Codes of Conduct and Bad Reputation *

Juan I. Block †

March, 2013

Abstract

We study bad reputation games from the perspective of self-referentiality. In self-referential games players have the possibility of understanding opponents' intentions, and this can mitigate the problem of bad reputation. We characterize the probabilities of the Stackelberg type required to overcome a bad reputation problem when there is a possibility that intentions can be observed directly. The complementarity between direct and indirect observation of opponents' intentions is shown to be qualitatively different from games where all agents are long-lived.

Keywords: Reputation, Self-referential Games, Stackelberg
JEL Classification: C72,C73,D82

*I am grateful to David Levine for guidance and encouragement. I also thank John Nachbar and Paulo Natenzon for helpful discussions.

†Department of Economics, Washington University in St. Louis, St. Louis, MO 63130-4899, USA. Email: juanblock@wustl.edu

1 Introduction

Most models of reputation establish that uncertainty about the long-run player's type implies a beneficial reputation effect as long as long-run player can exploit some commitment power. By way of contrast, Ely and Välimäki (2003) propose a novel example where reputation may be bad. With a bad commitment type, they find that in equilibrium the entire surplus is lost. Along the same line, an example which captures all features of their model is mitral valve surgeries. There are two types of surgeries, valve repair and valve replacement. Both surgical procedures are equally successful when appropriately conducted. The right type of surgery is known only by the surgeon. Think of the case where there are two broad categories: good and bad physicians. The bad physician always does valve replacement but the good one performs the necessary surgery. Since this example is analogous to theirs, the model predicts none of physicians would perform a surgery. In reality, however, we still observe the coexistence of bad and good physicians involve in long-run relationships with patients.[1]

Humans are rarely perfect liars. They reveal – even unintentionally – their states of mind either through micro expressions (facial gestures, body posture, etc.) or tools such as pen-drumming. In an evolutionary setting, Levine and Pesendorfer (2007) examine self-referential strategies which have the ability to recognize each other in the context of two player symmetric games. Intuitively, players have the chance of discerning whether opponents conform to a rule of behavior. To generalize this idea, Block and Levine (2012) define self-referential games in which players are able to understand opponents' intentions about chosen strategies by receiving informative signals. The self-referential nature of these games is characterized by the fact that players choose how they will play the game depending on signals they observe, and the choice of such strategies indeed determines the likelihood of those signals. (See also Kalai, Kalai, Lehrer, and Samet (2010).[2])

In this paper we analyze bad reputation games from the perspective of self-referentiality in order to study the connection between the two ideas mentioned at the beginning. We show how the possibility of observing opponents' intentions restricts the bad reputation effect, and we identify conditions under which such restriction applies for both weak and strong sources of information about intentions.

In our model, myopic players play against a long-lived opponent whose life span is stochastic. We show that the possibility of fathoming other players' plan of strategies and the occasional renewal of the long-run player mitigate the bad reputation effect. That is, it is generally better to have myopic players with permanent uncertainty about types for the long-run player because this weakens how informative public histories full of bad signals are. Moreover, we are interested in showing how information about opponents' intentions may complement re-

[1] In the US more than 40,000 mitral valve operations are performed every year.

[2] They study commitment devices which work very similarly to code-of-conduct in two-player games, however, they consider very precise signals.

newal. We characterize conditions on the self-referential game and on the relative likelihood of commitment types to assure that the bad reputation effect will not arise.

Bad reputation games exhibit distinctive features from the bulk of the literature starting with the classic works of Kreps and Wilson (1982) and Milgrom and Roberts (1982). We model bad reputation games borrowing the set-up developed by Ely, Fudenberg, and Levine (2008) who give the first characterization of the limit between good and bad reputation identifying properties of this class of games. A long-lived opponent plays against a sequence of different myopic players. Participation of short-run players takes place if friendly actions are likely to be played. Since long-run player's actions are imperfectly observed, friendly actions may generate bad signals that can be interpreted as evidence of unfriendly actions. In addition, there exist temptation actions which result in good signals more often but they may be unfriendly. A numerous amount of bad signals points to an unfriendly type, hence a patient normal long-run player eventually chooses temptation actions. In contrast to Ely et al. (2008) we focus on costly temptation actions, that is, it is costly to play actions that are more likely to generate good signals. While they find general conditions under which reputation is bad, their characterization does not include all possible commitment type priors. In contrast, we identify the whole set of priors where bad reputation is overcome.

Our main departure from their setting is that the long-run player might be replaced every period. Short-run players remain ignorant about the long-lived player's type since they can observe neither the renewal nor the type. For instance, during a valve operation surgeons may shift every other couple of hours but the patient only knows his primary care surgeon. The possibility of switching types has been previously studied in reputation models but not in the class of bad reputation games. Renewal has a dual effect. First, since the long-run player is likely to be gone by tomorrow he has less incentives to play friendly and harvests the fruits of reputation. Second short-run players are unaware whether a replacement has occurred, this implies that a very unlucky history of bad signals weighs less frequently in the updating of prior probabilities. The latter is prevalent in reputation games, as in Ekmekci, Gossner, and Wilson (2012). In Mailath and Samuelson (2001), the long-run player does not lean on a favorable history because players are uncertain whether he might have been replaced by a bad type. In contrast, in our model, the possibility of being renewed partially "cleans up" a history plenty of bad signals. The reason why renewal may restore the short-run players' beliefs about the long-run player's type is similar to that in Liu and Skrzypacz (2009), consider short-run players having bounded recall. (See also Liu (2011).) Most of the literature on switching types is concerned about equilibrium dynamics (see for example, Holmström (1999), Phelan (2006) and Wiseman (2008)).

We first consider perfect information about opponents' intentions, and show that the pre-commitment friendly action outcome can be induced by a self-referential Nash equilibrium. The self-referential game might have distinctive features compared to the base game. This case may look like rather extreme but it helps us understand how players behave when they have the chance of agreeing upon behavior without noise. In this setup, there is no role for reputation building in the self-referential game and it eliminates the incentives to regain opinion of short-run players when the long-lived is likely to be tagged as a bad type. We should also stress that the result is independent of either the level of patience and the probability of renewal of the long-lived player. Basically, the long-lived player is locked in the stage game in terms of incentives.

When this source of information is weak, the reputation might be perverse but under more restrictive conditions. In this case, our main result is that self-referentiality strongly complements the probability of renewal of the long-run player. We show that there exists a self-referential equilibrium in which the long-run player obtains his pre-commitment friendly action. The construction of the equilibrium in the self-referential game is more demanding and it requires to balance players' incentives to deviate that were not present in the base game. Having imprecise signals makes the self-referential game share at least all characteristics of the base game. Again, this illustrates that when dealing with self-referential games there exists a trade-off between the agreement about players' behavior and the cost of backing up this consensus.

2 Model

2.1 Basic Setup

A long-run player, player 1, faces a sequence of different short-run players, player 2. Starting at period $t = 1$, each period players choose simultaneously actions from their finite action space, and the stage-game is played repeatedly for finite $T(< \infty)$ periods. Let the action space of the long-run player be A with element a and of short-lived player be B with element b.

The long-run player discounts future at the discount factor $\delta \in (0, 1)$. There are different types θ for the long-run player that are described by the finite set Θ and are private information. The set of types of the long-run player consists of type $0 \in \Theta$ which is defined as normal type, and a commitment type $\theta(a)$ that always plays action $a \in A$. In the repeated game, the normal type maximizes the discounted sum of expected payoffs whereas commitment types have the trivial strategy of repeating the stage action. In the stage-game, the payoff functions of the short-run players and of the normal long-run player are denoted by $u_i : A \times B \to \mathbb{R}$ for all $i = 1, 2$. Let the common prior distribution over types at time 0 be the probability measure $\mu_0 \in \Delta(\Theta)$ with full support, thus, $\mu_0(\theta) > 0$ for all $\theta \in \Theta$.

There exists the possibility of renewal of the long-run player, and it is exponentially distributed with exogenous probability $\lambda \in (0, 1)$. Every period, the long-run player might be replaced with probability λ by either a normal type or a commitment type. The new type θ is drawn according to the probability measure μ_0. Once the long-run player has been replaced, he will never enter the game again. Both the renewal and the type of the upcoming long-run player are unknown to short-run players. Our interpretation of renewal goes beyond examples like retirement, and we do not exclusively associate renewal with the mere fact of a physical replacement.[3] We define the effective discount factor $\tilde{\delta} = (1-\lambda)\delta$.

Here we follow the setup developed by Ely et al. (2008) when characterizing bad reputation games. At the end of the stage game players observe a public signal from finite space Y. Given an action profile (a, b) in the stage game, the probability of the public signal $y \in Y$ is characterized by the probability distribution $\rho(\cdot|a, b) \in \Delta(Y)$. The history of realizations of these public signals is observed by all players. We assume that short-run players are restricted to observe public signals, and they are not able to observe neither the past play of the long-run player nor of the other short-run players. Let $H^t = Y^t$ be the set of all t-length public history whose element $h^t = (y_1, \ldots, y_t)$ is the public history through the end of period t, and by $h^0 = \emptyset$ the null history. We write $H = \bigcup_{t=1}^{T} Y^t$ for the set of public histories. The private history of the long-run player is represented by $h_1^t = (a_1, \ldots, a_{t-1})$[4] which belongs to the set of his t-length histories written as $H_1^t = A^t$, and let $H_1 = \bigcup_{t=1}^{T} H_1^t$ be the set of private histories for the long-run player.

The behavior strategy of the long-run player is a sequence of maps from public history, private history and type to the set of probability distributions on the set of actions A. That is, player 1's strategy is a mapping $\sigma_1 : H \times H_1 \times \Theta \to \mathcal{A} := \Delta(A)$. A strategy for the short-run player is a sequence of maps from public histories to probability distribution on the set B, $\sigma_2(h^t) \in \Delta(B) := \mathcal{B}$. Then, a short-run action $\beta \in \mathcal{B}$ is a Nash response to action $\alpha \in \mathcal{A}$ if $u_2(\alpha, \beta) \geq u_2(\alpha, \hat{\beta})$ for all $\hat{\beta} \in \mathcal{B}$. Let $\mathbf{B}$ be the best-response correspondence and let $\mathbf{B}(\alpha)$ be the set of short-run Nash responses to α.

We focus on bad reputation games that are a subclass of participation games. In participation games short-run players have the option not to participate in the game. If they do not participate in the game, they choose exit actions defined as pure actions $e \in E \subseteq B$. Alternatively, entry actions belong to $B - E$. Let $Y^E \subseteq Y$ be a set of public signals which are called exit signals. For each exit action $e \in E$, the probability distribution over public signals satisfy these two conditions: $\rho(y|a, e) = \rho(y|e)$ for all $a \in A, y \in Y$, and $\rho(Y^E|e) = 1$. That is, given an exit action the distribution of public signals is independent of the long-run player's actions, moreover, only exit signals are possible. In addition, when entry actions are chosen none of the exit signals can be observed. Formally if short-run player chooses any entry action $b \notin E$, then $\rho(Y^E|a, b) = 0$ for all $a \in A$. Note that exit signals are very informative since these are driven by exiting short-run players. A game is a participation game if the exit action set is non-empty $E \neq \emptyset$ and there exists some action $\alpha \in \mathcal{A}$ with $\mathbf{B}(\alpha) \cap E \neq \emptyset$.

We use $\beta\{E\}$ to denote the probability assigned to the set of exit actions, $E \subseteq B$, by Nash

[3] For example, many German car companies have in their service centers the car remotely connected to the headquarter. The device inside the car interchanges information with the mechanics/computer in the headquarter. More importantly, as a result of this interaction mechanics find out where the problem is and what kinds of repairs may be needed. Again, motorists do not know exactly about modifications in diagnostic algorithms or mechanics in headquarters, though the brand does not change.

[4] Possibly it might also include the actions of the short-run players in previous interactions.

response action $\beta \in \mathbf{B}(\alpha)$. A non-empty finite set of pure actions $F \subseteq A$ for long-run player is friendly if there is a number $\gamma > 0$ such that, for all player 1's actions $\alpha \in \mathcal{A}$ if the short-run player strategy is a Nash response $\beta \in \mathbf{B}(\alpha)$ and $\beta\{E\} < 1$ then the long-run player assigns positive probability $\alpha(f) \geq \gamma$ for some friendly action $f \in F$. On the other hand, an unfriendly set N corresponding to friendly action set F is any non-empty subset of $A \setminus F$. This definition says that Nash response of short-run players puts positive probability to non-exit actions when friendly actions are played with high probability. Any pure non-friendly action makes short-run players choose an exit action. Commitment types associated with friendly set F is denoted by the subset $\Theta(F) \subseteq \Theta$ and are called friendly commitment types. Analogously, let the subset $\Theta(N) \subseteq \Theta$ be the unfriendly commitment types corresponding to actions in unfriendly set N. Because of the definition of the sets F and N, we have that these two corresponding commitment sets are disjoint $\Theta(N) \cap \Theta(F) = \emptyset$. Let a mixed action $\alpha \in \mathcal{A}$ for the long-run player be enforceable if for every other action $\tilde{\alpha} \in \mathcal{A}$ for all short-run player's action $\beta \in \mathcal{B}$ with $\beta \in \mathbf{B}(\alpha)$ and $\beta\{E\} < 1$, such that $u_1(\tilde{\alpha}, \beta) > u_1(\alpha, \beta)$ then $\rho(\cdot|\tilde{\alpha}, \beta) \neq \rho(\cdot|\alpha, \beta)$.[5]

A set of signals $\bar{Y} \subseteq Y$ is evidence for a set of actions $N \subseteq A$ if N is non-empty and $\rho(\bar{y}|n, b) > \rho(\bar{y}|a, b)$ for all $b \notin E, \bar{y} \in \bar{Y}, n \in N$ and $a \notin N$. In words, each action in the set N gives rise to a higher probability for every signal in $\bar{Y}$ than any action not in N. An action $a \in A$ is vulnerable to temptation relative to a set of signals $\bar{Y}$ if there exist numbers $\underline{\rho}, \tilde{\rho} > 0$ and an action $d \in A$ such that (i) If $b \notin E, \bar{y} \in \bar{Y}$, then $\rho(\bar{y}|d, b) \geq \rho(\bar{y}|a, b) - \underline{\rho}$; (ii) If $b \notin E$ and $y \notin \bar{Y} \cup Y^E$ then $\rho(y|d, b) \geq (1 + \tilde{\rho})\rho(y|a, b)$; and (iii) For all $b \in E, u_1(d, b) \geq u_1(a, b)$. The action d is called a temptation whose temptation bounds are the largest possible $\underline{\rho}, \tilde{\rho}$ satisfying (i) and (ii) for action a. It says that by playing the temptation action d the long-run player reduces the probability of bad signals by at least $\underline{\rho}$ and increases of all signals but in the set $\bar{Y} \cup Y^E$ by factor $1 + \tilde{\rho}$.

We say an action d is a *costly temptation* if it is a temptation and $u_1(a, b) - u_1(d, b) \geq c$ with $c > 0$ for all $b \in B - E$. When we examine self-referential games with imperfect detection, we are interested in bad reputation games with costly temptation. A participation game has exit minmax if

$$\max_{b \in E \cap \mathbf{B}} \max_{a \in A} u_1(a, b) = \min_{\beta \in \mathbf{B}} \max_{a \in A} u_1(a, \beta)$$

A participation game is a bad reputation game if it has exit minmax, and there is a friendly set F and corresponding non-empty unfriendly set N and a set of signals $\bar{Y}$ that are evidence for N, such that every enforceable friendly action $f \in F$ is vulnerable to temptation relative to the set of signals $\bar{Y}$. All the signals in $\bar{Y}$ are called bad signals. A bad reputation game with costly temptation is a bad reputation game where the set of temptations is costly.

Let the constant κ be interpreted as a measure of how revealing the evidence is, and defined by $\kappa = \min_{n \in N, a \notin N, \beta\{E\}<1, \bar{y} \in \bar{Y}} \frac{\rho(\bar{y}|n,\beta)}{\rho(\bar{y}|a,\beta)}$. Note that k is finite with $\kappa > 1$. Let $\varphi > 0$ be the minimum of the temptation bounds $\underline{\rho}$ and finally the signal lag given by $\eta = -\log(\gamma\varphi) \backslash \log \kappa$.

[5] The idea of identification of actions was proposed by Fudenberg, Levine, and Maskin (1994).

2.2 Self-referential Game

We consider a generic base game $\Gamma = \{(S_i, u_i)_{i \in \mathcal{I}}, \mathcal{I}\}$ where there is a finite set of players $\mathcal{I} = \{1, \ldots, N\}$. Each player i selects a strategy from a finite set S_i with strategy profile $s \in S \equiv \prod_i S_i$. We allow for mixed strategies, but finitely many of them (e.g. spinning the roulette wheel). Let the payoff function be $u_i : S \to \mathbb{R}$ for each player i.

Next, we define the *self-referential game*. The set of players is the same as in the base game, $\mathcal{I}$. We assume that there is a finite set of private signals $Z_i \ni z_i$ for each player i, the strategy of player i is defined as *code of conduct* denoted r^i which is an $|\mathcal{I}| \times 1$ vector whose jth element corresponds to the mapping from the set of player j's private signals to player j's strategies in the base game, namely, $r^i_j : Z_j \to S_j$. Given a profile of codes of conduct $r \in \mathcal{R} \equiv \prod_i R_0$, the joint probability distribution of private signals is given by $\pi(z|r)$ with $z \in Z \equiv \prod_i Z_i$. The space of codes of conduct takes the form $R_0 := \{r^i | r^i_j : Z_j \to S_j$ for all players j and any player $i\}$. It requires each player to decide how he will play the base game after receiving private signals and how all the other players will play conditional on their signals. This is the self-referential nature of this class of games. We say a space of code-of-conduct R_0 is *complete* if all profiles of map $r^i_j : Z_j \to S_j$ is represented in R_0–we use complete code-of-conduct throughout the paper. The profile of codes of conduct $\hat{r}$ is a Nash equilibrium of the self-referential game if for all $i \in \mathcal{I}$, we have that

$$\hat{r}^i \in \underset{\tilde{r}^i}{\operatorname{argmax}} \sum_{z \in Z} u_i(r^1_1(z_1), \ldots, r^N_N(z_N))\pi(z|r)$$

The timing of the self-referential game is as follows: before observing any signal and playing the base game, all players simultaneously choose codes of conduct. Given this choice, a profile of private signals is drawn from the probability distribution $\pi(z|r)$. After observing private signals players execute codes of conduct.

3 Bad Reputation and Detection

3.1 Perfect Identification

We start by examining bad reputation games when players have perfectly revealing signals about deviations from codes of conduct in the self-referential version of the game. We show that we can sustain "good" equilibria regardless of the long-run player's patience and how frequently he leaves the game.

Let us define more precisely the notion of perfect detection. We say a self-referential game permits detection if for each player $i \in \mathcal{I}$ there exists some player $j \in \mathcal{I} \setminus \{i\}$ with non-empty subset of private signals $\bar{Z}_j \subset Z_j$ such that for every profile of codes of conduct $r \in \mathcal{R}$ and any code-of-conduct $\tilde{r}^i \in R_0$ with $\tilde{r}^i \neq r^i$, we have $\pi_j(\bar{Z}_j|\tilde{r}^i, r^{-i}) = 1$ and $\pi_j(\bar{Z}_j|r) = 0$. This definition says that if any player deviates from the code, there always exists some other player would point at this deviation. Given that our base game is of bad reputation kind we assume that all short-run players obtain the same private signals about the long-lived player's intentions in turn.[6] Let $f^* \in F$ be the precommitment friendly action of the long-lived player.

Theorem 1 *Given any finite bad reputation game with probability of renewal λ. If the self-referential game permits detection, there exists a code-of-conduct profile $r \in \mathcal{R}$ such that in the self-referential Nash equilibrium the normal long-run player gets a normalized discounted payoff of $u_1(f^*, \beta)$, where short-run players participate in the game, i.e. $\beta \in \mathbf{B}(f^*), \beta \notin E$.*

All proofs are relegated to the Appendix. This result illustrates that the self-referential game with perfect detection may have different features compared with the base game, and to prove it we use ideas from Levine and Pesendorfer (2007).

Observe that in this environment there is no room for reputation building. In bad reputation games short-run players care about whether his opponent will play a friendly or unfriendly action, had he entered the game. Thus they use public histories to make inference about long-lived player's forthcoming actions. Here, however, histories are irrelevant to short-run players' behavior because self-referential private signals provide precise information about what players would do in the current stage game. Consequently, the short-run player chooses his action conditional on what the long-run player will be doing irrespective of the history. The short-run player trusts the long-run player because he knows wether the long-lived opponent adheres to the code-of-conduct or not.

In this case, the long-run player would not need to regain the short-run players' faith after a history of bad signals but he might find profitable not to play a friendly action. The long-run player will not pursue this action since any deviation from the code-of-conduct will be punished immediately and thereafter.

3.2 Identification with Noise

We turn to the analysis of the situation in which players cannot perfectly identify deviations from the code-of-conduct. As we saw, when players detect deviations with certainty long-lived player lacks of reputation concerns in the self-referential game. Then, we ask to what extent even in the case of imperfect identification a small probability of detecting deviations from the code of conduct would still mitigate the bad reputation effect.

In this section, we consider the case where the private signals induced by the self-referential game are very noisy. The notion of detection we use says that evidence of deviation is modelled by the likelihood of observing a subset of private signals. Formally, we say a self-referential game E, D permits detection with constants satisfying $0 \leq E, D \leq 1$ and $E + D \leq 1$ if for every player $i \in \mathcal{I}$ there exists some player $j \in \mathcal{I} \setminus \{i\}$ and a nonempty subset of private signals $\bar{Z}_j \subset Z_j$, such that for any profile of codes of conduct $r \in \mathcal{R}$, any signal $\bar{z}_j \in \bar{Z}_j$

[6] This keeps results clear since players do not infer anything about other short-run players' signals, moreover, it turns out there are no qualitative changes.

and any code-of-conduct $\tilde{r}^i \in R_0$ where $\tilde{r}^i \neq r^i$ we have $\pi_j(\bar{z}_j|\tilde{r}^i, r^{-i}) - \pi_j(\bar{z}_j|r) \geq D$ and $\pi_j(\bar{z}_j|r) \leq E$.[7] In words, D measures the probability of detection if any player i deviates, and E can be interpreted as the probability of accusing someone who is being honest. Observe that E, D permits detection implies that the same signal may have different interpretations but its likelihood depends on the choice of codes of conduct.

It is necessary to impose uniformity to avoid the possibility of a relatively too high prior probability of unfriendly types. We say a bad reputation game with friendly set F and unfriendly set N has *uniformly friendly commitment size* ψ, χ with $\chi > 0$ if the prior probability of friendly and unfriendly types $\mu_0[\Theta(N)], \mu_0[\Theta(F)]$ satisfy

$$\mu_0[\Theta(N)] \leq \psi(1 - \mu_0[\Theta(F)] + \mu_0[\Theta(F)]^{\frac{1+\eta}{\eta}}) - \chi$$

The constant ψ reflects the uniformity of friendly types $\mu_0[\Theta(F)]$ relative to unfriendly types $\mu_0[\Theta(N)]$. Players use their private signals to update their prior probabilities about commitment types to a limited extent, therefore a relatively high likelihood of friendly commitment types is necessary to guarantee participation.

In the next result we show that reputation would have a negative effect in arbitrarily long discounted finitely repeated games when the long-run player's probability of renewal is big enough and his discount factor tends to 1, and the self-referential version of the game provides certain detection technology. We formally state the main result of this section:

Theorem 2 *Suppose a finite bad reputation game with uniformly friendly commitment size $1 - \gamma, \chi$ for some $\chi > 0$ and with costly temptation. If the self-referential game E, D permits detection with sufficiently high $D > 0$, for precommitment friendly action $f^* \in F$ and discount factor $\delta \to 1$ we can find a threshold $\tilde{\lambda}$ such that for all probabilities of renewal $\lambda \in (\tilde{\lambda}, 1)$ there exists a profile of codes of conduct $r \in \mathcal{R}$ such that in the self-referential Nash equilibrium the normal long-run player gets approximately a discounted average payoff of $u_1(f^*, \beta)$ with $\beta \in \mathbf{B}(f^*)$ and β is not an exit action.*

Short-run players cannot only rely on the code-of-conduct in order to know whether friendly actions are likely to be played. This implies that short-lived agents combine public information with private self-referential signals to update their beliefs about long-run player's type. When updating beliefs they take into account both the public history and the chance of facing a new long-lived opponent jointly with the information about intentions. Because with this detection technology histories are relevant in the self-referential game, reputation can be potentially perverse.

[7] Block and Levine (2012) use a stronger version of E, D permits detection because in that setting there exists the issue of mutual accusation – an ambiguity that arises, for instance, with many long-run players. The idea is that if two players point at each other we do not know who is guilty unless we strengthen the notion of permit detection. We have a long-run player facing a sequence of short-run players so this stronger definition is not needed.

Since we are interested in arbitrary long horizon T, we may find a probability of renewal such that for any t-length history of just bad public signals short-run players participate in the game as long as they perceive intentions of playing a friendly action from the long-lived opponent. This suffices to make the long-lived opponent elude the play of a temptation action. Notice that information about long-lived agent's intentions is relatively scarce to public histories. While the games evolves, short-run players put large weight on public signals history. Thus, the chance of facing a new long-lived opponent casts doubts on this history. Simultaneously, the likelihood of renewed friendly types must be sufficiently high so that posterior beliefs on unfriendly types stay below the threshold that induces exit of short-lived players. The salient fact is that information about intentions strongly complements the probability of renewal which means that the required probability is relatively small.

Given that the long-lived opponent is very patient, $\delta \to 1$, his long-run incentives are mainly determined by the probability of renewal. However, since the required probability of renewal comes from the posterior beliefs of short-run agents we find a detection technology that works at the stage-game level to avoid its dependence on renewal. We assumed that temptation actions are costly so there might be other actions which are profitable and not necessarily unfriendly. Then a high enough probability of detecting deviations from the code-of-conduct would inhibit the long-lived player to play these types of actions. Basically, once temptation actions are not longer a problem we do have the issue of profitable actions which was not an issue in the bad reputation game.

3.3 Ely-Välimäki Example

The bad reputation game in Ely and Välimäki (2003) consists of a long run player, a mechanic and a sequence of short-run players, the customers. There are two equally likely i.i.d. states of the world $\omega \in \{\mathcal{E}, \mathcal{T}\}$ that are observed only by the mechanic. These two states are what type of repair the car needs: engine replacement $\mathcal{E}$, or tune-up $\mathcal{T}$. The action space of the mechanic is $A = \{ee, et, te, tt\}$ where te reads for tune-up in state $\mathcal{E}$ and engine replacement in state $\mathcal{T}$. The customer's action space is $B = \{In, Out\}$ with In stands for hire the mechanic and Out stands for not hire the mechanic. The public outcomes are $Y = \{e, t, Out\}$ with distribution ρ described by $\rho(Out|(\cdot, Out)) = 1$, and the corresponding announcements $\rho(e|(et, In)) = \rho(e|(te, In)) = \frac{1}{2}$, $\rho(e|(ee, In)) = 1$ and $\rho(e|(tt, In)) = 0$.

If the mechanic performs the correct repair in each state and the costumer plays In, both receive a payoff of u. Otherwise, given participation, both get $-w$ with $w > u > 0$. Alternatively, if the customer plays Out both players get utility 0. We consider the finite horizon version of this game.

We briefly state elements of this bad reputation game. The set of friendly and unfriendly actions are $F = \{et\}$ and $N = \{ee\}$, respectively. Where e is evidence for the set $\{ee\}$. The enforceable friendly action et is vulnerable relative to e and the costly temptation is tt.

They point out that there exists a Nash equilibrium in which the long-run player chooses the correct repair and the costumer hires the mechanic (this also holds for the finite version). However, if we introduce a bad type of the long-run player that always plays the action ee, then with a patient enough mechanic the bound of his payoff converges to zero. Even though the stage game is played infinitely often we can show that if we take a sufficiently long but finite horizon the bad reputation effect results. When a bad type exists, short-run players update their posterior probability on this type with every realization of the bad signal e. Ely and Välimäki (2003) do not consider renewal. It may seem possible that with sufficiently high renewal the bad reputation effect is eliminated. Roughly, even when the discount factor tends to one the mechanic would not play the temptation action tt because he cares about short-run payoffs and customers would hire him given that public histories have low impact on posteriors. Adding self-referentiality has a multiplicative effect on renewal – there exists strong complementarity.

Next, we examine this example with self-referential games. The set of signals for the mechanic is $\{c, nc\}$ and for the customer is $\{m, g\}$, so the space of codes of conduct R_0 is the set of all mappings from $\{c, nc\} \times \{m, g\}^T$ to $\{ee, et, te, tt\} \times \{In,\ Out\}^T$. Denote $\mathbf{m}$ the $T \times 1$ vector with all entries filled by m.

Suppose first the perfect information case in which the probability distribution is given by $\pi((nc, \mathbf{m})|\tilde{r}^i, r^{-i}) = 1$ and $\pi((nc, \mathbf{m})|r) = 0$ for all players $i \in \mathcal{I}$. The following result states that if we have perfect information we can recover the good equilibrium in Ely-Välimäki's example with bad type.

Proposition 1 *Suppose there are normal and unfriendly commitment types for the mechanic. In the self-referential Nash equilibrium, the mechanic gets a normalized discounted payoff of u.*

This result is an immediate consequence of Theorem 1, and its intuition is as follows. The code-of-conduct requires the mechanic to "do the right thing," that is, to play the friendly action et whenever he observes c. For short-run players, the code requires them to hire the mechanic if g is observed, do not hire otherwise. This code is a self-referential Nash equilibrium.

In the imperfect information case, we do not state the analogous result to Proposition 1. Instead, we discuss some features of the one-shot version of this example. To simplify the analysis we assume that $\pi((nc, \mathbf{m})|r) = p$ and $\pi((nc, \mathbf{m})|\tilde{r}^i, r^{-i}) = q$ with $q \geq p$ for all players i. This type of signal structure could be interpreted as the situation in which the profile of signals $(c, \mathbf{g})$ is more likely to be observed if all players adhere to the same code-of-conduct. We restrict attention to three types: normal, unfriendly and friendly (Stackelberg). Stackelberg type plays et. Let μ^* be the prior probability of unfriendly type that induces entry of short-run player when there are only Stackelberg and unfriendly types.

Proposition 2 *Assume that there are three types of mechanic in the one-shot version of the game. Given the above self-referential characterization, there exists a code of conduct r such that $\mu^* \leq \mu^*(r)$ and the mechanic is hired.*

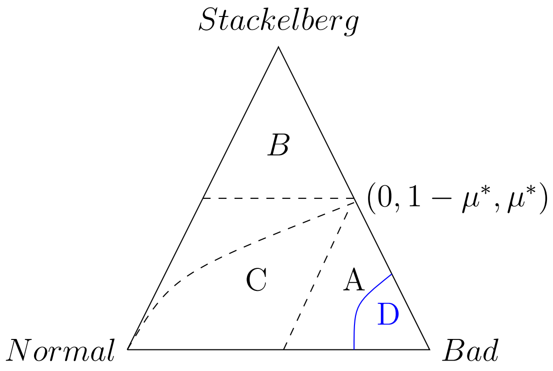

Figure 1: Space of prior distributions.

While the requirement on the prior of the bad type is reduced, the code cannot be a self-referential Nash equilibrium in the repeated version of the game. As long as the detection is imperfect, the updating of the likelihood of the bad type increases with the realization of bad signals – engine replacements. Hence, with sufficiently long horizon short-run players will not participate.

Now we illustrate the main result in Ely et al. (2008), and discuss how it is related to ours. Figure 1 represents the set of possible prior probabilities on normal, bad and Stackelberg types. The upper triangle region (region B) depicts the cases where short-run players enter irrespective of the behavior of the normal type long-run player. In this case reputation is always good. On the other hand, region A represents the cases when the probability of the bad type is so high that none of short-run players participate. Ely et al. (2008) show that in the region below the dashed curve (region C) bad reputation effect also arises. Moreover, the curve asymptotically reaches the lower left vertex. This implies that for any arbitrary perturbation of the complete information case, the bad reputation effect occurs. With self-referentiality and renewal, we can preclude reputation from being bad for any epsilon neighborhood of the complete information case. Note that from their result we cannot predict what would happen in the area between region B and region C. The completeness of our characterization comes from the fact we delimit the regions where reputation is bad or good for the entire space of prior probability distributions. In addition, we show that the cases where bad reputation effect occurs with self-referentiality (Region D) is a subset of their region A. We should stress that the higher the precision of the probability of detecting deviation from the code-of-conduct, the easier is to overcome the bad reputation effect in the self-referential game.

3.4 Bad Reputation Games without Renewal

In this section we discuss our assumption of renewing the long-run player. If detecting deviation from the code of conduct is imperfect and there is *no* renewal, we show that the bad reputation effect persists for arbitrary long horizon.

At the beginning of each period t, short-run player makes inference about the likelihood of friendly actions by combining the public history and information about long-lived agent's intentions of play. For arbitrary long horizon T, an unlucky history plenty of bad signals would make short-run players exit the game because this type of histories are very informative. The reason is that self-referential private signals modify posterior beliefs of short-lived players just a little bit.

We restrict attention to short-run players' updating beliefs since here the assumption of replacement is crucial. For a code-of-conduct to be a Nash equilibrium we require short-run players to participate in the game if the posterior probability of unfriendly types incorporating the self-referential information is sufficiently low.

Proposition 3 *Assume a finite bad reputation game with costly temptation and uniformly friendly commitment size $1-\gamma$ with constant $\chi > 0$, and that for some constants E, D the self-referential game E, D permits detection with $D > 0$. Consider the code-of-conduct profile used in Theorem 2. If h^t is a positive probability history in the self-referential Nash equilibrium with k signals in $\bar{Y}$ then*

(a) For all private signals lie in $Z_2 \setminus \bar{Z}_2$, we have $\mu(h^t, z_2)[\Theta(N)] < \mu(h^t)[\Theta(N)]$;

(b) Given all private signals $\bar{z}_2 \in \bar{Z}_2$, $\mu(h^t, \bar{z}_2)[\Theta(N)] > \mu(h^t)[\Theta(N)]$;

(c) For arbitrary long horizon T, there exists a number k^* of bad signals such that short-run players do not participate in the self-referential game.

Note that no matter how accurate the signals might be, the self-referential posterior probability weighs a bad signal as evidence of an unfriendly commitment type. That is, the code-of-conduct does not allow us to completely undermine the posterior when a bad signal is observed, but it certainly reduces the weigh on unfriendly commitment types if that realization is complemented with good private signal. In essence with a history of signals lying in $\bar{Y}$, observation of $Z_j \setminus \bar{Z}_j$ allows higher number of bad signals to "convince" short-run players to exit. Conversely, private signals in $\bar{Z}_j$ jointly with bad signals imply a greater posterior probability.

Statement (c) points out a potential problem on pursing to sustain friendly actions by using the code of conduct in the proof of Proposition 2. Since short-run players increase the posterior probability of unfriendly commitment types every time a bad signal is observed, eventually $\mu(h, z)[\Theta(N)] > 1 - \gamma$. This proposition shows one limitation of the applicability of self-referential games to bad reputation games.

4 Conclusion

We have analyzed self-referential games in the context of bad reputation games. Our results say that self-referentiality with renewal allows us to mollify the perverse effect of reputation in this class of games. We have also identified conditions on the relatively likelihood of unfriendly and friendly commitment types for such results to hold considering all possible distributions over types. In that sense our characterization is more complete than previous research. We assume that there exists the possibility of renewing the long-lived player. As we mentioned, we do not observe shut-down of markets where bad reputation effect is strong. We think one reason is that we have impermanent types of the long-run player. Yet how often this replacements occur do not seem to be frequent. We view plausible the idea that the probability of detecting deviation from code-of-conduct requires then a lower frequency of renewal. Both ideas are used to reconcile predictions from bad reputation models and the existence of markets with those characteristics. It is also important that our results hold for arbitrarily long finite horizon games.

A Appendix

Proof of Theorem 1 Let us define the minmax payoff $\underline{u}_2$ for short-run player 2 given by $\underline{u}_2 = \min_{\alpha \in \mathcal{A}} \max_{b \in B} u_2(\alpha, b)$, and let α^2 be the long-lived player's strategy that minimizes the short-run player in the stage-game. Pick the profile of code-of-conduct $r \in \mathcal{R}$ such that for all players $i = 1, 2$, $r^i \in R_0$ prescribes:

$$r_1^i(z_1) := \begin{cases} f^* & \text{for all } z_1 \in Z_1 \setminus \bar{Z}_1, \\ \alpha^2 & \text{otherwise,} \end{cases} \quad \text{and} \quad r_2^i(z_2) := \begin{cases} \beta \in \mathbf{B}(f^*) & \text{for all } z_2 \in Z_2 \setminus \bar{Z}_2, \\ e \in E & \text{otherwise.} \end{cases}$$

With some abuse of notation, we denote by f^* the strategy for long-run player which prescribes the play of friendly action f^* every period, similarly, for strategy α^2. It remains to show that this profile of codes of conduct $r \in \mathcal{R}$ constitutes a Nash equilibrium in the self-referential game. Note that the long-run player gets an expected payoff of $u_1(f^*, \beta)$ which is the most he can get by playing his Stackelberg friendly action, any other action will cause the short-run player to exit game so he does not have incentives to deviate from this code. Similarly for short-run player, by adhering to this code he expects a friendly action to be played and avoids being minmaxed by the long-lived player. ∎

Proof of Theorem 2 For any positive probability public history $h^t \in H^t$, let $\mu(h^t)[\Theta_0]$ be the posterior beliefs over types $\Theta_0 \subseteq \Theta$. For any player $i \in \mathcal{I}$ and any private signal $z_i \in Z_i$, let $\mu(h^t, z_i)[\Theta(N)]$ be the posterior beliefs on unfriendly types after incorporating the information of the self-referential game using Bayes' rule, and let $\mu_0(z_i)[\Theta_0]$ be the posterior beliefs for the null history h^0. We write $\tilde{\mu}(h^t, z_i)[\Theta(N)]$ for the posterior beliefs at the beginning of period $t+1$ taking into account private signals and the probability of renewal of the long-lived player, and its formal expression can be found below. We next construct the profile of code-of-conduct

$r \in \mathcal{R}$ with $r^i \in R_0$ for all players i such that for the long-run player

$$r_1^i(z_1) := \begin{cases} f \in F & \text{for all } z_1 \in Z_1 \setminus \bar{Z}_1, \\ a \notin F & \text{otherwise.} \end{cases}$$

Same disclaimer about notation as in the previous proof applies here. For short-run players we have

$$r_2^i(z_2) := \begin{cases} \beta \in \mathbf{B}(f^*) & \text{if } \tilde{\mu}(h^t, z_2)[\Theta(N)] < 1 - \gamma \text{ for all } z_2 \in Z_2 \setminus \bar{Z}_2, \\ e \in E & \text{otherwise.} \end{cases}$$

Suppose that all players i adhere to the proposed code-of-conduct $r^i \in R_0$. Consider any positive probability history $h^t \in H^t$, and assume $z_2 \in Z_2 \setminus \bar{Z}_2$. Then the posterior beliefs of short-run player on the unfriendly commitment types at the beginning of period t can be written as

$$\tilde{\mu}(h^t, z_2)[\Theta(N)] = \lambda \mu_0(z_2)[\Theta(N)] + (1-\lambda)\mu(h^t, z_2)[\Theta(N)]$$

The posterior beliefs is a linear combination of two components. The first component takes into account the possibility of renewal of the long-run player. The second component combines the information in the public history up to period t and the information in the private signal. We define the constant $\Lambda = (1-E)/(1-(E+D))$ for natational convenience, and note that $\Lambda > 1$ as $D > 0$. By Bayes' rule we obtain

$$\begin{aligned} \tilde{\mu}(h^t, z_2)[\Theta(N)] = {} & \frac{\lambda(1-\pi_2(\bar{z}_2|\tilde{r}^1, r^{-1}))\mu_0[\Theta(N)]}{(1-\pi_2(\bar{z}_2|\tilde{r}^1, r^{-1}))\mu_0[\Theta(N)] + (1-\pi_2(\bar{z}_2|r))(1-\mu_0[\Theta(N)])} \\ & + \frac{(1-\lambda)(1-\pi_2(\bar{z}_2|\tilde{r}^1, r^{-1}))\mu(h^t)[\Theta(N)]}{(1-\pi_2(\bar{z}_2|\tilde{r}^1, r^{-1}))\mu(h^t)[\Theta(N)] + (1-\pi_2(\bar{z}_2|r))(1-\mu(h^t)[\Theta(N)])} \end{aligned}$$

Observe that all unfriendly commitment types would be violating the code of conduct r. Pick the history $\hat{h}^T \in H$ where all signals $y_t \in \hat{h}^T$ belong to the set of bad signals $\bar{Y}$. Since the self-referential game has uniformly friendly commitment size $1-\gamma, \chi$ for some constant $\chi > 0$, and probability of replacement is λ the posterior beliefs then can be bounded by

$$\tilde{\mu}(\hat{h}^T, z_2)[\Theta(N)] \leq \frac{\lambda[1-(\gamma+\chi)]}{1-(\gamma+\chi)(1-\Lambda)} + \frac{(1-\lambda)[1-\lambda(\gamma+\chi)]}{1-\lambda(\gamma+\chi)(1-\Lambda)}$$

By adhering to the code r, short-run players do not exit the game as long as the posterior beliefs satisfy $\tilde{\mu}(\hat{h}^T, z_2)[\Theta(N)] < 1-\gamma$. Using the last expression, this is equivalent to

$$-\lambda^2(\gamma+\chi)\Lambda + \lambda(\gamma+\chi)b - \gamma(1-(\gamma+\chi)(1-\Lambda)) > 0$$

where the constant b is given by $b \equiv (\gamma+\chi)[\Lambda - (\gamma+\chi)(1-\Lambda)(\Lambda(1-\gamma)+\gamma) + \gamma + \Lambda(1-\gamma)]$. Thus, from the second-order polynomial we obtain

$$\tilde{\lambda} = \frac{-b - \sqrt{b^2 - 4\Lambda\gamma(1-(\gamma+\chi)(1-\Lambda))}}{-2(\gamma+\chi)\Lambda}$$

For all probabilities of renewal λ with $\lambda \geq \tilde{\lambda}$ we have $\tilde{\mu}(h^t, z_2)[\Theta(N)] < 1 - \gamma$ for any history $h^t \in H$ that guarantees short-run player would participate in the game. Observe that since all short-run players draw the same signal about long-lived player's intentions, his incentives to follow the code are driven at a stage-game level. Suppose that all short-run player adhere to the proposed code-of-conduct. Let $M = \max_{i \in \mathcal{I}} u_i$ and $m = \min_{i \in \mathcal{I}} u_i$ be the highest and lowest payoffs in the stage-game. If the normal long-run player adheres to the code-of-conduct r he obtains at least

$$u_1(f^*, \beta) - (1 - (1 - E)^2)(u_1(a, \beta) - u_1(f^*, \beta) + M - m) - \pi_2(\bar{z}_2 | r)(u_1(f^*, \beta) - u_1(f^*, e))$$

On the other hand, if he optimally deviates and chooses the code-of-conduct $\tilde{r}^1$ in which he plays some action $\tilde{a} \in A$ he gets at most

$$\begin{aligned} & u_1(\tilde{a}, \beta) - (1 - (1 - E)^2)(u_1(a, \beta) - u_1(\tilde{a}, \beta) + M - m) \\ & \quad - (\pi_2(\bar{z}_2 | r) + D)(u_1(f^*, \beta) - u_1(f^*, e) + u_1(a, \beta) - u_1(\tilde{a}, \beta)) \end{aligned}$$

Then the gain to deviation can be bounded above by

$$\begin{aligned} & u_1(\tilde{a}, \beta) - u_1(f^*, \beta) - (1 - (1 - E)^2)(2u_1(a, \beta) - u_1(f^*, \beta) - u_1(\tilde{a}, \beta) + 2(M - m)) \\ & \quad - (\pi_2(\bar{z}_2 | r) + D)(u_1(\tilde{a}, \beta) - u_1(a, \beta)) - D(u_1(f^*, \beta) - u_1(f^*, e)) \\ & \quad \leq u_1(\tilde{a}, \beta) - u_1(f^*, \beta) + 2E(2u_1(a, \beta) - u_1(f^*, \beta) - u_1(\tilde{a}, \beta) + 2(M - m)) - D\mathcal{C}_1 \end{aligned}$$

where $\mathcal{C}_1 = (u_1(\tilde{a}, \beta) - u_1(a, \beta) + u_1(f^*, \beta) - u_1(f^*, e))$. Adherence to the code-of-conduct r by the normal long-run player requires

$$D > \frac{u_1(\tilde{a}, \beta) - u_1(f^*, \beta) + 2E\mathcal{C}_2}{\mathcal{C}_1}$$

we have defined constant $\mathcal{C}_2 = (2u_1(a, \beta) - u_1(f^*, \beta) - u_1(\tilde{a}, \beta) + 2(M - m))$. This shows that the code-of-conduct r is a self-referential Nash equilibrium. ■

Proof of Proposition 2 If there are only Stackelberg and bad types in Ely-Välimäki example the prior probability of bad type is bounded above by $\mu^* \leq 2u/w + u$. For the case with three types the code-of-conduct profile r states that the long-run player plays et for any signal in $\{c, nc\}$, and the short-run player participates if he receives the signal g and stays out if b is realized. The prior probability of bad type required for having short-run player participating is given by $\mu^*(\hat{r}) \leq (1 - p)2u/(w + u - q(w - u) - p2u)$. It is immediate that $\mu^*(\hat{r}) \geq \mu^*$ as $q \geq p$. ■

Proof of Proposition 3 We begin with part (a) of the Proposition. Suppose that the short-run players observe only signals $z_2 \in Z_2 \setminus \bar{Z}_2$. Recall that

$$\begin{aligned} \mu(h^t, z_2)[\Theta(N)] &= \frac{(1 - \pi_2(\bar{z}_2 | \tilde{r}^1, r^{-1}))\mu(h^t)[\Theta(N)]}{(1 - \pi_2(\bar{z}_2 | \tilde{r}^1, r^{-1}))\mu(h^t)[\Theta(N)] + (1 - \pi_2(\bar{z}_2 | r))(1 - \mu(h^t)[\Theta(N)])} \\ &\geq \frac{1}{\mu(h^t)[\Theta(N)] + \Lambda(1 - \mu(h^t)[\Theta(N)])} \mu(h^t)[\Theta(N)] \end{aligned}$$

Observe that $\mu(h^t, z_2)[\Theta(N)] < \mu(h^t)[\Theta(N)]$ as $\Lambda > 1$. Suppose that for some $\epsilon > 0$, $\mu(h^t)[\Theta(N)] = 1 - \gamma + \epsilon$ and $\mu(h^{t-1})[\Theta(N)] < 1 - \gamma$. Thus

$$\mu(h^t, z_2)[\Theta(N)] \geq \frac{1}{1 - \gamma + \Lambda\gamma - \epsilon(\Lambda - 1)}\mu(h^t)[\Theta(N)]$$

Next, we borrow from Lemma 2 in Ely et al. (2008) the following lower bound on the posterior probability of unfriendly commitment types[8]

$$\mu(h^t)[\Theta(N)] \geq \left(\frac{1}{1 - \gamma + \frac{\gamma}{\kappa}}\right)^k \mu_0[\Theta(N)]$$

This shows that the lower bound for $\mu(h^t, z_2)[\Theta(N)]$ is below the lower bound found in the last expression (without self-referentiality) which implies that signals in $Z_2 \setminus \bar{Z}_2$ allow for a greater number of bad signals $\bar{Y}$ given the same history.

Suppose short-lived players' private signals $z_2 \in \bar{Z}_2$. By similar arguments, it must be the case that $\mu(h^t, \bar{z}_2)[\Theta(N)] > \mu(h^t)[\Theta(N)]$. That is, for a given number of bad signals, observation of private signals in $\bar{Z}_2$ pushes the posterior probability upward relatively more to the case without code-of-conduct.

Finally, note that the updating beliefs formula with self-referential signals is characterized by the factor $\Upsilon \equiv 1/\mu[\Theta(N)] + \Lambda(1 - \mu([\Theta(N)])$, from which we stress $\Upsilon \to 1$ as $\mu \to 1$. ∎

References

Block, J. I., and Levine, D. K. (2012). *Codes of conduct, private information and repeated games* (mimeo). Washington University in St. Louis.

Ekmekci, M., Gossner, O., and Wilson, A. (2012). Impermanent types and permanent reputations. *Journal of Economic Theory*, *147*(1), 162-178.

Ely, J. C., Fudenberg, D., and Levine, D. K. (2008). When is reputation bad? *Games and Economic Behavior*, *63*(2), 498-526.

Ely, J. C., and Välimäki, J. (2003). Bad reputation. *The Quarterly Journal of Economics*, *118*(3), 785-814.

Fudenberg, D., and Levine, D. K. (1989). Reputation and equilibrium selection in games with a patient player. *Econometrica*, *57*(4), 759-78.

Fudenberg, D., and Levine, D. K. (1992). Maintaining a reputation when strategies are imperfectly observed. *Review of Economic Studies*, *59*(3), 561-79.

Fudenberg, D., Levine, D. K., and Maskin, E. (1994). The folk theorem with imperfect public information. *Econometrica*, *62*(5), 997-1039.

Holmström, B. (1999). Managerial incentive problems: A dynamic perspective. *Review of Economic Studies*, *66*(1), 169-82.

[8] To find the number k they use the argument in Fudenberg and Levine (1989, 1992).

Kalai, A. T., Kalai, E., Lehrer, E., and Samet, D. (2010). A commitment folk theorem. *Games and Economic Behavior*, *69*(1), 127-137.

Kreps, D. M., and Wilson, R. (1982). Reputation and imperfect information. *Journal of Economic Theory*, *27*(2), 253-279.

Levine, D. K., and Pesendorfer, W. (2007). The evolution of cooperation through imitation. *Games and Economic Behavior*, *58*(2), 293-315.

Liu, Q. (2011). Information acquisition and reputation dynamics. *The Review of Economic Studies*, *78*(4), 1400-1425.

Liu, Q., and Skrzypacz, A. (2009). *Limited records and reputation* (Research Paper 2030). Stanford University, Graduate School of Business.

Mailath, G. J., and Samuelson, L. (2001). Who wants a good reputation? *Review of Economic Studies*, *68*(2), 415-41.

Mailath, G. J., and Samuelson, L. (2006). *Repeated games and reputations: Long-run relationships.* Oxford University Press.

Milgrom, P., and Roberts, J. (1982). Predation, reputation, and entry deterrence. *Journal of Economic Theory*, *27*(2), 280-312.

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

Wiseman, T. (2008). Reputation and impermanent types. *Games and Economic Behavior*, *62*(1), 190-210.