results below

**Proposition B.2** (**Topology of $\boldsymbol{H_i}$**). *The set of hierarchies of beliefs $H_i$ for $i \in \{p,d\}$ as defined in B.12 are Hausdorff and compact topological spaces, with the point-wise convergence in each level:*

$$\mathbf{h}_n = \left(\pi_n^k\right)_{k\in\mathbb{N}} \to \mathbf{h} = \left(\pi^k\right)_{k\in\mathbb{N}} \iff \pi_n^k(\cdot \mid h) \rightsquigarrow \pi^k(\cdot \mid h) \text{ for all } k \in \mathbb{N}, h \in \mathcal{H}_i \tag{B.13}$$

*Moreover, for all hierarchies $\mathbf{h} = \left(\pi^k\right)_{k\in\mathbb{N}}$ and all $k \in \mathbb{N}$, we have that $\pi^k \in \Delta^{\mathcal{H}_i}\left(X_j^k\right)$ is a regular CPS*

B.3. **Construction of the Universal Type Space.** Not all hierarchies of beliefs will be "rational", in the sense that upper level beliefs (say, $k-$order beliefs) may not be consistent with lower level beliefs. We say that a hierarchy $h \in H_i$ is *coherent* when different levels of beliefs are consistent with each other. The formal definition is given by **?** (Definition 1) :

**Definition B.1** (**Coherency**)**.** A hierarchy of beliefs $\mathbf{h} \in H_i$ is *coherent* if and only if

$$\text{(B.14)} \qquad \text{mrg}_{X_j^{k-1}} \pi^{k+1} (\cdot \mid h) = \pi^k (\cdot \mid h) \text{ for all } h \in \mathcal{H}_i, k \in \mathbb{N}$$

where $\text{mrg}_{X^{k-1}} \pi^{k+1}$ is the marginal of measure $\pi^{k+1}$ on the projection $X^{k-1}$.

Definition B.1 is also identical to the definition of projective sequence of regular Borel probability measures, as in **?** (Definition 7), since by B.2 we know that all measures (and their projections) involved in the hierarchies are regular probability measures. We write $H_i^*$ to mean the set of coherent hierarchies of beliefs for agent $i$. Is easy to see that coherency is a closed restriction of the space $H_i$, which readily implies that $H_i^*$ is itself a Hausdorff and compact subspace of $H_i$.

Define

$$\text{(B.15)} \qquad T_p^* = H_p^*, T_d^* = \Theta_d \times H_d^*$$

So, the universal type sets are simply the sets of all coherent hierarchies of beliefs for each agent. See also that the identity mappings make sense, in that an element in $H_i^*$ is precisely a coherent CPS on the elements of $H_j^*$. In order to understand the maximality property of the type space we want to construct, in the sense that all other type spaces are in some way embedded into it, I need to define the concept of *type-morphisms*. Given two type spaces $\mathcal{T} = \left(T_p, T_d, \hat{\theta}(\cdot), \hat{\pi}_p(\cdot), \hat{\pi}_d(\cdot)\right)$ and $\mathcal{T}' = \left(T'_p, T'_d, \hat{\theta}'(\cdot), \hat{\pi}'_p(\cdot), \hat{\pi}'_d(\cdot)\right)$ and a function $\varphi_i : \Sigma_j \times T_j \to \Sigma_j \times T'_j$ with $i \in \{p, d\}$ and $j \neq i$, define $\tilde{\varphi}_i : \Delta^{\mathcal{H}_i}(\Sigma_j \times T_j) \to \Delta^{\mathcal{H}_i}\left(\Sigma_j \times T'_j\right)$ as the function associating to each CPS $\pi_i$ on $\Sigma_j \times T_j$ the induced CPS $\tilde{\varphi}_{ii}(\mu_i)$ over $\Sigma_j \times T'_j$, as defined in **?** (Subsection 3.1). Formally, given $\pi_i \in \Delta^{\mathcal{H}_i}(\Sigma_j \times T_j)$ we have

$$\text{(B.16)} \qquad \hat{\varphi}_i(\mu_i)(A \mid h) = \pi_i\left(\varphi_i^{-1}(A) \mid h\right) \text{ for all measurable } A \subset \Sigma_j \times T'_j, h \in \mathcal{H}_i$$

i.e. it gives events in $\Sigma_j \times T'_j$ the probability according to $\mu_i$ in the pre-image of that event in $\Sigma_j \times T_j$.

**Definition B.2** (**Type-morphisms (?)**)**.** Given two type spaces $\mathcal{T} = \left(T_p, T_d, \hat{\theta}(\cdot), \hat{\pi}_p(\cdot), \hat{\pi}_d(\cdot)\right)$ and $\mathcal{T}' = \left(T'_p, T'_d, \hat{\theta}'(\cdot), \hat{\pi}'_p(\cdot), \hat{\pi}'_d(\cdot)\right)$ we say that a pair of functions $\varphi = (\varphi_p, \varphi_d)$ where $\varphi_i : T_i \to T'_i$ is a **type morphism** from $\mathcal{T}$ to $\mathcal{T}'$ if and only if the functions $\varphi_i$ are continuous, and satisfy

$$\text{(B.17)} \qquad \hat{\pi}'_i[\varphi_i(t_i)] = \tilde{\varphi}_i[\hat{\pi}_i(t_i)] \text{ for all } t_i \in T_i, i \in \{p, d\}$$

and

$$\text{(B.18)} \qquad \hat{\theta}'[\varphi_d(t_d)] = \hat{\theta}(t_d) \text{ for all } t_d \in T_d$$

When $\varphi$ is a homeomorphism we say that $\mathcal{T}$ and $\mathcal{T}'$ are *type-isomorphic*.

Conditions B.17 and B.18 state that the beliefs and utility parameters (respectively) of all types in $\mathcal{T}$ can be mapped (in a continuous way) into beliefs and parameters of $\mathcal{T}'$ . The intuitive idea of this definition is that $\mathcal{T}$ is "smaller" than $\mathcal{T}'$, since every type in $T_i$ can be mapped to a subset of types in $T_i'$ (i.e. the image of $\varphi$) that have essentially the same epistemic properties: same beliefs and same utility parameters.

The following Theorem is a simple consequence of Theorem 8 in **?** and Proposition 3 in **?**, adapted to the modified topological assumptions of this model.

**Theorem B.1** (**Universal Type Space Theorem**). *The sets $T_p^*$ and $T_d^*$ defined in B.15 satisfy:*

$$T_p^* \text{ is homeomorphic to } \Delta^{\mathcal{H}_d}\left(T_d^* \times \Sigma_d\right) \tag{B.19}$$

*and*

$$T_d^* \text{ is homeomorphic to } \Theta_d \times \Delta^{\mathcal{H}_d}\left(T_p^* \times \Sigma_p\right) \tag{B.20}$$

*with homeomorphisms $Q_p : T_p^* \to \Delta^{\mathcal{H}_P}\left(T_d^* \times \Sigma_d\right)$ and $Q_d : T_d^* \to \Theta_d \times \Delta^{\mathcal{H}_d}\left(T_p^* \times \Sigma_p\right)$. The type space $\mathcal{T}^* = \left(T_p^*, T_d^*, \hat{\theta}^*(\cdot), \hat{\pi}_p^*(\cdot), \hat{\pi}_d^*(\cdot)\right)$ with $\left(\hat{\theta}^*(\cdot), \hat{\pi}_d^*(\cdot)\right) = Q_d(\cdot)$ and $\hat{\pi}_p^*(\cdot) = Q_p(\cdot)$ is called the **Universal Type Space**, and has the following property: for any other type space $\mathcal{T}$ there exists a (unique) type morphism $\varphi$ between $\mathcal{T}$ and $\mathcal{T}^*$*

**Proof.** Proposition B.2 tells us that all measures in a hierarchy are regular measures. This, together with Theorems 8 and 9 in **?** proves conditions B.19 and B.20, by applying the Generalized consistency theorem to each individual history $h \in \mathcal{H}$ and constructing the homeomorphism by defining it history by history. The universality condition is an almost direct application of Proposition 3 in **?** since we can easily replicate the proof step by step with our topological assumptions. □

B.4. **Topology of Rationalizable sets.** In this section I show that the set of rationalizable strategies for any compact type space is in fact, a compact subset of the set of strategies, which I characterized in subsection B.1. Moreover, the set of strongly rationalizable strategies will be in fact, a subset of the weak rationalizable strategy set, implying that strong rationalizability is a closed, stronger solution concept. This will be useful when using the structure theorems in **?**. The main tool I will be using to prove stated in **?**

**Theorem B.2** (**Berge's Theorem of the Maximum** (**?**)). *Let $X$ and $Y$ be topological spaces, with $Y$ regular, a continuous function $f : X \times Y \to \mathbb{R}$ and a continuous, non-empty and compact valued correspondence $\Gamma : X \rightrightarrows Y$. Then the function*

$$M(x) := \max_{y \in \Gamma(x)} f(x, y)$$

*is well defined and continuous, and moreover, the correspondence*

$$g(x) := \arg\max_{y \in \Gamma(x)} f(x, y)$$

*is non-empty, compact valued and upper hemi-continuous.*

The most important consequence of the theorem of the maximum is the continuity and upper hemicontinuity of the value and best response functions, respectively

**Proposition B.3** (**Continuity of Sequential Best Responses**). *For any type space* $\mathcal{T} = \left(T_p, T_d, \hat{\theta}(\cdot), \hat{\pi}_p(\cdot), \hat{\pi}_d(\cdot)\right)$ *and any* $t_i \in T_i$ *define the sequential best response correspondence* $SBR_i : T_i \to \Sigma_i$ *that gives the sequential best responses for type* $i$. *Then, if* $\hat{\pi}_i(\cdot)$ *for* $i \in \{p, d\}$ *are continuous functions,* $SBR_i(t_i)$ *is a non-empty, compact valued and upper-hemi continuous correspondence.*

**Proof.** I will only show the continuity of $SBR_d$, since $SBR_p$ follows a similar (and easier) argument. Corollary B.1 tells us that $\mathcal{W}_\theta(\sigma_d, \pi_d \mid h)$ is a continuous function of the CPS $\pi_d$, and Proposition B.1 tells us that it is also a continuous function of $\sigma_d$ (taking $\pi_d$ as given), which makes $\mathcal{W}_\theta(\cdot \mid h)$ a continuous function over $\Sigma_d \times \Delta^{\mathcal{H}_d}(T_p \times \Sigma_p)$ (with the product topology). Proposition B.1 also implies that the set $\Sigma_d$ is regular, Hausdorff and compact. The domain of the program is $\Sigma_d$, which is a constant correspondence, hence continuous, non-empty and compact valued (since $\Sigma_d$ is compact). Therefore, we can apply the theorem of the maximum B.2 to show that the correspondence

$$\phi(\theta, \pi_d \mid h) := \underset{\sigma_d \in \Sigma_d}{\arg\max}\ \mathcal{W}_\theta(\sigma_d, \pi_d \mid h)$$

is a continuous, non-empty and compact valued u.h.c correspondence of $(\theta, \pi_d)$ for all $h \in \mathcal{H}$, (continuity on $\theta$ comes for free with it's finiteness) which therefore implies that the correspondence $\hat{\phi}(\theta, \pi) = (\phi(\theta, \pi_d \mid h))_{h \in \mathcal{H}}$ is also a continuous, non-empty and compact valued u.h.c correspondence. The desired result then follows from the continuity of $\hat{\pi}_d$, since

$$SBR_d(t_d) = \hat{\phi}\left[\hat{\theta}(t_d), \hat{\mu}_d(t_d)\right]$$

a composition of an u.h.c correspondence with a continuous function, which is also a u.h.c correspondence, as we wanted to show. Compactness also follows from continuity (Weierstrass). □

Now I present the main result of this section

**Theorem B.3** (**Topological Properties of Weak and Strong rationalizability**). *Take a compact type space* $\mathcal{T} = \left(T_p, T_d, \hat{\theta}(\cdot), \hat{\pi}_p(\cdot), \hat{\pi}_d(\cdot)\right)$ *and recall the definitions of* $WCR_i^k(\mathcal{T}) \subset \Sigma_i$ *and* $SCR_i^k(\mathcal{T}) \subset \Sigma_i$ *as the set of weak and strong rationalizable strategies for type space* $\mathcal{T}$. *Then:*

(1) *The sets* $WCR_i^k(\mathcal{T})$ *and* $SCR_i^k(\mathcal{T})$ *are non-empty, compact, Hausdorff and regular spaces, and satisfy* $WCR_i^k(\mathcal{T}) \subseteq SCR_i^k(\mathcal{T})$ *for all* $k \in \mathbb{N}, i \in \{p, d\}$

(2) *The rationalizable sets* $WCR_i^\infty(\mathcal{T}) \subseteq SCR_i^\infty(\mathcal{T})$ *are also non-empty, compact, Hausdorff and regular spaces for* $i \in \{p, d\}$

(3) *The set of all weak rationalizable strategies* $WCR_i^{\infty} \subset \Sigma_i$ *and strong rationalizable strategies* $SCR_i^{\infty} \subset \Sigma_i$ *are non-empty, compact, Hausdorff and regular spaces, and satisfy* $WCR_i^{\infty} \subset SCR_i^{\infty}$ *for* $i \in \{p, d\}$.

**Proof.** (1) follows directly from Propositions 3.5 and 3.6 in **?**, since the strategy space $\Sigma_i$ is compact (Proposition B.1) and the best response correspondences are u.h.c, non-empty compact valued (Proposition B.3). The restrictions on the rationalizable sets can also be mapped as restrictions on the type space, as shown by **?**. The fact that $WCR_i^{\infty}(\mathcal{T})$ and $SCR_i^{\infty}(\mathcal{T})$ are non-empty follows from the compactness and non-emptiness proved in follows from (2) and the generalization of Cantor's Theorem, which states that the intersection of a decreasing sequence of non-empty compact sets is non-empty (Theorem 26.9 in **?**). Since $WCR_i^k(\mathcal{T})$ and $SCR_i^k(\mathcal{T})$ are compact, they are also closed sets, which make $WCR_i^{\infty}(\mathcal{T})$ and $SR_i^{\infty}(\mathcal{T})$ closed. Because $\Sigma_i$ is a Hausdorff space, this also implies that $WCR_i^{\infty}(\mathcal{T})$ and $SCR_i^{\infty}(\mathcal{T})$ are also compact spaces (Theorem 26.2 **?**). Regularity follows from regularity of $\Sigma_i$, and therefore we have shown (2). For (3) we use the universal type space theorem B.1 to be able to write

$$WCR_i^* \equiv \bigcup_{\mathcal{T}:\mathcal{T} \text{ is a type space}} WCR_i^{\infty}(\mathcal{T}) = WCR_i^{\infty}(\mathcal{T}^*) \tag{B.21}$$

and

$$SCR_i^* \equiv \bigcup_{\mathcal{T}:\mathcal{T} \text{ is a type space}} SCR_i^{\infty}(\mathcal{T}) = SCR_i^{\infty}(\mathcal{T}^*) \tag{B.22}$$

and we use again this theorem to recall that the type space $\mathcal{T}^*$ consists of compact type spaces $T_i^*$ with continuous belief functions $\hat{\pi}_i^*$. Therefore we can apply the result in (3) for the particular case of $\mathcal{T} = \mathcal{T}^*$. □

## 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 $h^{\tau}$ *precedes* $h^s$ and write $h^{\tau} \prec h^s$.

**Proof of Lemma 5.1.** The first part is a consequence of Lemma (C.3) in Appendix C. For the second result, take a robust and strong rationalizable strategy $\sigma_d$ 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 **?**), and hence she should be also certain that by following the robust strategy $\sigma_d$ 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 $\sigma_d$ at $h$), reaching a contradiction $\square$

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

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

*Then, there exists a pair $(\sigma_d^\nu, \pi_d^\nu)$ that also $\theta-$rationalizes $h^\tau$ and is such that*

$$\sigma_d^\nu\left(h^s\right) = \hat{\sigma}_d\left(h^s \sim h^\tau\right), \pi_d^\nu\left(\cdot \mid h^s\right) = \hat{\pi}_d\left(\cdot \mid h^s \sim h^\tau\right) \tag{C.2}$$

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

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

$$\sigma_d^\nu\left(\tilde{h}^s\right) := \begin{cases} \sigma_d\left(\tilde{h}^s\right) & \text{if } s < \tau \text{ or } \tilde{h}^s = h^\tau \\ \sigma_\theta^*\left(\tilde{h}^s \sim \tilde{h}^\tau\right) & \text{if } s \geq \tau \text{ and } h^\tau \not\prec \tilde{h}^s \\ \hat{\sigma}_d\left(\tilde{h}^s \sim h^\tau\right) & \text{if } s \geq \tau \text{ and } h^\tau \prec \tilde{h}^s \end{cases} \tag{C.3}$$

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

$$\pi_d^k\left(A \mid \tilde{h}^s\right) := \begin{cases} \pi_d\left(A \mid \tilde{h}^s\right) & \text{if } s < \tau \\ \underline{\pi}_\theta\left(A \mid \tilde{h}^s \sim \tilde{h}^\tau\right) & \text{if } s \geq \tau \text{ and } h^\tau \not\prec \tilde{h}^s \\ \hat{\pi}_d\left(A \mid \tilde{h}^s \sim h^\tau\right) & \text{if } s \geq \tau \text{ and } h^\tau \preceq \tilde{h}^s \end{cases} \tag{C.4}$$

so the pair $(\sigma_d^\nu, \pi_d^\nu)$ coincides with $(\sigma_d, \pi_d)$ for any histories of length less than $\tau - 1$, and strategies also do it up to time $\tau$. 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 $h'^\tau$, then type $\theta$ believes that she will switch to the optimal strong rationalizable strategy from then on, to which the best response is $\sigma_\theta^*$ 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 $h^\tau$, and so the pair $(\sigma_d^\nu, \pi_d^\nu)$ is rationalizable. Then, to finish our proof, we need to show that it is consistent with $h^\tau$ 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.1. The other cases are shown in a similar fashion. $\square$

**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 $(\sigma_d, \pi_d)$ 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 $(\sigma_w, \pi_w)$ 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

$$W_\theta^{\hat{\pi}_z}\left(\hat{\sigma}_z \mid h^0\right) = w(z) \tag{C.5}$$

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 $\theta$ 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 $\sigma_0$ 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 $\sigma_z$ 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

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

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 $h^\tau$ 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 $h$ :

$$\pi(A \mid h) = \begin{cases} 1 & \text{if } h \succ (h^\tau, \hat{c}) \text{ and } \sigma_p^{FB} \in A \\ 1 & \text{if } h \not\succ (h^\tau, \hat{c}) \text{and } \underline{\sigma}_p \in A \\ 0 & \text{otherwise} \end{cases} \tag{C.6}$$

where $\sigma_p^{FB}$ 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

$$\hat{\sigma}_d(h) = \begin{cases} \left(\hat{c}, r^{spot}_{old}(\cdot \mid \hat{c})\right) & \text{if } h = h^\tau \\ \left(0, r^{spot}_{old}(\cdot \mid c = 0)\right) & \text{if } h \succ (h^\tau, \hat{c}) \\ (\infty, r_g(\cdot)) & \text{if } h \not\succ (h^\tau, \hat{c}) \end{cases} \tag{C.7}$$

where $r^{spot}_\theta(z \mid c) = \underset{r\in(0,1)}{\operatorname{argmax}} (U_\theta - c)\, r$ 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

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

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 $(h^0, c_0)$ 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 $S_{old,k-1} = \max_{\tilde{r}\in\{0,1\}} (U_{old,k-1} - c_{k-1})\,\tilde{r} - (U_{old,k-1} - c_{k-1})\, r_{k-1} \leq S^{\max}_{old}$. This concludes the proof. □

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

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

$$\mathbb{E}_{z_\tau}\left\{r^{\sigma_d}(h^\tau, z_\tau)\left[U_p(z_\tau) - c^{\sigma_d}(h^\tau)\right]\right\} \geq 0 \tag{C.8}$$

**Proof.** The proof will follow from 2 steps:

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

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

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

Suppose $\underline{\mathbb{W}}_{new} < \beta \mathbb{E}_{z_\tau}[\max(0, U_p(z_\tau))]$. Then the following strategy would be strongly rationalizable: prohibit $r = 1$ at $h^\tau$ 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 $\underline{\mathbb{W}}_{new} \geq \beta \mathbb{E}_{z_\tau}[\max(0, U_p(z_\tau))]$ from the fact that $\underline{\mathbb{W}}_{new}$ is the maximum utility over robust strategies, and thus reaching a contradiction.

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

For any rationalizable strategy $\sigma_d$ we have

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

This also implies then that

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

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

using Step 1 in the last inequality. □

**Lemma C.4.** *There exist a non-zero measure set $\tilde{S} \subset [0, S^{\max}_{old}]$ such that $\frac{\partial \mathcal{W}_n}{\partial s}(s) > 0$ for all $s \in \tilde{S}$ and hence $\mathcal{V}(s, c^*(s)) = \underline{u}_p$ for all $s \in \tilde{S}$*

**Proof.** Notice first that for all $s$ we have $\mathbb{E}[r(z \mid s)] = \Pr[r(z \mid s) = 1] > 0$ . This is because if it wasn't, then utility of this policy at $s$ would give utility 0, whereas we could have chosen $c = \mathbf{c}(s)$ and get positive utility, together with positive probability of playing $r = 1$ . Suppose not, so that $\frac{\partial \mathcal{W}_n}{\partial s} = 0$ for all $s$ for which the derivative exists (which are almost everywhere). Pick a $s$ such that the constraint is not binding: i.e. $\underline{\mathcal{V}}(s, c^*(s)) > \underline{u}_p$ (which must necesarily exist given the characterization of the minimum cost function $\mathbf{c}(s)$. Take the optimal policy at that state, which is $r(z) = r^*(z)$, $s(z) = s^*(z)$ and $c = c^*$. We will construct a local feasible deviation: keep the same policy function $r(z)$ and only reduce the commitment cost to $\tilde{c} = c - \epsilon$, which implies that the next period sacrifice would now be

$$s(z, \epsilon) = \max_{\tilde{r} \in [0,1]} (U_{old}(z) - c + \epsilon)\tilde{r} - (U_{old}(z) - c + \epsilon) r(z)$$

The utility of the right hand side maximized problem was □

$$(1-\beta)\mathbb{E}[(U_{old}(z) - c) r(z)] + \beta \mathbb{E}\mathcal{W}[s(z)]$$

and with the deviation is

$$(1-\beta)\mathbb{E}[(U_{old} - c + \epsilon) r(z)] + \beta \mathbb{E}\mathcal{W}[s(z, \epsilon)]$$

we will show that it is a stricly increasing deviation:

$$(1-\beta)\mathbb{E}[(U_{old}(z) - c) r(z)] + \beta \mathbb{E}\mathcal{W}[s(z)] < (1-\beta)\mathbb{E}[(U_{old}(z) - c + \epsilon) r(z)] + \beta \mathbb{E}\mathcal{W}[s(z, \epsilon)] \iff$$

$$(1-\beta)\Pr\left[r(z)=1\right]\epsilon+\beta\mathbb{E}\left\{\mathcal{W}\left[s(z,\epsilon)\right]-\mathcal{W}\left[s(z)\right]\right\}>0 \tag{C.9}$$

Because $\mathcal{W}$ is differentiable almost everywhere, then for almost all $z\in Z$ we can make the differential approximation around $\epsilon=0$:

$$\mathcal{W}\left[s(z,\epsilon)\right]-\mathcal{W}\left[s(z)\right]\approx\frac{\partial\mathcal{W}}{\partial s}\left[s(z)\right]\left[\frac{\partial s(z,\epsilon)}{\partial\epsilon}\mid_{\epsilon=0}\right]\epsilon$$

and using the envelope theorem

$$\frac{\partial s(z,\epsilon)}{\partial\epsilon}=r_{old}^{spot}(z\mid c-\epsilon)-r(z)$$

so that evaluating it at $\epsilon=0$ we simplify this condition as

$$\mathcal{W}\left[s(z,\epsilon)\right]-\mathcal{W}\left[s(z)\right]\approx\frac{\partial\mathcal{W}}{\partial s}\left[s(z)\right]\left[r_{old}^{spot}(z)-r(z)\right]\epsilon$$

then for small enough $\epsilon$ condition C.9 is satisfied if and only if

$$(1-\beta)\int r(z)f(z)\,dz+\beta\int\left[r_{old}^{spot}(z)-r(z)\right]\frac{\partial\mathcal{W}}{\partial s}\left[s(z)\right]f(z)\,dz>0 \tag{C.10}$$

The assumption $\Pr(r(z)=1)>0$ implies that condition C.10 will necesarily hold if we can show

$$\int\left[r_{old}^{spot}(z)-r(z)\right]\frac{\partial\mathcal{W}}{\partial s}\left[s(z)\right]f(z)\,dz>0$$

Because the only potential mass-point for the implied distribution for $s'(z)$ is at $s=0$ (when there is no sacrifice, sacrifice is zero, and this can happen if $r_{\theta}^{spot}\neq r$ has positive probability) and we already know that $\mathcal{W}$ is locally constant in the interval $[0,\hat{s}]$ we also have that $\frac{\partial\mathcal{W}}{\partial s}(0)=0$. Therefore,

$$\frac{\partial\mathcal{W}}{\partial s}\left[s(z)\right]=0\text{ a.e in }z\in Z$$

which given the absolute continuity of $Z$ delivers the desired result.

**Lemma C.5.** *If $\underline{\mathcal{V}}(s,c^*(s))>\underline{u}_p$ for some $s$, then it also holds for all $s'\in(s,S_o^{\max})$*

**Proof.** It follows by inspection of the first order conditions of the lagrangian problem, since $s$ only enters the conditions through this constraint, which implies that if it is non-binding at $s$ it is also non-binding at $s'>s$, since increasing the sacrifice only relaxes this constraint, which was not binding in the optimum. □

**Corollary.** *There exist $\underline{s}>\hat{s}$ such that for all $s\leq\underline{s}$ we have $c^*(s)=\mathbf{c}(s)$ and for $s\geq\hat{s}$ we have $c^*(s)=\mathbf{c}(\hat{s})$*

**Lemma C.6.** *Under the increasing misalignment assumption 1, given $\epsilon,\delta>0$, the functions:*

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

*and*

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

*satisfies* $\frac{\partial G}{\partial a}\cdot\frac{\partial H}{\partial a} \leq 0$. *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 Leibnitz rule:

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

which is negative given our assumption. Moreover,

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

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

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

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

as we wanted to show. □

**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,

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

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 $\underline{\mathcal{V}}_{new}(S_{new,\tau-1},c_\tau) \geq 0 > \underline{u}_p$. Therefore, the implementation restriction

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

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

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

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

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

using the definitions in Lemma C.6, and hence it is decreasing in $c$, as we wanted to show. ☐

**Lemma C.7.** *$T$ as defined in 5.25 is a contraction mapping with modulus $\beta$*

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

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

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

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

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

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

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

which violates the definition of $c_0^*$

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

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

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

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

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

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

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

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

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 $h^\tau$ 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 $\delta$ 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:

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

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 $n\to\infty$ :

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

reaching a contradiction. □

## 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*

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

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

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

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

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

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

(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 (**?**)

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

where $\gamma \geq 0$ 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 $n(z) = \min\{0, c - U_p\}$. Then, given $\gamma$ we solve for $r(z \mid \gamma)$ and $n(z \mid \gamma)$, and we solve for $\gamma$ 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 $\gamma$ as a function of s, showing (3). □

Results are better explained using Figures (9) , (10) and (1) below.

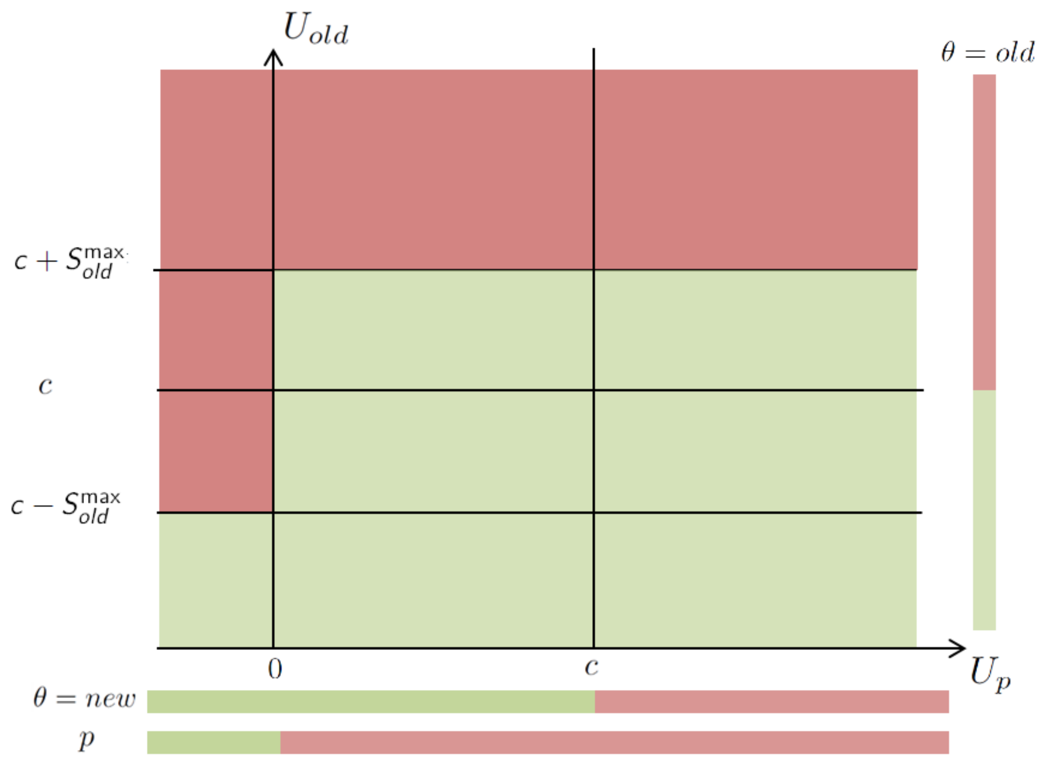


FIGURE 9. Rationalizable Evil Agent strategy, with $s \leq \hat{s}$

See that for $U_{old}$ above $c + S_{old}^{\max}$ and below $c - S_{old}^{\max}$ the unique rationalizable actions for $b$ are $\hat{r}(z) = 1$ (red) and $\hat{r}(z) = 0$ (green) respectively, as we have seen before. When $U_{old} \in (c - S_{old}^{\max}, c + S_{old}^{\max})$, both $r = 0$ and $r = 1$ are rationalizable for any $z$ in this region, by appropriately choosing the expected continuation payoffs. Therefore, the worst strategy that $p$ could expect would be one of a *contrarian*: whenever $p$ wants the green button to be played ($U_p < 0$), then the old type would play the opposite action. We can draw an obvious parallelism to the "evil agent" in the robustness literature of **?**, with the restriction that instead of a pure evil agent, the *rationalizable evil agent*, that is only contrarian at states in which the utility of doing her most desirable action is not too high.

See that being a rationalizable contrarian is costly for $\theta = old$, since there are regions in which both $p's$ and $\theta = old$ most desired action coincide, as we see in the next figure (regions stressed in darker colors)

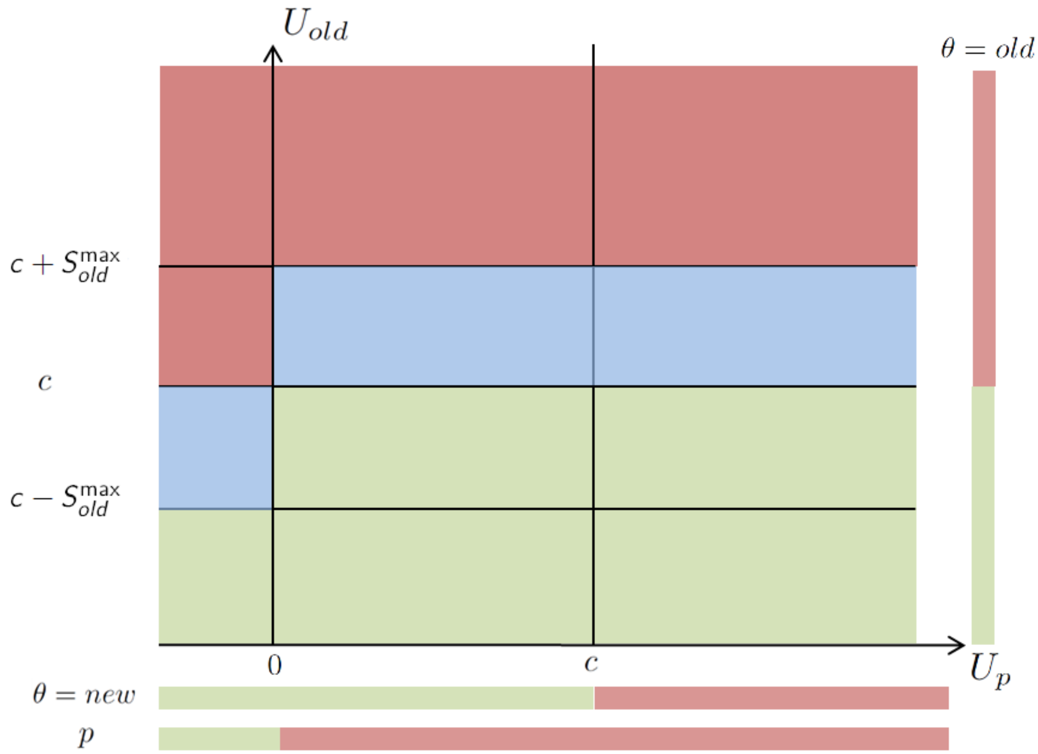


FIGURE 10. Self-contrarian regions for rationalizable evil agent

Then, when sacrifice is high enough, the disutility generated by the dark regions would not be consistent with the observed behavior. Therefore, to satisfy the "promise keeping" constraint, we must allow the "rationalizable evil agent" not to be fully contrarian, and play her desired action in some states, as we see in the figure below.

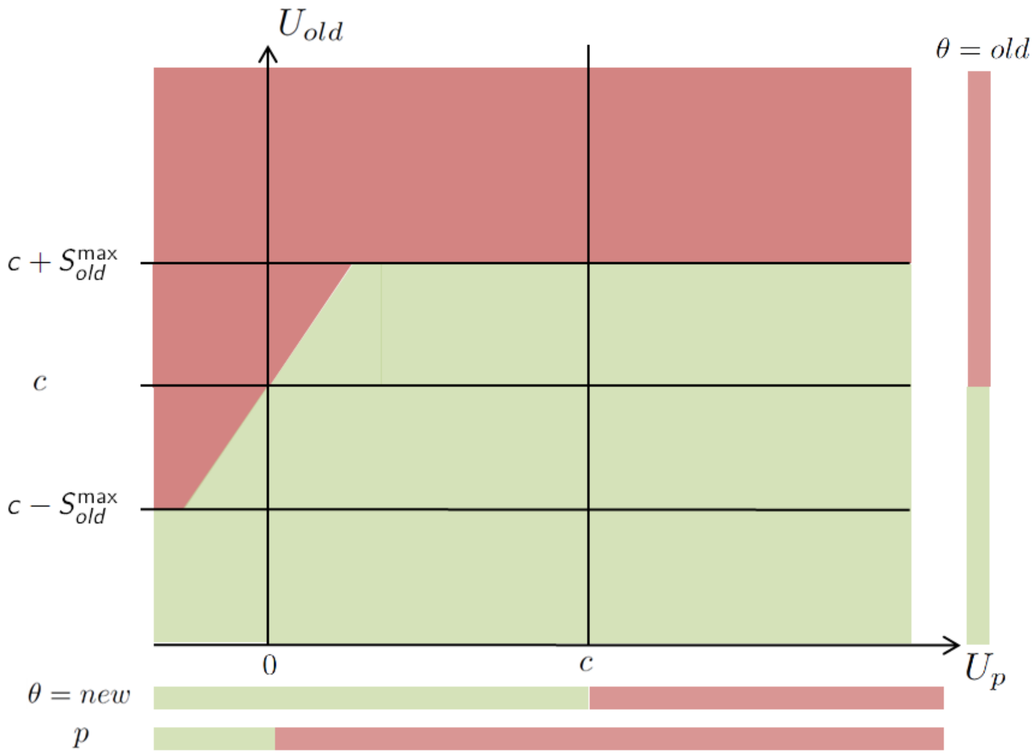


TABLE 1. Rationalizable evil agent strategy, with $s > \hat{s}$

Finally, is easy to see that as the promise keeping constraint becomes more and more binding, the worst type's policy $\hat{r}(.)$ resembles more and more the spot optimum policy $r_{old}^{spot}(.)$. Then, $\mathbf{c}(s) > \overline{c}$ and it approximates it as the promise keeping constraint becomes more binding.

## REFERENCES