ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

RE-THINKING REPUTATION

Andrew Mell

Number 565
September 2011

Manor Road Building, Oxford OX1 3UQ

Re-Thinking Reputation*

Andrew Mell†
Nuffield College
New Road
Oxford
OX1 1NF

September 2, 2011

Abstract

Economic models of reputation make strong assumptions about the information available to players. In particular, it is assumed that they know the entire history of the game to date. Such models can seldom reproduce the cycling of reputations we observe in the real world. We build a model of reputation with more realistic assumptions about the partial knowledge of the history that would be available and how it might be used. This new approach can explain cycles in reputations.

JEL Classification Numbers: D82, D84.

Keywords: Reputation, Monitoring, Expectations Formation.

*I would like to thank my supervisors, Peyton Young and Meg Meyer, and participants in the workshop on Learning in Games at Nuffield College, University of Oxford.

†andrew.mell@nuffield.ox.ac.uk

1 Introduction and Relevant Literature

Consider a good most of us will buy only once in our lives. Tertiary education provides a good example. Colleges would like their students to work hard. A student will benefit from doing so only if the professors and teaching assistants exert costly effort in designing and teaching the courses. Professors and teaching assistants would rather exert effort on their own research projects. However, whether they exert effort or not, they prefer it if their students work hard.

We could easily see this situation as a game played between the college and the student. Both players simultaneously choose whether to exert effort, 'E', or not, 'N'. An example payoff matrix, with best response payoffs in bold, is shown in Table 1 which shows a stage game used in Cripps et al. (2004).

		Student	
		E	N
College	E	**3** 2	2 0
	N	0 **3**	**1** **1**

Table 1: An education game.

The problem is that N is a dominant strategy for the college, so the only Nash equilibrium of the stage game involves both players choosing not to exert any effort. Yet both players would be better off if they both exerted effort. Indeed, this is what would happen if the game were played sequentially and the college got to move first. Absent any commitment technology, the only way to achieve this superior outcome is for the college to build and maintain a "reputation" for exerting effort. Then the student will expect effort and find it optimal to exert effort herself.

However, modeling the process of "building and maintaining a reputation" is no simple matter. Many different models exist with their own insights and shortcomings. An excellent summary of the existing models can be found in the Bar-Isaac and Tadelis (2008) survey article, and the reader interested in a full survey of the available literature on reputation is directed there. One regularity which we will observe here is that a reputation game is often thought of as the repeated play of a stage game. One long-lived player plays in every period in the position of player 1. A sequence of short-lived players take turns to play as player 2. Each short-lived player plays once. The long-lived player is the one who tries to build a reputation and the short-lived players are the ones who try to read it.

For our purposes, we will make a distinction between two types of reputation model. Broadly speaking we will call these folk theorem models and incomplete information models.

Folk Theorem models work through the logic of infinitely repeated games. Take the example of the game in Table 1. A folk theorem reputation result would look for a subgame perfect equilibrium that would induce perpetual play of {E,E}. The subgame perfect equilibrium (SPE) would rely on indefinite repetition and a sufficiently patient long-lived player. The SPE strategy profile has the college always exerting effort unless at a history where they have not exerted effort in the past; and each student exerting effort unless at a history where the college has failed to do so in the past. Kreps (1990) is a good example of a reputation model based on the folk theorem. As Selten (1978) pointed out, there is an intuitive paradox in the way such a model fails when the game is only repeated a large but finite number of times. We would intuitively expect reputation to be built up at the beginning of finite games, but under the folk theorem this is impossible because of the end point problem.

An incomplete information model would suppose that there is a small probability that the college is committed to exerting effort. Then the optimal strategy for the college that is not committed, provided they are sufficiently patient, is to exert effort in order to try and preserve that uncertainty. If there is a final period then they might switch to not exerting effort some short time before that final period. However this is precisely the advantage of incomplete information models over folk theorem models of reputation; they work in games with finite repetitions. Kreps and Wilson (1982); Milgrom and Roberts (1982); Kreps et al. (1982) are all classic examples of this approach. However such models assume impressive information gathering and processing capabilities on the part of short-lived players.

The reputation model I will present here is capable of producing reputation results, even in finite games, without the need to introduce different types of long-lived players. Furthermore we do not require exceptional information acquisition and processing abilities in the short-lived players. We have more realistic assumptions about the sources players use to gather information about a potential opponent and the nature of the information that can be gathered. A third benefit of our method is that cycles in reputations follow quite naturally. The poor behavior of the long-lived player is observed and punished in some periods and the reputation is built up again later. The reputations in the papers mentioned above, and in the more general setting of Fudenberg and Levine (1989), are lost forever should the long-lived player every take any action other than the Stackleberg action.

Imperfect monitoring technologies can go some way towards resolving the problem of implausibly perfect reputations. Green and Porter (1984) show this for folk theorem models and Fudenberg and Levine (1992) show this for incomplete information models. However, an objection to the model presented by Green and Porter (1984) is that when punishments do occur on the equilibrium path, everyone knows that no actual cheating occurred. It was simply a failure of the monitoring technology. Further, Cripps et al. (2004) show, when monitoring is imperfect in an incomplete information reputation model, the reputation must always break down in the very long run. A repeated Nash equilibrium of the stage game will be played in the very long run. The reputation

breaks down because as the short-lived players become more and more certain they are playing the commitment type, they will update that probability more and more slowly in response to bad signals. So the long-lived player's temptation to cheat on their reputation becomes greater and greater until it becomes irresistible.

There are two different strands of the literature that seek to resolve the problem identified by Cripps et al. (2004). The first is exemplified by by Mailath and Samuelson (2001); Phelan (2006); Wiseman (2008) and involves allowing for some exogenous probability of a hidden change of type. This bounds the probability of a commitment type away from unity and ensures it will always adjust reasonably quickly in response to bad news. A second and on the whole more recent strand rejects the hidden nature of these type changes as implausible and bounds the memory of short-lived players to achieve a similar result. This strand is exemplified by Ekmekci and Wilson (2005); Monte (2007); Ekmekci and Wilson (2007); Liu (2010); Liu and Skrzypacz (2011).

My approach is more similar to this second strand. However I do not suppose that short-lived players interacting for one period only will have the ability, inclination or time to bring Bayesian inference to bear on formulating their expectations. Instead I incorporate some of the recent developments in evolutionary game theory, specifically the adaptive play formulated in Young (1993, 1998). I apply that to the behavior of the short-lived players. They draw some sample from the recent actions of the long-lived player and best respond to the (possibly mixed) strategy that sample suggests the long-lived player is using. What we have in mind is a process of 'asking around'. In the education game example above we could imagine prospective students asking recent graduates of the college what they thought of their educational experience. However we still assume that the long-lived player behaves like a traditional *homo economicus*. Since they will be playing in this interaction for a very long time, they are more dependent on it and have a greater incentive to understand it. This is not the first model to examine the consequences of interaction between the limited rationality agents from evolutionary game theory and the fully rational players of standard game theory. Ellison (1997) examines whether a rational player would have an incentive to try and shift the equilibrium if playing in a population of evolutionary game theory's fictitious players from Fudenberg and Levine (1993).

The next section will explain more fully how this model works through the use of an example. Section 3 looks at more general reputation games and derives the results proving that a reputation will be built under certain conditions. Section 4 looks at a more complex example and examines the optimal strategy of a long-lived player and the phases where they build a reputation and the phases where they 'coast' on the basis of that reputation. Section 5 examines the incentive to 'coast' on a reputation in a more general setting. Section 6 takes the example from Section 4 and compares and contrasts my approach with the one that would be taken by an incomplete information model . Section 7 gives some brief consideration to how this model might work if a dynamic element were to be added to the stage game. Section 8 concludes.

2 A Different Information Set

Consider the education game in Table 1. Below, we reproduce two versions of this game. The left-hand panel shows how the game appears to the college. The right-hand panel shows how the game appears to one of the students when they are called to move. We will call the college player 1 with action set $A_1 = \{E, N\}$ and the student player 2 with action set $A_2 = \{E, N\}$.

		Student	
		E	N
College	E	**3** 2	2 0
	N	0 **3**	**1** **1**

		Student	
		E	N
College	E	**3** ?	2 ?
	N	0 ?	**1** ?

Table 2: The stage game for the long and short-lived players.

This is the stage game and it is played repeatedly. The college is (for the moment) infinitely lived and plays in every period.[1] The student playing the game is different in each period. There is an infinite sequence of students who each play once. However, before playing, the new student might meet some of the college's recent graduates and take the opportunity to ask them what the college did when they played. We model this as the student called to play in period t drawing a random sample (without replacement) of size s from the last m actions of the college. Following Young (1993, 1998), we call s the sample size and m the available history.

Each student when they are called to move has absolutely no idea about anything else in the game. They do not know the payoffs to the long-lived player, nor do they know what the past students' actions were when they were called to move. They do not even have a Bayesian prior on which to base any future inferences.

So we imagine the students naively believing that the frequency of actions by the college that appears in their sample is the probability with which each action will be taken when they play this game with the college. They best respond to the implied mixed strategy.[2,3] By contrast, we continue to model the college as a fully rational player.

Let p^t be the probability with which the student at time t believes the college will choose effort and play E. p^t will be equal to the proportion of the time that the college exerted effort in the sample that the student at time t has drawn from the available history. The expected payoff to the student from exerting effort is $3p^t$. Meanwhile the expected

[1] We will briefly consider finite reputation games later.

[2] One difference between this behavioural rule and the one used by Young (1993) is that we do not impose any limits on the size of the sample, s, except that it be strictly less than the available history, m.

[3] Shafir et al. (2008) found evidence of similar decision-making in humans under uncertainty and used it to explain a reversed certainty effect sometimes observed in experiments.

payoff from choosing not to exert any effort is $2p^t + 1 - p^t = 1 + p^t$. Exerting effort will therefore be a best response provided that:

$$\begin{aligned} & 3p^t \geq 1 + p^t, \\ \Rightarrow \quad & p^t \geq \frac{1}{2}. \end{aligned} \tag{2.1}$$

The probability that the student draws a sample such that $p^t \geq 1/2$ is determined by the distribution of the last m actions of the college.

For the sake of a tractable example, we will assume that $m = 4$ and $s = 1$. We will define the set of states, Ω as the set of all possible permutations and combinations of the last four moves of the college. The students' last four moves are irrelevant and so are not included in the state. A state is completely summarized by a four dimensional vector where each element takes one of two possible values, E or N, indicating that the action of the college at this point was "Effort" or "No Effort" respectively. So one possible state, $\omega^t \in \Omega$ might be:

$$\omega^t = \begin{array}{cccc} t-4 & t-3 & t-2 & t-1 \\ E & N & E & E \end{array}. \tag{2.2}$$

The transition rule, $\tau : \Omega \times A_1 \to \Omega$, is that given an initial state at time t, $\omega^t \in \Omega$, and an action by the long-lived player at time t, $a_1^t \in A_1 = \{E, N\}$:

- The entry in ω^t at $t-4$ is "dropped".
- All other entries move one place to the left.
- The action a_1^t, whatever it was is added in the now vacant right-most slot.

The transition rule is denoted by $\omega^{t+1} = \tau\left(\omega^t, a_1^t\right)$. For example, suppose that ω^t is as described in (2.2) and that effort is exerted by the college in period t, then:

$$\omega^{t+1} = \tau\left(\omega^t, E\right) = \begin{array}{cccc} t-3 & t-2 & t-1 & t \\ N & E & E & E \end{array}. \tag{2.3}$$

With $s = 1$, the student called to play at time t will either assess $p^t = 1$ if they sample a period which involved the play of "Effort" by the college, or $p_t = 0$ if the one period they sample involved the play of "No effort" by the college. In the former case, their best response, as found in (2.1) will be to exert effort themselves and in the latter case it will be not to exert effort. Let $n^t : \Omega \to \{0, 1, 2, 3, 4\}$ be a counting function such that $n^t\left(\omega^t\right)$ indicates the number of occasions in ω^t that involve the play of "Effort". So returning to the example from (2.2), $n^t\left(\omega^t\right) = 3$.

So in any particular state $\omega^t \in \Omega$, the probability that the student's sampling of past moves returns $p^t = 1$ is given by

$$\Pr\left(p_t = 1 | \omega^t \in \Omega\right) = \frac{n^t\left(\omega^t\right)}{4}. \tag{2.4}$$

Given the student's best response function, (2.4) is also the probability with which the student at time t exerts effort.

We now exploit the additive separability of the college's payoffs in the example game we are using in Table 2.[4]

- The immediate marginal benefit to the college from shirking instead of exerting effort is 1 and is independent of the action that the student at time t might happen to play in that period.
- The immediate marginal cost to the college of the student at time t deciding to shirk instead of exert effort is 2 and is independent of the action that the college might take in that period.

Let I_s be an indicator function that takes the value of 1 if the student exerts effort and 0 otherwise. Let I_c be an indicator function that takes the value of 1 if the college exerts effort and 0 otherwise. Then we can summarize the payoff to the College as

$$\pi_c = 1 + 2I_s - I_c.$$

The benefits and costs of each players' actions are separable for the college. For example, the benefit to the college of exerting no effort is independent of the decision of the student as to whether to exert effort.

This additive separability combined with the $s = 1$ sampling of the student has a useful feature. It means that the future costs and benefits of any action the college might take will be independent of both the college's recent actions and the college's future plans. The $s = 1$ sampling means that the impact of an extra N or E in the available history on the probability of a good or bad sample is independent of the current and future contents of the memorable history. So the opportunity cost of playing N rather than E is that the probability that the student draws a sample of N instead of E is 1/4 higher than it otherwise would have been for the next four periods. So the probability that they play N instead of E is 1/4 higher than it otherwise would have been over the next four periods. This is true irrespective of the current state. Furthermore the additive separability of the college's payoffs means that the payoff cost of this extra probability of the student 'doing the wrong thing' is independent of what actions the college plans to take over the next four periods.

This allows us to find the best response for the college with relative ease by comparing the marginal cost and benefit of playing E rather than N. They will be the same whatever state we happen to be in.

With $s = 1$ and $m = 4$, the marginal opportunity cost to the college from choosing No Effort is paid over the subsequent four periods. The probability that the students

[4]Additive separability refers to the way in which there is a constant cost to the college from exerting effort that is independent of the action of the student and a constant benefit to the college of the student's effort independent of their own action.

will draw a sample that contains only Effort, and so assign $p^{t+i} = 1$ for $i \in \{1, 2, 3, 4\}$ will be $1/m = 0.25$ lower than it would otherwise have been. So the probability that the students will exert effort will be 0.25 lower than it would otherwise have been. The cost of this reduction of the students' probability of exerting effort in the 4 subsequent periods discounted to today can be calculated as:

$$\begin{aligned} C &= 0.25 \times 2\delta + 0.25 \times 2\delta^2 + 0.25 \times 2\delta^3 + 0.25 \times 2\delta^4, \\ C &= \frac{1}{2}\delta\left(1 + \delta + \delta^2 + \delta^3\right). \end{aligned} \tag{2.5}$$

As observed above, the immediate benefit of playing "No Effort" instead of "Effort' is 1. So the college would choose to exert effort instead of shirk whenever:

$$\begin{aligned} \frac{1}{2}\delta\left(1 + \delta + \delta^2 + \delta^3\right) &> 1, \\ \delta\left(1 + \delta + \delta^2 + \delta^3\right) &> 2, \\ \delta &> \delta^* \approx 0.7412. \end{aligned} \tag{2.6}$$

So provided the college is patient enough, they will always choose "Effort" instead of "No Effort". If they are insufficiently patient, they will always choose "No effort" instead of "Effort". We have not needed either an expectation of indefinite continuation to achieve this reputation result. Neither have we needed an incentive to pool with some commitment type or separate from some inept type. It comes about because of the adaptive expectations of the short-lived players.

This example has not led to cycling reputations, but instead to reputations that are constant along the equilibrium path. This is the result of the same factors that made the model so easy to solve. When we complicate the model a bit more, by allowing for fuller, but still incomplete sampling, we will get cycles in reputations. However we will first record when there is an incentive to build a reputation under more general conditions on the stage game, sample size and available history.

3 General Reputation Games

We now examine more general reputation games. These are played between a long-lived player 1 who is fully rational and knows the entire game and an infinite sequence of short-lived players in the position of player 2. However, the short-lived players are only able to partially monitor the actions of the long-lived player. With no other information, they behave naively by responding optimally to their partial knowledge about the past actions of the long-lived player. I use the term "partial monitoring" to refer to the way in which the short-lived players only know about some of the recent past moves of the long-lived player. We will prove that, given the behavior of the short lived players, the equilibrium strategy of the sufficiently patient long lived player cannot constitute their part of a repeated Nash equilibrium of the stage game.

We make a distinction between reputation games of partial but perfect monitoring and reputation games of partial imperfect monitoring. In the former, the past moves of the long-lived player that the short-lived players do observe, they observe perfectly and know which action the long-lived player took. In the latter case, those past actions of the long-lived player that are sampled are observed only imperfectly. We will explain the nature of these imperfections later.

The proof is simplest and easiest for games of partial but perfect monitoring. So we will start with that case. The strategy is to show that from any Nash equilibrium convention the sufficiently patient long-lived player would prefer to move to a Stackleberg convention.[5] For games of partial and imperfect monitoring the complication is that the imperfect observation of the long-lived player's actions means that there is positive probability of short-lived players failing to appreciate the change in strategy. However if the sample size s is large enough (and the available history m is correspondingly larger) then the law of large numbers applies and a similar proof goes through.

3.1 Games of Partial and Perfect Monitoring

In order to prove that the long-lived player would prefer to move away from the Nash convention to the Stackleberg convention, we first put a bit of structure on the stage game. We then show that the strategy of adaptive short-lived players means that the problem of finding the best response for the long-lived player can be viewed as an optimal control problem. Finally we show that for a sufficiently patient long-lived player the best response cannot involve playing a repeated Nash equilibrium of the stage game.

The stage game

The stage game is played between the short-lived player 2 in that period and the long-lived player 1. Each player has a set of possible actions they can take, A_i, with a particular action denoted $a_i \in A_i$. The action sets are common knowledge. So the set of possible outcomes is $A = A_1 \times A_2$ with $a \in A$ representing a particular outcome. Each player has a V-NM payoff function $\pi_i : A \to \mathbb{R}$, the outcome of which is always finite, $-\infty < \pi_i(a) < \infty$. Player 1 knows both payoff functions while player 2 only knows her own.

Let B_i be the set of action profiles where player i is best responding to the action of player $j \neq i$. Then

$$B_i = \{a : a_i = \arg\max_{a_i \in A_i} \pi_i(a_i, a_j)\}.$$

So the set of outcomes $E = B_1 \cap B_2$ is the set of Nash equilibrium outcomes. Clearly, $E \subseteq B_2 \subseteq A$.

[5] By convention we mean a situation where the same action profile has been played for the last m periods.

In order to put ensure that reputation concerns are non-trivial we need to put a bit more structure on the stage game. It must be the case that

$$\exists a' = \arg\max_{a \in B_2} \pi_1(a) \quad \text{and} \quad a' \notin E.$$

We assume this profile to be the unique maximand of π_1 over B_2. It is the Stackleberg profile and a'_1 is the Stackleberg leader action and a'_2 is the Stackleberg follower action.

The optimal control problem

Where the available history from which the short-lived player will draw her sample consists of the last m periods, let $\Omega = A_1^m$ be the set of possible states. A typical member is $\omega \in \Omega$ which is a vector of the last m actions of the long-lived player. Denote the state in period t by ω^t. It is important that the state space is finite. Let $\tau : \Omega \times A_1 \to \Omega$ be the transition function which picks out the state to which the system moves from a given initial state and for a given action by the long lived player. So $\omega^t = \tau\left(\omega^{t-1}, a_1^{t-1}\right)$.

Let σ be a strategy for the long-lived player specifying an action in every state, and denote the specified action in state ω as $\sigma(\omega)$. Let $\Sigma = \{\sigma^{t+k}\}_{k=0}^{\infty}$ be an infinite sequence of such strategies, we call Σ a 'grand strategy'. Order the states $\omega_1, \omega_2, ..., \omega_{|\Omega|}$, and let $\mathbf{Q}(\sigma)$ be a $|\Omega| \times |\Omega|$ matrix whose (i,j) element is 1 if $\tau(\omega_i, \sigma(\omega_i)) = \omega_j$, and 0 otherwise. Let $\mathbf{Q}_k(\Sigma) = I_{|\Omega|}\mathbf{Q}\left(\sigma^t\right)\mathbf{Q}\left(\sigma^{t+1}\right)...\mathbf{Q}\left(\sigma^{t+k-1}\right)$ so that the (i,j) element of $\mathbf{Q}_k(\Sigma)$ is 1 if k periods after following strategy Σ from state ω_i, the system is in ω_j, and 0 otherwise.

Denote by $\Delta A_1 \in \Delta$ the point in the simplex over A_1 that the short-lived player 2 draws as her sample. The set of possible samples, $\Delta = A_1^s$ is finite because any sample that the short-lived player draws can only put probabilities that are multiples of $1/s$ on any member of A_1. Denote by $b_2 : \Delta \to A_2$ the best response function of player 2.[6] This function is common knowledge. For every state there will be a probability distribution over the samples that player 2 might draw, denoted by $\theta(\Delta A_1 | \omega)$.

For each action of player 2, $a_2 \in A_2$, define R_{a_2} as the set of samples that player 2 can draw as a best response,

$$R_{a_2} = \{\Delta A_1 : b_2(\Delta A_1) = a_2\}.$$

This collection of sets will be mutually exclusive and collectively exhaustive over the set of all possible samples that could be drawn. So they constitute a partition of the set of all possible samples that could be drawn by the short-lived player 2, Δ.

[6]For simplicity we assume that no mixing that places probabilities that are multiples of $1/s$ on the actions of player 1 can lead to ties in the payoffs of player 2. This guarantees that the best response of player 2 is unique.

So player 1 knows that from any given state in period t, ω^t, the probability that player 2 will play action $a_2 \in A_2$ is given by

$$p\left(a_2|\omega^t\right) = \sum_{\Delta A_1 \in R_{a_2}} \theta\left(\Delta A_1|\omega^t\right). \tag{3.1}$$

So the expected payoff to player 1 in period t from any given state, ω^t and any given action, a_1 can be characterized as

$$v_1\left(a_1|\omega^t\right) = \sum_{a_2 \in A_2} p\left(a_2|\omega^t\right) \pi_1\left(a_1, a_2\right). \tag{3.2}$$

Let $\mathbf{v}_1\left(\sigma^t\right)$ be a $|\Omega| \times 1$ column vector whose i^{th} element is $v_1\left(\sigma^t\left(\omega_i\right), \omega_i\right)$.

Suppose that player 1's rate of time preference is defined by the discount factor δ. So the best response of the long-lived player 1 when facing an infinite sequence of adaptive short-lived player 2s is to choose a 'grand strategy', Σ to maximize every element of the vector

$$\mathbf{V}_1\left(\Sigma\right) = \sum_{i=0}^{\infty} \delta^i \mathbf{Q}_i\left(\Sigma\right) \mathbf{v}_1\left(\sigma^{t+i}\right). \tag{3.3}$$

This is a special case of the problem solved by Blackwell (1962), who shows that there is a stationary solution to this problem. Stationarity means that the action called for by the optimal strategy is dependent only on the state and not on which period we are in. So it can be summarized by a mapping σ which does not depend on which period we are in. It is quite intuitive that the optimal strategy be stationary. Each time a state is visited, player 1 faces the same trade off between the opportunities for high payoffs today and the impact this will have on payoffs over the next m periods. The optimal way in which to trade off these two concerns should not change.

The stationarity of the solution to this problem has important consequences for the equilibrium outcome. It means that the equilibrium strategy of the long-lived player must eventually induce a cycle over some subset of the state-space. Consider any state that is visited more than once. The same action must be played the second time the state is visited as was played the first time the state was visited. This means we must transit to the same state in the next period as we did the last time the state was visited. Once there, we must take the same action as we did before, and so on, until we come to a state that has been visited twice already. Then it begins again. So the optimum strategy must result in cycles over some subset of the state space.

Avoiding the repeated Nash equilibrium

The definition of play being locked into a particular action profile is as follows:

Definition 1 *Play has become locked into the play of a particular strategy profile if and only if that strategy profile would be the only strategy profile seen if play continued indefinitely.*

Before considering Proposition 2 it may help to refresh on some of the other definitions and make some new ones:

$$\begin{aligned} a^* &= \arg\max_{a \in E} \pi_1(a), \\ a' &= \arg\max_{a \in B_2} \pi_1(a), \\ \underline{a} &= \arg\min_{a \in A} \pi_1(a), \\ \overline{a} &= \arg\max_{a \in A} \pi_1(a). \end{aligned}$$

And for notational simplicity, we will say that:

$$\begin{aligned} \pi_1^e &= \pi_1(a^*), \\ \pi_1^s &= \pi_1(a'), \\ \underline{\pi}_1 &= \pi_1(\underline{a}), \\ \overline{\pi}_1 &= \pi_1(\overline{a}). \end{aligned}$$

We can then make the following proposition:

Proposition 2 *If*

$$\delta > \left(\frac{\pi_1^s - \underline{\pi}_1}{\pi_1^e - \underline{\pi}_1} \right)^{\frac{1}{m}} \in (0,1), \tag{3.4}$$

Then the equilibrium strategy for a sufficiently patient player 1, playing an infinite sequence of adaptive player 2s will not involve being locked into any Nash equilibrium of the stage game.

Proof. See Appendix A ■

We have consigned the proof of Proposition 2 to the appendix, but the intuition is relatively simple. The worst possible stream of payoffs that the long-lived player could endure in moving from the best possible Nash equilibrium convention to the Stackleberg convention involves getting the worst possible stage game payoff for m periods in a row. However after that he gets the Stackleberg payoff forever which is greater than the best possible Nash payoff. For a sufficiently patient player, this will eventually overtake the payoff from continuing with the Nash convention. The longer the available history, the longer the transition takes. So the long-lived player will have to endure low payoffs for longer. So they will have to be more patient in order to find moving worthwhile.

Proposition 2 demonstrates that long-lived players who are sufficiently patient will have no incentive to perpetually play *any* Nash equilibrium of the stage game. This means that they must build some kind of reputation. Proposition 2 does not claim that they will *always* play the Stackleberg action. It simply claims that always playing it is better than always playing any Nash equilibrium. It provides a sufficient, not a necessary condition.

3.2 Games of Partial and Imperfect Monitoring

We will briefly discuss how the structure of the stage game in an imperfect monitoring setting differs from that in a perfect monitoring setting. Then we will discuss how the nature of the optimal control problem changes when we have to allow for imperfect monitoring. Finally we will demonstrate that in the presence of imperfect monitoring a version of Proposition 2 holds with an additional requirement on the sample size and available history being large enough.

Changes to the stage game for imperfect monitoring

The key change we need to make for the stage game with imperfect monitoring is that player 2 cannot see exactly what player 1 did in the past events she knows about. Instead she sees an informative signal and her payoff depends on this signal rather than the actual move. Let the set of possible signal outcomes be Y with typical element $y \in Y$. We assume the set Y to be countably finite. Each action for player 1 induces a different probability distribution over the signal space, $f_{a_1}(y)$. So $\forall a_1 \neq a_1'$, $\exists y$ such that $f_{a_1}(y) \neq f_{a_1'}(y)$.

Player 2 never finds out what action was taken. Her payoff function must depend only on the signal outcome and her action and not (directly) on the action of player 1, $\pi_2 : Y \times A_2 \to \mathbb{R}$. This means we must formulate her best response to the action of player 1 slightly differently. She must maximize her expected payoff given the action of player 1 and the probability distribution over signal outcomes that this generates. So

$$B_2 = \{a : a_2 = \arg\max_{a_2 \in A_2} E_{f_{a_1}(y)}\{\pi_2(y, a_2)\}.$$

The best response function for player 1 is still defined as before as are the set of Nash equilibria and the Stackleberg action profiles.

An optimal control problem

What is now relevant in terms of the state is the past m signal outcomes rather than the past m actions of the long-lived player. So it will now be the case that $\omega \in \Omega$ is a vector of the last m signal outcomes and $\Omega = Y^m$. The transition function will also have changed. It will now be probabilistic and will generate a probability distribution over the state to which we will move next. The transition function $\tau : \Omega \times A_1 \to 1 \times |\Omega|$ now maps to a row vector whose j^{th} element is the probability that for a given starting state and a given action we transition to state ω_j.

The matrix $\mathbf{Q}(\sigma)$ can now be constructed by 'stacking' the row vectors $\tau(\omega_i, \sigma(\omega_i))$ so that the (i, j) element gives the probability of transitioning from state i to state j when taking the action prescribed by the strategy σ. Where Σ represents the grand strategy composed of a sequence of strategies $\{\sigma^{t+i}\}_{i=0}^{\infty}$, we can say that $\mathbf{Q}_k(\Sigma) =$

$I_{|\Omega|}\mathbf{Q}\left(\sigma^t\right)\mathbf{Q}\left(\sigma^{t+1}\right)...\mathbf{Q}\left(\sigma^{t+k-1}\right)$. The (i,j) element of $\mathbf{Q}_k\left(\Sigma\right)$ will record the probability of finishing in state j having been in state i k periods ago and followed grand strategy Σ since.

The sample that the short-lived player draws can now be captured by a point in the simplex ΔY. We again assume there are no ties in payoffs when the probability of each signal outcome must be a multiple of $1/s$. So we can again partition the set of possible samples, Δ based on the action from player 2 that they induce as a best response. This partition is summarized by

$$R_{a_2} = \{\Delta Y : b_2\left(\Delta Y\right) = a_2\}.$$

The probability of any particular sample being drawn given any particular state at time t can be denoted by $\theta\left(\Delta Y|\omega^t\right)$. So the probability that the short-lived player takes any particular action $a_2 \in A_2$ given a particular state at time t is given by

$$p\left(a_2|\omega^t\right) = \sum_{\Delta Y \in R_{a_2}} \theta\left(\Delta Y|\omega^t\right). \tag{3.5}$$

So the expected payoff to player 1 from any particular action $a_1 \in A_1$ is

$$v_1\left(a_1|\omega^t\right) = \sum_{a_2 \in A_2} p\left(a_2|\omega^t\right)\pi_1\left(a_1,a_2\right). \tag{3.6}$$

As before we stack these to generate the column vector $\mathbf{v}_1\left(\sigma\right)$, whose i^{th} element is given by $v_1\left(\sigma\left(\omega_i\right)|\omega_i\right)$.

So the problem for the long-lived player is to maximize $\mathbf{V}_1\left(\Sigma\right)$ given below by choosing the optimal Σ.

$$\mathbf{V}_1\left(\Sigma\right) = \sum_{i=0}^{\infty}\delta^i\mathbf{Q}_i\left(\Sigma\right)\mathbf{v}_1\left(\sigma^{t+i}\right). \tag{3.7}$$

This is exactly the problem solved by Blackwell (1962). So we know that the solution will be stationary again.

However a stationary optimal strategy does not have the same implication about inducing deterministic cycles through the state space as before. Arriving at a state that the long-lived player has already visited does mean that he will take the same action in this state as he did the last time he was there. However this will not necessarily mean moving to the same subsequent state as before. The same action from the same state might induce a different signal and so a transition to a different state. However it will be the case that the probability of being in any particular state at any future period will be conditional only on the current state and will be the same whether that state is reached in period t, or in period $t+10$ or in period $t+1000$.

Avoiding the repeated Nash equilibrium

We start by proving the following Lemma.

Lemma 3 *Suppose that we are at the beginning of period t and the long-lived player plans to play the same action $a_1 \in A_1$ for the next m periods. $\forall \epsilon > 0$, there exists an $\overline{s}$ sufficiently large but finite such that, for $m > s > \overline{s}$; and $\forall y \in Y$, the proportion of occasions in which y is expected to appear in the short-lived player's sample in m periods time, denoted by $f_{S^{t+m-1}}(y)$, will be such that $|f_{S^{t+m-1}}(y) - f_{a_1}(y)| \leq \epsilon$ with probability $1-\epsilon$.*

Proof. See Appendix B ■

We have consigned the proof to the Appendix. However the proof is quite intuitive. After m instances of the same action by the long-lived player, the number of times a particular signal outcome occurs in the sample of the short lived player is a Bernoulli random variable. For a large enough s we can apply the law of large numbers. This shows that the distribution of the short-lived player's sample in m periods time must be "close" to the distribution of the signal outcomes under the action in question.

Recall that the payoffs to all players are bounded, and that there are no ties in player 2's best response function. Then Lemma 4 follows.

Lemma 4 *Suppose that m and s are large enough that Lemma 3 holds for a small ϵ. Consider an arbitrary state at period t. If player 1 were to play a_1 for the following m periods, then with probability $1-\epsilon$, in period $t+m-1$, the ordering of expected payoffs to each action for Player 2 will be the same under $f_{S^{t+m-1}}(y)$ as it is under $f_{a_1}(y)$.*

Proof. See Appendix C. ■

The proof has been confined to the appendix, but the intuition is relatively simple. If m and s are large enough that Lemma 3 holds then with very high probability the sample distribution will be very close to the true distribution from action a_1. So the expected payoff to each action for player 2 cannot be that far apart under the two distributions. We only need them to be sufficiently close that the rankings will be the same.

This would be an appropriate juncture to remind the reader of some definitions:

$$a^* = \arg\max_{a \in E} \pi_1(a),$$
$$a' = \arg\max_{a \in B_2} \pi_1(a),$$
$$\underline{a} = \arg\min_{a \in A} \pi_1(a),$$
$$\overline{a} = \arg\max_{a \in A} \pi_1(a).$$

And for notational simplicity, we will say that:

$$\begin{aligned}\pi_1^e &= \pi_1(a^*),\\ \pi_1^s &= \pi_1(a'),\\ \underline{\pi}_1 &= \pi_1(\underline{a}),\\ \overline{\pi}_1 &= \pi_1(\overline{a}).\end{aligned}$$

Suppose that we are at the beginning of period t and the long-lived player plans to play a_1^* continuously from now on. Assume that Lemmas 3 and 4 both hold. The expectation at t of the best possible expected net present value of payoffs that will accrue from continued play of a_1^* from $t+m$ onwards discounted to $t+m$ will be

$$V_1(a_1^*) = \frac{(1-\epsilon)\pi_1^e + \epsilon\overline{\pi}_1}{1-\delta}. \tag{3.8}$$

Suppose that we are at the beginning of period t and the long-lived player plans to play a_1^s continuously from now on. Assume that Lemmas 3 and 4 both hold. The expectation at t of the worst possible expected net present value of payoffs that will accrue from continued play of a_1^s from $t+m$ onwards discounted to $t+m$ will be

$$V_1(a_1^s) = \frac{(1-\epsilon)\pi_1^s + \epsilon\underline{\pi}_1}{1-\delta}. \tag{3.9}$$

(3.8) and (3.9) describe the best and worst case scenarios respectively. However, what should be clear is that we can always find an ϵ small enough that $V_1(a_1^s) > V_1(a_1^*)$. The following notation will be convenient:

$$V_1^N = (1-\epsilon)\pi_1^e + \epsilon\overline{\pi}_1 \qquad \Rightarrow V_1(a_1^*) = \frac{V_1^N}{1-\delta}, \tag{3.10}$$

$$V_1^S = (1-\epsilon)\pi_1^s + \epsilon\underline{\pi}_1 \qquad \Rightarrow V_1(a_1^s) = \frac{V_1^S}{1-\delta}. \tag{3.11}$$

We briefly define the process of being locked in for games of imperfect monitoring as follows:

Definition 5 *Play has become locked into the play of a particular strategy profile if and only if player 1 would always plays his part in that strategy profile if play continued indefinitely, and the sequence of player 2s play their part in that profile with probability at least* $1-\epsilon$.

We can now prove the following proposition,

Proposition 6 *If*

$$\delta > \left(\frac{\overline{\pi}_1 - \underline{\pi}_1}{(\overline{\pi}_1 - \underline{\pi}_1) + \left(V_1^S - V_1^N\right)} \right)^{\frac{1}{m}} \in (0,1)\,, \tag{3.12}$$

then for large but finite m and s, the equilibrium strategy for a finitely patient player 1 playing adaptive player 2s will not involve being locked into any Nash equilibrium of the stage game.

Proof. See Appendix D. ■

The strategy for the proof of Proposition 6 is similar to the proof of Proposition 2.

Suppose that a sufficiently patient long-lived player plays against short-lived adaptive players with sufficiently large samples. Proposition 6 shows that he can always do better than perpetually playing his part in a Nash equilibrium of the stage game. This means that they must build some kind of reputation. Proposition 6 is not claiming that the best strategy is always to play the Stackleberg action. It simply claims that always playing it is better than always playing any Nash equilibrium.

The only reason to play away from the stage game Nash equilibrium is to manage the expectations of the short-lived players. By doing so the long-lived player hopes to influence their strategy for his own benefit. What is this if not building a reputation.

However there is one sense in which Proposition 6 is weaker than Proposition 2. In both cases, the level of patience required is increasing in m. This is because for larger available histories it takes longer for the long-lived player to move to the Stackleberg convention. So the maximum potential cost of moving is higher. The background to Proposition 6 requires m and s to be relatively large, whereas Proposition 2 places no restrictions on these variables. However we must remember that these are sufficient and not necessary conditions. The costs of moving conventions in terms of foregone payoffs might actually be much smaller than in the calculations of these sufficient conditions. Further, it is highly unlikely that they would have to be paid for all m periods of the move.

3.3 Finite Games

Propositions 6 and 2 were proved for games that continue indefinitely. In both cases, the key observation was that the Stackleberg payoff for the long-lived player was very much higher than that from the best possible Nash equilibrium from the stage game. This meant that it would be better to accept very bad payoffs in the short term in order to enjoy the long term benefits of the higher Stackleberg payoffs. There is nothing crucial about the indefinite nature of the game in this chain of logic. Provided that the game is expected to last long enough that the Stackleberg payoffs net of the associated short term costs exceed the Nash equilibrium payoffs, a reputation will be built.

Hence the following corollary:

Corollary 7 *Propositions 2 and 6 apply equally well to finite games provided they are sufficiently long.*

For simplicity the rest of this paper will be concerned with the case of infinite games of partial but perfect monitoring.

3.4 Knowledgeable Short Lived Players

As a check on robustness, we will consider what happens if one of the short-lived players becomes more aware and more rational. This short-lived player becomes a traditional *homo economicus*. We will assume that the long-lived player becomes aware of this development when the knowledgeable short-lived player is called to move. It is then common knowledge between the long-lived player and the knowledgeable short-lived player. What is also common knowledge between the two of them is that all of the other short-lived players remain the sort of adaptive players we have been talking about.

What the short-lived player does not know is the state of the system when she is called to play. She only knows some subset (of size s) of the last m actions of the long-lived player and she does not know when precisely these took place. This limitation to her knowledge is also common knowledge among the long-lived player and the short-lived player in question. The Knowledgeable short-lived player will have enough information to derive the long-lived player's best response to the adaptive population. Furthermore it is common knowledge among the long-lived player and the knowledgeable short-lived player that this is the case. However what she does not know is exactly where in that cycle she has been called to play. We shall assume that her sample provides no information on this matter so she maintains a uniform prior as to her location in the best response cycle used by the long-lived player.[7]

Consider the following strategy profile. The long-lived player does not change his plans and continues to play the next action in his cycle which constitutes a best response to the population of adaptive players. The knowledgeable short-lived player plays the Stackleberg follower action. This might well constitute an equilibrium strategy profile in the game between the long-lived player and the knowledgeable short-lived player. Suppose there are a large number of instances of the Stackleberg leader action in the long-lived player's best response cycle. Then the Stackleberg follower action may well be the short-lived player's *ex ante* best response. The next action in the long-lived player's best response cycle (whatever it is) might not be a static best response to the Stackleberg follower action. However it may well be a best response bearing in mind the future costs and benefits from today's action turning up in the samples of tomorrows' short-lived players.

[7]This will be true if the length of the best response cycle employed by the long-lived player is m or some $m/\mathbb{N}$. For all cycle lengths $< m$, or not much $> m$ this will be very nearly true. In the event that the cycle length is very much greater than m, Bayesian inferences on the sample drawn might be somewhat informative. However this would not dramatically change the conclusions that are to follow, and so we do not consider the tedious issues of the relevant Lebesgue spaces and σ-algebras here.

This discussion has, of necessity, been somewhat imprecise. However it is difficult to be more precise without putting more structure on the stage game, available history and sample size. This is precisely what we are trying to avoid. However, this discussion should be sufficient to reassure us that the reputational conclusions are robust to the occasional arrival of a knowledgeable short-lived player.

4 Living Off Reputations - An Example

Consider the following example in Table 3. It is another education game similar to that in Table 2, except that the payoffs to the long-lived player have changed.

		Student	
		E	N
College	E	**3** 2	2 0
	N	0 **9**	**1** **1**

		Student	
		E	N
College	E	**3** ?	2 ?
	N	0 ?	**1** ?

Table 3: The stage game for the long and short-lived players.

Imagine the game in Table 3 were to be played as an infinitely repeated stage game of a reputation game. The College takes the place of a long-lived player 1 and the students represent an infinite stream of short lived player 2s. Table 4 then records the equilibrium strategy of the long-lived player 1 under certain parameter assumptions. The left hand column specifies each of the 16 states in turn. The central column records the action that is called for in that state by the best response strategy of the long-lived player. The right hand column then specifies to which state the reputation game would then move.

We assumed an available history of length $m = 4$ and a sample size of $s = 3$. We have assumed $\delta = 0.85$. The stage game is that recorded in Table 3.

Some kind of reputation is clearly being built since E is played from the state $NNNN$. However the long-lived player will not always exert effort. There are many states from which the equilibrium strategy prescribes that the long-lived player should play N, in particular from the state $EEEE$.

We should note that playing N from the state $EEEE$ is costless for the long-lived player. From the state $EEEN$, the worst possible sample of three recent moves that the short-lived player might draw would include two instances of E and one instance of N. The short-lived player would conclude that the probability the long-lived player will play N is only $1/3$ and so play E themselves. So even a perfectly patient player will play N in that state. I refer to this as the "I certainly won't be punished" incentive to cheat.

However that is not the only reason a long-lived player might not keep his reputation perfectly clean. In this example, the benefit to the long-lived player from cheating when

State	Action	Destination
$EEEE$	N	$EEEN$
$EEEN$	N	$EENN$
$EENE$	N	$ENEN$
$EENN$	E	$ENNE$
$ENEE$	N	$NEEN$
$ENEN$	E	$NENE$
$ENNE$	E	$NNEE$
$ENNN$	E	$NNNE$
$NEEE$	N	$EEEN$
$NEEN$	E	$EENE$
$NENE$	E	$ENEE$
$NENN$	E	$ENNE$
$NNEE$	E	$NEEE$
$NNEN$	E	$NENE$
$NNNE$	E	$NNEE$
$NNNN$	E	$NNNE$

Table 4: The best response to Adaptve Players. That this was the best strategy has been shown in an Excel Spreadsheet model which calculates the NPV of each strategy for the long-lived player and finds the unique strategy that gets the highest payoff from every initial state. The file is available from the author on request.

the short lived player is certain to play E is very large. So the long-lived player is willing to tolerate positive probability of the future short-lived players playing N in order to enjoy the benefits of cheating on their reputation more frequently. For example, from the state $EEEN$, the equilibrium strategy of the long-lived player is to play N and move to $EENN$. From here there is positive probability $(1/2)$ that the short-lived player will play N. I think of this as the "I probably won't be punished" incentive to cheat.

We can see from the diagram in Figure 1 that depending on the initial state, the strategy in Table 4 can produce two different cycles through the state space. Over the course of both cycles, the long-lived player oscillates between having one and two instances of not exerting effort in the available memory. Or to put it another way, all of the recurrent states have either 2 or 3 instances of E in the available history and all states with either 2 or 3 instances of E in their available history are recurrent states.

The cycles are deterministic, and once an initial state has been picked out, all future states have effectively been determined. However the play of the game is not fully deterministic. In the states with only 2 instances of the college playing E in the available history, there is a fifty per cent probability that the short-lived player will play N and a fifty per cent probability that the short-lived player will play E.

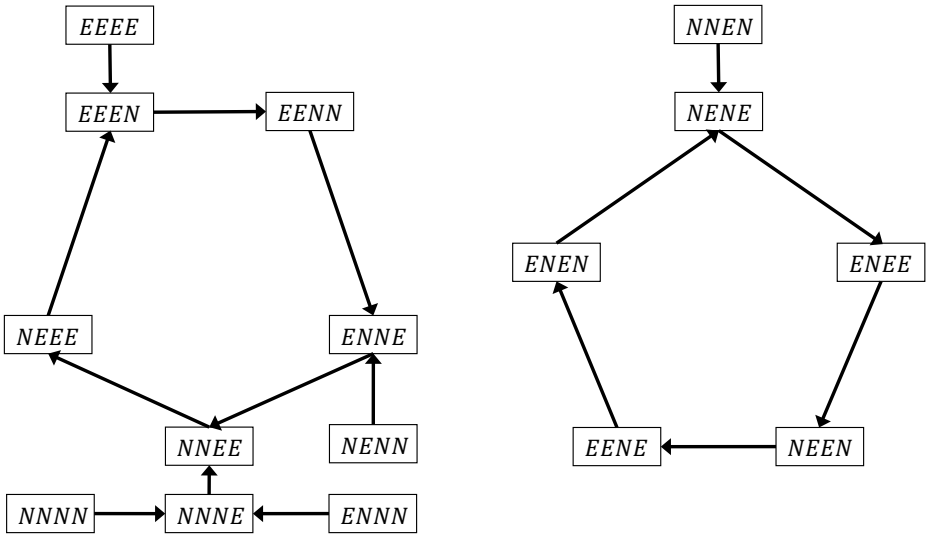

Figure 1: The path through the state space constituting a best response to adaptive players.

5 Living Off Reputations

In order to consider living off reputations more generally, without getting bogged down, we shall restrict attention to reputation games with 2×2 stage games. Consider the general 2×2 stage game in Table 5. Each player has two actions, they can play their Nash action, N or their Stackleberg action T. The payoffs are shown in Table 5 and stage game best responses are in bold.

		S-L player T	S-L player N
L-L player	T	$\mathbf{t}_f$, t_l	a_f, a_l
	N	c_f, $\mathbf{c}_l$	$\mathbf{n}_f$, $\mathbf{n}_l$

Table 5: A general 2×2 stage game of a reputation game.

For the long-lived player we assume that N is a dominant strategy, so that $c_l > t_l$ and $n_l > a_l$. In order to ensure that there is an incentive for the long-lived player to build a reputation, we assume that the long-lived player prefers the outcome {T,T} to the outcome {N,N}. So $t_l > n_l$. This means we can completely order the payoffs to the

long-lived player, $c_l > t_l > n_l > a_l$.

For the short-lived player we assume that the best response to N is N and the best response to T is T so that $n_f > c_f$ and $t_f > a_f$. However we also assume that *ceteris paribus* they benefit from the long-lived player taking the action T, so $t_f > c_f$ and $a_f > n_f$. So we can completely order the payoffs to the short-lived players as $t_f > a_f > n_f > c_f$.

5.1 "I'll never get punished"

Suppose that the short-lived player assessed the probability that the long-lived player would play T as p. Then the expected payoffs from T and N for the short lived player would be given by

$$E\{\pi(T)|p\} = c_f + p(t_f - c_f),$$
$$E\{\pi(N)|p\} = n_f + p(a_f - n_f).$$

So T would be a best response for the short-lived player provided that

$$c_f + p(t_f - c_f) \geq n_f + p(a_f - n_f),$$
$$\Leftrightarrow p \geq \frac{n_f - c_f}{(t_f - c_f) - (a_f - n_f)}. \tag{5.1}$$

The ordering we discussed above ensures that the denominator and numerator of (5.1) are positive. Indeed we can re-write (5.1) as

$$p \geq \alpha = \frac{(n_f - c_f)}{(t_f - a_f) + (n_f - c_f)} \in (0,1). \tag{5.2}$$

Recall that s denotes the sample size of the short-lived players and that m denotes the length of the available history from which that sample is drawn. Let $n^* = \lfloor (1-\alpha)s \rfloor$, the largest integer smaller than or equal to $(1-\alpha)s$. Then n^* is the largest number of incidents of the play of N in the available history that the long-lived player can have without running *any* risk of reputational failure. I define *reputational failure* as being where the long-lived player's record in the sample drawn by the short-lived player is so bad that the short-lived player plays N.

Let $n^t(\omega^t)$ be the number of occasions where N was played in the last m periods in state ω^t. Suppose $n^t(\omega^t) \leq n^*$. Then even if all of plays of N in state ω^t appeared in the sample drawn by the short-lived player, that would not be sufficient to induce reputational failure. So if $n^* > 0$, then some cheating is "safe" from the perspective of the long-lived player. They can cheat on their reputation with no chance of reputational failure.

No matter how small $1-\alpha$ is as a fraction, we know from (5.2) that it must be strictly greater than zero. This means that we can always find an s sufficiently large that $(1-\alpha)s \geq 1$ and so $n^* \geq 1$. This leads us to the following proposition:

Proposition 8 *For any reputation game with a 2×2 stage game, there exists an s' and an $m \geq s'$ large enough that the optimal strategy, even for a perfectly patient long-lived player, is not to play the Stackleberg action all the time.*

Proof. The proposition follows from the fact that if it is possible to costlessly cheat on that reputation, then even a perfectly patient player should do so. And if $s \geq 1/(1-\alpha)$ and $m > s$, then it will be possible to cheat on the reputation costlessly. This has been shown above. ■

The kind of cheating I have described in Proposition 8 is what I referred to earlier as the "I'll never get punished" incentive to cheat. We now want to turn attention to the "I'll probably not get punished" incentive to cheat in this more general setting.

5.2 "I'll probably not get punished"

Suppose that a long-lived player is using the "I'll never get punished" strategy. This strategy will seek to keep the number of instances of N in the available history at n^*. Suppose that we are in period t and the strategy calls for the long-lived player to play E in this period and N in period $t+1$. Consider the deviation to playing N in period t and E in period $t+1$. This will bring forward the benefit of cheating on the reputation by one period. The cost is that there will be n^*+1 occasions in the available history where the long-lived player has played N in period $t+1$. This creates a small chance of reputational failure if *all* of those occasions are drawn in the short-lived player's sample at period $t+1$. For now we will denote this probability β, and we will leave the definition of β for later.

It is only the payoffs over the next two periods that will differ in this deviation. Not deviating earns $t_l + \delta c_l$. Deviating will earn $c_l + \delta\left((1-\beta)\, t_l + \beta a_l\right)$. All subsequent payoffs will be the same. So the deviation will be a profitable one provided that

$$\begin{aligned} c_l + \delta t_l - \delta\beta\left(t_l - a_l\right) &\geq t_l + \delta c_l, \\ \delta\beta\left(t_l - a_l\right) &\leq \left(1-\delta\right)\left(c_l - t_l\right), \\ \Rightarrow \qquad \beta \leq \beta^* &= \left(\frac{1-\delta}{\delta}\right)\left(\frac{c_l - t_l}{t_l - a_l}\right). \end{aligned} \tag{5.3}$$

The two-period deviation from the "I'll never get punished" strategy that we have identified will not necessarily be the best possible deviation. So we have found a sufficient condition for choosing a strategy other than the "I'll never get punished" strategy rather than a necessary one.

The condition in (5.3) places a maximum on β. Recall that β represents the probability of reputational failure that a long-lived player will accept in order to bring forward a point in their strategy where they can live off their reputation. This expression also shows that maximum to be decreasing as the long-lived player becomes more patient, and increasing in the benefit from cheating on that reputation.

The parameter β will be defined by the hypergeometric distribution. This tells us the probability that the short lived player at $t+1$ draws each of the n^*+1 occasions in which the long-lived player played N in her sample of size s. Hence

$$\beta = \frac{\left(\begin{array}{c} n^*+1 \\ n^*+1 \end{array} \right) \left(\begin{array}{c} m-(n^*+1) \\ s-(n^*+1) \end{array} \right)}{\left(\begin{array}{c} m \\ s \end{array} \right)}. \tag{5.4}$$

Where

$$\left(\begin{array}{c} l \\ k \end{array} \right) = \prod_{i=1}^{k} \frac{l-k+i}{i}, \tag{5.5}$$

which is the coefficient on the x^k term in the expansion of $(x+1)^l$. This means that

$$\left(\begin{array}{c} n^*+1 \\ n^*+1 \end{array} \right) = 1.$$

So we can say that

$$\beta = \frac{\left(\begin{array}{c} m-n^*-1 \\ s-n^*-1 \end{array} \right)}{\left(\begin{array}{c} m \\ s \end{array} \right)}. \tag{5.6}$$

Hence, using (5.5),

$$\begin{aligned} \left(\begin{array}{c} m \\ s \end{array} \right) &= \prod_{i=1}^{s} \frac{m-s+i}{i}, \\ \left(\begin{array}{c} m-n^*-1 \\ s-n^*-1 \end{array} \right) &= \prod_{i=1}^{s-n^*-1} \frac{m-n^*-1-s+n^*+1+i}{i}. \\ &= \prod_{i=1}^{s-n^*-1} \frac{m-s+i}{i}. \end{aligned}$$

Hence (5.6) becomes

$$\beta = \frac{\begin{pmatrix} m-n^*-1 \\ s-n^*-1 \end{pmatrix}}{\begin{pmatrix} m \\ s \end{pmatrix}},$$

$$\beta = \frac{\prod_{i=1}^{s-n^*-1} \frac{m-s+i}{i}}{\prod_{i=1}^{s} \frac{m-s+i}{i}},$$

$$\beta = \frac{1}{\prod_{i=s-n^*}^{s} \frac{m-s+i}{i}},$$

$$\beta = \prod_{i=s-n^*}^{s} \frac{i}{m-s+i}. \tag{5.7}$$

As sampling becomes more incomplete, $m-s$ increases and each of the fractions in (5.7) decreases. This will lead the whole of β to decrease as well. For a fixed sample size, as $m \to \infty$, each of the fractions in the product of (5.7) will $\to 0$. So for a fixed s, as $m \to \infty$, $\beta \to 0$. We therefore assert the following proposition:

Proposition 9 *For any reputation game with a fixed 2×2 stage game; and a fixed sample size, s; and a fixed discount factor for the long lived player, δ: there exists an available history, m sufficiently large that the long lived player will accept some positive probability of reputational failure at some point in the cycle that forms his best response to the strategy of the adaptive players.*

Proof. The expression in (5.3) gives the maximum value of β as a sufficient condition for the long-lived player to prefer strategies other than those that will ensure that they avoid reputational failure. This will always be strictly positive. The equation (5.7) shows that for a fixed sample size, as the available history becomes larger, β becomes smaller, and asymptotes towards 0. So for any fixed stage game; fixed discount factor, δ; and sample size, s: there will be some large but finite available history, m such that the probability of being punished for cheating on one's reputation is sufficiently low that the long-lived player will choose to do so. ■

The important thing about Proposition 9 is that it will hold *for any level of δ*. So it will hold even for those long-lived players who are sufficiently patient to build a reputation in the first place.

6 Comparison With Incomplete Information Models

It is worth pausing here to compare this model of reputation with the incomplete information models. For simplicity we will restrict ourselves to one example using the stage

game from Table 3. We will suppose that the stage game is to be repeated indefinitely, and that the discount factor of the rational long-lived player is $\delta = 0.85$.

In the incomplete information case, we will suppose the rational short-lived players have a prior, $\lambda \in (0, 1)$ on the probability that the long-lived player is a commitment type who can only play E. There are only two types of long-lived player, the commitment type and the rational type. The reputation result in this example is the demonstration of a pooling equilibrium where the rational long-lived player always plays E (unless they have played N in the past in which case they play N); and the short lived players always play E unless there exists some past period where the long-lived player played N, in which case they play N themselves.

Given the strategy of the long-lived player, the strategy of the short-lived players is trivially optimal. Treat the strategies of the short-lived players as given. The rational long lived player can get a Net Present Value discounted payoff of $3/(1-\delta)$ by following the proposed equilibrium strategy we have described. Or they can deviate by playing N for the first time and getting $9+\delta/(1-\delta)$. The deviation can be shown to be unprofitable for any $\delta \geq 0.75$, and is therefore unprofitable at $\delta = 0.85$. Trivially the commitment type of long-lived player has no possible deviation.

Now consider the partial information case outlined in Section 4. Once again we will assume $m = 4$ and $s = 3$. Then the best response of the long-lived players to adaptive play by the short-lived players is the Markov strategy outlined in Table 4. This induces the paths of play outlined in Figure 1.

So the reputation model that uses incomplete information will see reputation as a pooling equilibrium in which the rational long-lived player imitates a committed long-lived player. Once lost, such a reputation can never be recovered. However by viewing reputation through the lens of partial information and adaptive play by short-lived players, we can produce reputations that are run down and built up over an equilibrium cycle. Such a view allows for more natural interpretation of reputation as an asset that is built up and consumed out of. If we model reputations through incomplete information and try to interpret reputations as an asset, then the metaphor will become muddled. We will have to think of a very special kind of asset that forms a natural endowment (the uncertainty and the expectation of a pooling equilibrium). This asset can only be destroyed; can never be rebuilt; and is costly to maintain.

7 Dynamic Stage Games

Before concluding, we should consider the situation where the long-lived player's strategy might be unobservable. In particular where the stage game itself has a dynamic element, such as in the entry deterrence game from Selten (1978), reproduced in extensive form in Figure 2.

The dynamic structure and payoffs of this stage game are shown in Figure 2. We imagine

the incumbent as the long-lived player 1 and entrants form the sequence of short-lived player 2s. Suppose all the entrants in the previous periods sampled by today's entrant have played OUT. Then today's entrant has no data on which to base an expectation of the action the incumbent would take should the entrant play IN.

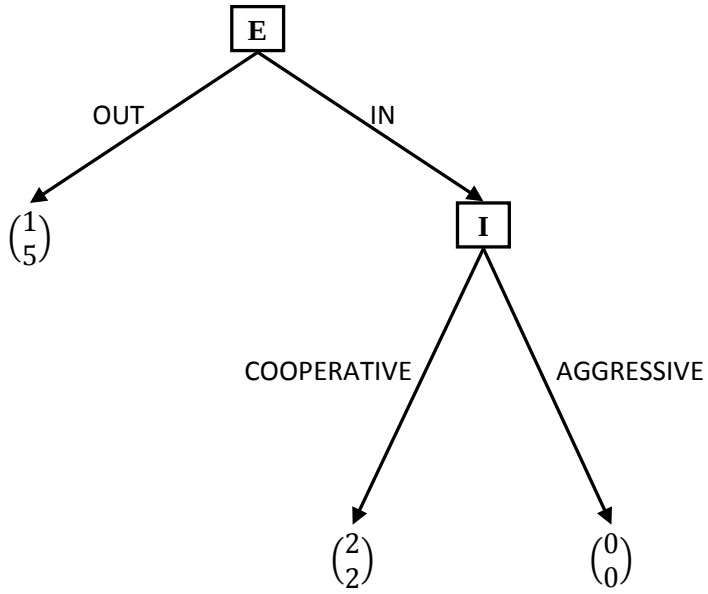

Figure 2: Selten's entry deterrence game.

However, we must take it as a given that the only actions a player in a game can take are the actions that they know about. This raises the question of where they find out about these actions. One entirely plausible and logical source is the actions of people who have been in a similar position in the past. In that case, once an incumbent has established a reputation for being aggressive towards entrants who enter, entry will become something that simply "is not done". It is almost as if potential entrants forget about their ability to enter. In such a situation forgotten actions might occasionally be rediscovered by some player willing to experiment.[8]

[8]The legal case of Ashford v Thornton (1818) provides a nice example of forgotten actions being rediscovered. After Thornton was acquitted of Mary Ashford's murder, her brother, William launched a private appeal to the King's Bench. At the King's Bench, Thornton claimed the right to trial by combat, a legal right which (as far as anyone can find) had last been used in England in 1446. Although appeals to the King's Bench at this time were rare they were not unheard of, and there are doubtless others who

Suppose that an incumbent has cooperated in response to entry for as long as anyone can remember. Suppose also that there are $T \gg m$ periods of the repeated game left in each of which the incumbent will face a potential entrant. Suppose that once an action has been forgotten, the probability that it will be re-discovered is negligible.

An entrant who is aware of his options will stay out so long as the probability of an aggressive response is greater than or equal to 1/2. Suppose that the incumbent switches to a strategy of always responding aggressively to entry from time t onwards. The payoff this gives to the incumbent must be strictly greater than $5\delta^m \left(1-\delta^{T-m}\right)/(1-\delta)$. By contrast continuing to respond cooperatively to entry will get $2\left(1-\delta^T\right)(1-\delta)$. So it will be profitable to switch to aggressive responses to entry provided that:

$$\begin{aligned}
\frac{5\delta^m\left(1-\delta^{T-m}\right)}{1-\delta} &\geq \frac{2\left(1-\delta^T\right)}{1-\delta}, \\
5\delta^m\left(1-\delta^{T-m}\right) &\geq 2\left(1-\delta^T\right), \\
5\delta^m - 5\delta^T &\geq 2 - 2\delta^T, \\
5\delta^m - 3\delta^T &\geq 2. \qquad (7.1)
\end{aligned}$$

It isn't possible to solve this generally however we can examine a parameterized example more closely. Take the example where $T = 20$ and $m = 5$.[9] Then we can show numerically that (7.1) will be satisfied for any $\delta \geq 0.8401$.

In the limit as $T \to \infty$, we will be able to find values of δ which will satisfy (7.1) as $\delta^T \to 0$ and (7.1) collapses to $\delta \geq (2/5)^{1/m}$.

8 Conclusion

Economists traditionally consider reputation as the explanation for actions which forgo the opportunity to reap benefits at the expense of others. By forgoing such opportunities, we encourage others to trust us and the benefits of that trust, over the long term, exceed the opportunity cost of not cheating people.

However models of this process which rely on incomplete information and the possibility of commitment types can be problematic. They lead to results where reputation can be instantly enjoyed and is then kept perfectly forever, or where reputation is built up and then lost forever. However we often think of reputation as an intangible asset. It is a strange asset indeed that cannot be brought, sold, built up and run down in any particular order. Furthermore, in order to build such models we must imbue transitory players with a lot of information about past actions and impressive computational abilities to interpret those actions. In reality the effort required to gather the knowledge of

would have benefited from lawyers as studious as Mr. Thornton's. In 1819 Parliament abolished private appeals to the King's Bench and with it, the right to trial by combat.

[9] $T = 20$ is following Selten (1978).

past outcomes and extract as much information as possible from them may well exceed any possible gains and losses from the interaction itself.

The model of reputation I have put forward here does not require such impressive knowledge on the part of the transitory actors. Neither does it require them to deploy such impressive computational abilities. My model also produces outcomes that are more in-line with the reputations we observe in the real world. We see the reputations of politicians and political parties falling and recovering in the opinion polls; we see the brand values of particular firms improve and depreciate as they invest in them and consume out of them; and we see out own opinions of our friends and colleagues wax and wane (and worry about whether the same happens to their opinions of us).

This model also presents tantalizing prospects for further research. The key insight is that the information structure is crucial to a reputation game. By information structure, I mean the way in which the experiences of past short-lived players are communicated to the current short-lived player. This is what creates the link for the long-lived player between his current behavior and future rewards. For example we might think that word of mouth communication which often carries reputation should be modeled through a game played on a network rather than random sampling. If the long-lived player knew the network structure and the place occupied by his current opponent, he might condition his action on the number of people likely to find out about it. A well connected short-lived player will then be more likely to receive good treatment than her poorly connected counterparts.[10] This model is also well equipped to look at the consequences of the perception that bad news hangs around longer. We could build this into the model by biasing the sampling process toward bad news.

Thinking about reputation in this way could also further the economic understanding of advertising and public relations. These can be seen as attempts to influence the samples that short-lived players draw from the available history. This model would be particularly suited to thinking of a specific kind of advertising campaign which industry practitioners refer to as the "poster child" campaign. The strategy is often used to solicit donations to charitable causes, particularly those combating disease. The example of Jolene Worley as the first National Muscular Dystrophy Poster Child is a case in point.[11] Similar campaigns are also sometimes used to reduce the stigma associated with some diseases. Firms who invite prospective employees to meet current employees may well have similar motives. The firm controls which employees are met and can ensure that only good things are heard. Of course such a strategy would rely on the naiveity of the prospective employees (or at least some of them), or at the very least a lack of access to other sources of information.

This model provides a way of considering aspects of reputation which previous models have not emphasized. The cost of doing so is that we are no longer considering the short-lived players to be rational actors in the traditional economic sense. However it

[10] This point demonstrates the wisdom of the adage "It's who you know, not what you know."

[11] See her obituary at http://community.seattletimes.nwsource.com/archive/?date=19930422slug=1697184.

is intuitive that people engaged in the interaction once and only once would limit the resources they devote to calculating their best strategy.

References

Bar-Isaac, H. and S. Tadelis (2008), *Seller reputation*, Now Pub.

Blackwell, D. (1962), 'Discrete dynamic programming', *The Annals of Mathematical Statistics* **33**(2), 719–726.

Cripps, Martin W., George J. Mailath and Larry Samuelson (2004), 'Imperfect monitoring and impermanent reputations', *Econometrica* **72**(2), 407–432.
URL: *http://www.jstor.org/stable/3598908*

Ekmekci, Mehmet and Andrea Wilson (2005), Sustainable reputations with rating systems. mimeo.

Ekmekci, Mehmet and Andrea Wilson (2007), Reputation games with finite automata. mimeo.

Ellison, Glenn (1997), 'Learning from personal experience: One rational guy and the justification of myopia', *Games and Economic Behavior* **19**(2), 180 – 210.
URL: *http://www.sciencedirect.com/science/article/B6WFW-45M8X0Y-P/2/250bc8461dc299d6bd8fb0a38147366b*

Fudenberg, Drew and David K Levine (1989), 'Reputation and equilibrium selection in games with a patient player', *Econometrica* **57**(4), 759–78.
URL: *http://ideas.repec.org/a/ecm/emetrp/v57y1989i4p759-78.html*

Fudenberg, Drew and David K. Levine (1992), 'Maintaining a reputation when strategies are imperfectly observed', *The Review of Economic Studies* **59**(3), 561–579.
URL: *http://www.jstor.org/stable/2297864*

Fudenberg, Drew and David K. Levine (1993), 'Steady state learning and nash equilibrium', *Econometrica* **61**(3), pp. 547–573.
URL: *http://www.jstor.org/stable/2951717*

Green, Edward J. and Robert H. Porter (1984), 'Noncooperative collusion under imperfect price information', *Econometrica* **52**(1), pp. 87–100.
URL: *http://www.jstor.org/stable/1911462*

Kreps, David M. (1990), Corporate culture and economic theory, *in* J. E.Alt and K. A.Shepsle, eds, 'Perspectives on Positive Political Economy', Cambridge University Press, Cambridge, UK.

Kreps, David M., Paul Milgrom, John Roberts and Robert Wilson (1982), 'Rational cooperation in the finitely repeated prisoners' dilemma', *Journal of Economic Theory* **27**(2), 245–252.
URL: *http://ideas.repec.org/a/eee/jetheo/v27y1982i2p245-252.html*

Kreps, D.M. and R. Wilson (1982), 'Reputation and imperfect information', *Journal of economic theory* **27**(2), 253–279.

Liu, Qingmin (2010), Information acquisition and reputation dynamics. mimeo.

Liu, Qingmin and Andrzej Skrzypacz (2011), Limited records and reputation. mimeo.

Mailath, George J. and Larry Samuelson (2001), 'Who wants a good reputation?', *The Review of Economic Studies* **68**(2), pp. 415–441.
URL: *http://www.jstor.org/stable/2695935*

Milgrom, Paul and John Roberts (1982), 'Predation, reputation, and entry deterrence', *Journal of Economic Theory* **27**(2), 280 – 312.
URL: *http://www.sciencedirect.com/science/article/B6WJ3-4CYGCVT-10S/2/03366ab7e493bf276847007975f9c86a*

Monte, Daniel (2007), Reputation and bounded memory: A theory of inertia and skepticism. mimeo.

Phelan, Christopher (2006), 'Public trust and government betrayal', *Journal of Economic Theory* **130**(1), 27 – 43.
URL: *http://www.sciencedirect.com/science/article/B6WJ3-4HC6M65-1/2/e0e69ce6bab4de25cdeb978643c9084c*

Selten, Reinhard (1978), 'The chain store paradox', *Theory and Decision* **9**, 127–159. 10.1007/BF00131770.
URL: *http://dx.doi.org/10.1007/BF00131770*

Shafir, Sharoni, Taly Reich, Erez Tsur, Ido Erev and Arnon Lotem (2008), 'Perceptual accuracy and conflicting effects of certainty on risk-taking behaviour', *Nature* **453**, 917–920.

Wiseman, Thomas (2008), 'Reputation and impermanent types', *Games and Economic Behavior* **62**(1), 190 – 210.
URL: *http://www.sciencedirect.com/science/article/B6WFW-4NJ7WBD-2/2/a687afe7e0824b8255cb71eae64f6269*

Young, H. Peyton (1993), 'The evolution of conventions', *Econometrica* **61**(1), 57–84.
URL: *http://www.jstor.org/stable/2951778*

Young, H. Peyton (1998), *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*, Princeton University Press, Princeton, NJ.

A Proof of Proposition 2

Again we should remind ourselves of some of the definitions of important action profiles and their associated payoffs:

$$\begin{aligned}
a^* &= \arg\max_{a\in E} \pi_1(a), \\
a' &= \arg\max_{a\in B_2} \pi_1(a), \\
\underline{a} &= \arg\min_{a\in A} \pi_1(a), \\
\overline{a} &= \arg\max_{a\in A} \pi_1(a), \\
\pi_1^e &= \pi_1(a^*), \\
\pi_1^s &= \pi_1(a'), \\
\underline{\pi}_1 &= \pi_1(\underline{a}), \\
\overline{\pi}_1 &= \pi_1(\overline{a}).
\end{aligned}$$

We also restate Proposition 2 below.

Proposition 2 *If*

$$\delta > \left(\frac{\pi_1^s - \underline{\pi}_1}{\pi_1^e - \underline{\pi}_1}\right)^{\frac{1}{m}} \in (0,1), \tag{3.4}$$

Then the optimal strategy for a sufficiently patient player 1,will not involve being locked into any Nash equilibrium of the stage game.

Proof. Suppose that player 1 is in the state ω', where she has played her part in the best possible Nash equilibrium profile for the last m periods. The expected net present value of continuing to play her part in this Nash equilibrium forever will be $\pi_1^e/(1-\delta)$.

Suppose that player 1 is in the state ω'' where she has played her Stackleberg leader action for the last m periods. Her expected net present value of continuing to play her Stackleberg leader action will be $\pi_1^s/(1-\delta)$.

From the state ω', she can move to the state ω'' by playing her Stackleberg leader action for m periods. The worst possible net present value of payoff she would get over those m periods would be $\underline{\pi}_1 + \delta\underline{\pi}_1 + ... + \delta^{m-1}\underline{\pi}_1$. So from the state ω', the net present value of the strategy of continuing to play the Nash equilibrium action will be: $\pi_1^e/(1-\delta)$. The net present value of the strategy of continually playing the Stackleberg action will be at least

$$\left(\frac{1-\delta^m}{1-\delta}\right)\underline{\pi}_1 + \frac{\delta^m \pi_1^s}{1-\delta}.$$

So player 1 will find it best to play the Stackleberg action provided that:

$$\begin{aligned}
\frac{1-\delta^m}{1-\delta}\underline{\pi}_1 + \frac{\delta^m\pi_1^s}{1-\delta} &\geq \frac{\pi_1^e}{1-\delta}, \\
(1-\delta^m)\underline{\pi}_1 + \delta^m\pi_1^s &\geq \pi_1^e, \\
\delta^m(\pi_1^s - \underline{\pi}_1) &\geq \pi_1^e - \underline{\pi}_1, \\
\delta^m &\geq \frac{\pi_1^e - \underline{\pi}_1}{\pi_1^s - \underline{\pi}_1}.
\end{aligned}$$

Which becomes

$$\delta \geq \left(\frac{\pi_1^e - \underline{\pi}_1}{\pi_1^s - \underline{\pi}_1} \right)^{\frac{1}{m}} \in (0,1), \tag{3.4}$$

as required. ■

B Proof of Lemma 3

We start by re-stating the Lemma.

Lemma 3 *Suppose that we are at the beginning of period t and the long-lived player plans to play the same action $a_1 \in A_1$ for the next m periods. $\forall \epsilon > 0$, there exists an $\overline{s}$ sufficiently large but finite such that, for $m > s > \overline{s}$; and $\forall y \in Y$, the proportion of occasions in which y is expected to appear in the short-lived player's sample in m periods time, denoted by $f_{S^{t+m-1}}(y)$, will be such that $|f_{S^{t+m-1}}(y) - f_{a_1}(y)| \leq \epsilon$ with probability $1-\epsilon$.*

Proof. For each $y \in Y$, let z_y be a Bernoulli random variable such that $z_y^t = 1$ when $y^t = y$ and $z_y^t = 0$ otherwise. So the success probability of this random variable is $f_{a_1}(y)$ where a_1 is the action player 1 will play for the next m periods. We can treat the next m periods for which the long-lived player will play the action a_1 as m i.i.d random Bernoulli experiments. So we can treat the s periods that will be sampled from the available history without replacement in m periods time as s i.i.d random Bernoulli experiments.

Suppose we are in period $t+m-1$. Let S^{t+m-1} be the set of past periods $t+m-1-k$, where $k \leq m-1$, which are in the short-lived player's sample. Now for each $y \in Y$, let x_y^{t+m-1} be the Binomial random variable such that

$$x_y^{t+m-1} = \sum_{t+m-1-k \in S} z_y^{t+m-1-k}. \tag{B.1}$$

So we can say that the proportion of occasions on which y appears in the sample will be given by $f_{S^{t+m-1}}(y) = x_y^{t+m-1}/s$. By the Weak Law of Large Numbers, we know that $\forall \epsilon_y > 0$,

$$\lim_{s \to \infty} \Pr\left(x_y^{t+m-1} - f_{a_1}(y) < \epsilon_y \right) = 1. \tag{B.2}$$

We also know that the variance of x_y^{t+m-1}/s is given by $f_{a_1}(y)(1 - f_{a_1}(y))/s$, which is clearly decreasing in s. The probability that the random variable x_y^{t+m-1}/s is any fixed distance from its true mean must be increasing in the random variable's variance. So we can say that for any fixed ϵ_y, $\Pr\left(\frac{x_y^{t+m-1}}{s} - g_{a_1}(y) < \epsilon_y \right)$ is non-decreasing in s. This, combined with the result from the Weak Law of Large Numbers means that for any ϵ_y, there must be some large, but finite s^* such that:

$$\Pr\left(\frac{x_y^{t+m-1}}{s} - f_{a_1}(y) < \epsilon_y \right) \geq 1 - \epsilon_y \qquad \forall s > s^* \tag{B.3}$$

(B.3) will be true $\forall y \in Y$. The large but finite s^* clearly requires a larger, but still finite $m^* \geq s^*$ from which to draw the sample.

For a fixed large but finite s^*, let $\epsilon = \max_{y \in Y} \epsilon_y$. Then we can say that the distribution of signals in the sample is ϵ-close to the distribution $f_{a_1}(y)$ since the probability that the frequency of any particular signal in the available history is more than ϵ away from $f_{a_1}(y)$ is greater than or equal to $1-\epsilon$. ■

C Proof of Lemma 4

We start by re-stating the Lemma.

Lemma 4 *Suppose that m and s are large enough that Lemma 3 holds for a small ϵ. Consider an arbitrary state at period t. If player 1 were to play a_1 for the following m periods, then with probability $1-\epsilon$, in period $t+m-1$, the ordering of expected payoffs to each action for Player 2 will be the same under $f_{S^t}(y)$ as it is under $f_{a_1}(y)$.*

Proof. The expected payoffs to Player 2 from any particular action $a_2 \in A_2$ if she knew that Player 1 was going to play $a_1 \in A_1$, would be:

$$E_{f_{a_1}(y)}\{\pi_2(y,a_2)\} = \sum_{y\in Y} f_{a_1}(y)\,\pi_2(y,a_2)\,. \tag{C.1}$$

The expected payoffs under the sample S^t will be:

$$E_{f_{S^{t+m-1}}(y)}\{\pi_2(y,a_2)\} = \sum_{y\in Y} f_{S^{t+m-1}}(y)\,\pi_2(y,a_2)\,. \tag{C.2}$$

So the difference in expected payoffs must be:

$$E_{f_{S^{t+m-1}}(y)}\{\pi_2(y,a_2)\} - E_{f_{a_1}(y)}\{\pi_2(y,a_2)\} = \sum_{y\in Y} \left(f_{S^{t+m-1}}(y) - f_{a_1}(y)\right)\pi_2(y,a_2)\,. \tag{C.3}$$

By Lemma 3, with probability $1-\epsilon$, the difference $|f_{S^{t+m-1}}(y) - f_{a_1}(y)| < \epsilon$ for all of the signal outcomes $y \in Y$.

The finite nature of the sample bounds the smallest possible ϵ away from zero. However we also know from the Law of Large Numbers that the smallest possible $\epsilon \to 0$ as $s \to \infty$. Hence there is some large, but finite sample size s' such that the smallest possible ϵ, is small enough that these differences in expected payoffs will be so small that the ordering of actions according to expected payoffs will not change. ■

D Proof of Proposition 6

We start by re-stating the proposition.

Proposition 6 *If*

$$\delta > \left(\frac{\overline{\pi}_1 - \underline{\pi}_1}{(\overline{\pi}_1 - \underline{\pi}_1) + \left(V_1^S - V_1^N\right)}\right)^{\frac{1}{m}} \in (0,1)\,, \tag{3.12}$$

then for large but finite m and s, the equilibrium strategy for a finitely patient player 1 playing adaptive player 2s will not involve being locked into any Nash equilibrium of the stage game.

Proof. From any arbitrary state (including all those that could conceivably result from having played a_1^* for the last m periods), the best possible payoff that could result from continuous play of a_1^* is

$$\overline{\pi}_1 + \delta\overline{\pi}_1 + ... + \delta^{m-1}\overline{\pi}_1 + \delta^m V_1(a_1^*) = \left(\frac{1-\delta^m}{1-\delta}\right)\overline{\pi}_1 + \delta^m V_1(a_1^*)\,. \tag{D.1}$$

From any arbitrary state, the worst possible payoff that could result from continuous play of a_1^s is

$$\underline{\pi}_1 + \delta\underline{\pi}_1 + ... + \delta^{m-1}\underline{\pi}_1 + \delta^m V_1\left(a_1^s\right) = \left(\frac{1-\delta^m}{1-\delta}\right)\underline{\pi}_1 + \delta^m V_1\left(a_1^s\right). \tag{D.2}$$

It is better to continuously play the Stackleberg action if the expression in (D.2) is greater than that in (D.1).

$$\begin{aligned}
\left(\frac{1-\delta^m}{1-\delta}\right)\underline{\pi}_1 + \delta^m V_1\left(a_1^s\right) &> \left(\frac{1-\delta^m}{1-\delta}\right)\overline{\pi}_1 + \delta^m V_1\left(a_1^*\right), \\
\left(\frac{1-\delta^m}{1-\delta}\right)\underline{\pi}_1 + \frac{\delta^m V_1^S}{1-\delta} &> \left(\frac{1-\delta^m}{1-\delta}\right)\overline{\pi}_1 + \frac{\delta^m V_1^N}{1-\delta}, \\
\delta^m\left(\left(\overline{\pi}_1 - \underline{\pi}_1\right) + \left(V_1^S - V_1^N\right)\right) &> \overline{\pi}_1 - \underline{\pi}_1,
\end{aligned}$$

And so:

$$\begin{aligned}
\delta^m &> \frac{\overline{\pi}_1 - \underline{\pi}_1}{\left(\overline{\pi}_1 - \underline{\pi}_1\right) + \left(V_1^S - V_1^N\right)}, \\
\delta &> \left(\frac{\overline{\pi}_1 - \underline{\pi}_1}{\left(\overline{\pi}_1 - \underline{\pi}_1\right) + \left(V_1^S - V_1^N\right)}\right)^{\frac{1}{m}} \in (0,1)
\end{aligned} \tag{3.12}$$

Since we have worked from any arbitrary state, this must include every state that could prevail after the long-lived player has played a_1^* m times.

This completes the proof. ■