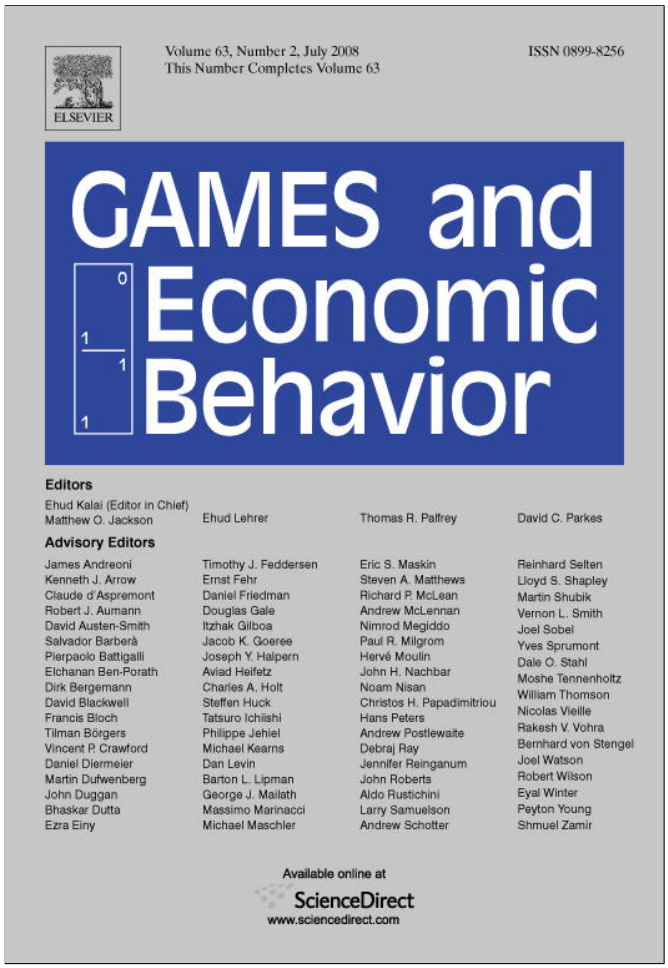

Available online at www.sciencedirect.com
ScienceDirect
Games and Economic Behavior 63 (2008) 498–526
GAMES and Economic Behavior
www.elsevier.com/locate/geb

When is reputation bad?

Jeffrey Ely [a], Drew Fudenberg [b,*], David K. Levine [c]

[a] *Department of Economics, Northwestern University, IL 60208, USA*
[b] *Department of Economics, Harvard University, Cambridge, MA 02138, USA*
[c] *Department of Economics, UCLA, CA, USA*

Received 18 April 2005

Available online 17 October 2006

Abstract

In traditional reputation models, the ability to build a reputation is good for the long-run player. In [Ely, J., Valimaki, J., 2003. Bad reputation. NAJ Econ. 4, 2; http://www.najecon.org/v4.htm. Quart. J. Econ. 118 (2003) 785–814], Ely and Valimaki give an example in which reputation is unambiguously bad. This paper characterizes a class of games in which that insight holds. The key to bad reputation is that participation is optional for the short-run players, and that every action of the long-run player that makes the short-run players want to participate has a chance of being interpreted as a signal that the long-run player is "bad." We allow a broad set of commitment types, allowing many types, including the "Stackelberg type" used to prove positive results on reputation. Although reputation need not be bad if the probability of the Stackelberg type is too high, the relative probability of the Stackelberg type can be high when all commitment types are unlikely.

JEL classification: C72; D82

Keywords: Game theory; Reputation; Stackelberg; Commitment

1. Introduction

A long-run player playing against a sequence of short-lived opponents can build a reputation for playing in a specific way and so obtain the benefits of commitment power. To model these "reputation effects," the literature following Kreps and Wilson (1982) and Milgrom and Roberts

* Corresponding author.
E-mail address: dfudenberg@harvard.edu (D. Fudenberg).

0899-8256/$ – see front matter
doi:10.1016/j.geb.2006.08.007

(1982) has supposed that there is positive prior probability that the long-run player is a "commitment type" who always plays a specific strategy.[1] In "Bad reputation," Ely and Valimaki (2003) (henceforth EV) construct an example in which introducing a particular commitment type hurts the long-run player. When the game is played only once and there are no commitment types, the unique sequential equilibrium is good for the long-run player. This remains an equilibrium when the game is repeated without commitment types, regardless of the player's discount factor. However, when a particular "bad" commitment type is introduced, the only Nash equilibria are "bad" for a patient long-run player.[2]

In the EV example, the short-run players' only decision action is whether to "enter" (trust the long-run player) or stay out; if they stay out their payoff and the public signals are both independent of the long-run player's strategy or type. The essence of the EV example is that the short-run players choose to exit unless they anticipate a "friendly" action by the long-run player, and that the friendly actions have a positive probability of generating signals that suggest the long-run player is unfriendly. If the long-run player is patient, he may be tempted to change his play to avoid generating these "bad signals," but the actions that reduce the probability of bad signals may themselves be unfriendly.

The EV example is intriguing but it leaves open many questions about what conditions are needed for the bad reputation result. This paper extends the analysis of bad reputation in two important ways. First, EV consider only distributions with two types; we extend the analysis to general distributions over all "commitment types." The bad-reputation result does not hold for all such distributions, but we show it does hold provided that the probability of the "Stackelberg type" is sufficiently low. This finding requires a very different method of analysis than in EV, as it requires keeping track of the relative probability of various commitment types.

Second, this paper extends the ideas in EV to a more general class of games, allowing for multiple short-run (SR) players, multiple signals, many actions, more general payoff functions, and a more general signal structure. In particular the SR players can get signals of the payoffs of previous SR players, which may be important in applications. These extensions allow us to more clearly identify the properties that are needed for the bad reputation result. There are several such properties, notably that the short-run players can either individually or collectively choose to exit. Also, the result requires that either exit is the minmax outcome for the long run player or that the signals of the long run player's action have only two possible values; the EV example had both of these properties. Games which satisfy our conditions are called *bad reputation games*, and we show that in these games any equilibrium payoff of a sufficiently patient long-run player is close to his value from exit.

The main substantive assumption in this paper is that the actions and signals satisfy an extension of the "friendly/unfriendly" dichotomy mentioned above. Loosely speaking, we require that there is a set of "friendly" actions that must receive sufficiently high probability in order to induce entry by the short-run players, a disjoint set of unfriendly actions, and a set of "bad signals"

[1] See Sorin (1999) for a recent survey of the reputation effects literature, and its relationship to the literature on merging of opinions.

[2] It is obvious that incomplete information about the long-run player's type can be harmful when the long-run player is impatient, since incomplete information can be harmful in one-shot games. Fudenberg and Kreps (1987) argue that a better measure of the "power of reputation effects" is to hold fixed the prior distribution over the reputation-builder's types, and compare the reputation-building scenario to one in which the reputation builder's opponents do not observe how the reputation builder has played against other opponents. They discuss why reputation effects might be detrimental in the somewhat different setting of a large long-run player facing many simultaneous small, long-run, opponents.

that are more likely under the unfriendly actions. Finally, each friendly action must be vulnerable to temptation, in the sense that some other action would decrease the probability of all of the bad signals and increase the probability of all other signals. This condition is never satisfied by games of perfect monitoring, and our result gives only a very weak bound when monitoring is almost perfect.

In addition to extending the applicability of the bad reputation result, these extensions allow us to better understand the demarcation between "bad" and "good" reputation. To further contribute to this understanding, we provide a more thorough explanation of how the EV conclusions relate to past work on reputation effects. Reputation effects are most powerful when the long-run player is very patient, and Fudenberg and Levine (1992) (FL) provided upper and lower bounds on the limiting values of the equilibrium payoff of the long-run player as that player's discount factor tends to 1. The upper bound corresponds to the usual notion of the "Stackelberg payoff." The lower bound, called the "generalized Stackelberg payoff," weakens this notion to allow the short-run players to have incorrect beliefs about the long-run player's strategy, so long as the beliefs are not disconfirmed by the information that the short-run players observe. When the stage game is a one-shot simultaneous-move game, actions are observed, payoffs are generic, and commitment types have full support, these two bounds coincide,[3] so that the limit of the Nash equilibrium payoffs as the long-run player's discount factor tends to one is the single point corresponding to the Stackelberg payoff. For extensive-move stage games, with public outcomes corresponding to terminal nodes, the bounds can differ. However, although FL provided examples in which the lower bound is attained, in those examples the upper bound was attained as well, and we are not aware of past work that determines the range of possible limiting values for a fairly general class of games.

We first present a number of examples to illustrate the results in the paper.

Illustrative examples

The basic setting is that of a repeated game between a long-run player and one or more "dynasties" of short-run players. Play in each period's *stage game* generates public signals that are observed by subsequent players.

We call the stage games *participation games* if they have the following extensive form. First, the short-run players simultaneously decide whether to exit or to participate in an interaction with the long-run player. The rules for relating the individual participation decisions to whether an interaction takes place can be arbitrary: individual short-run players may have veto power; either over exit or participation, or there can be majority rule, and so forth. If the short-run players have chosen to exit, the stage game ends and the payoffs and the public signals are unaffected by any choice made by the long-run player (although they may depend on which exit actions the short-run players selected.) When the short-run players participate, the long-run player's action potentially affects the payoffs of all players in the stage, and generates a public signal conveying information about the action choices of all players.

The original EV example is a participation game: There is a single short-run player who is a motorist with a car in need of repair. Participation means hiring the mechanic in which case the mechanic performs a diagnosis and makes a repair. The repair is publicly observed, but the diagnosis is not. More generally, even when entry occurs the short-run players need not perfectly

[3] For this result there must be types that play certain mixed strategies.

observe the long-run player's choice of action. Exit means choosing not to hire the mechanic; when this occurs future short-run players observe only this exit decision and not the repair that would have been chosen.

We identify conditions on the payoffs and information structure of participation games that make them susceptible to bad reputation; we call these "bad reputation games." Bad reputation games generalize the properties of the EV game but have application to stage games with quite different structure. Our results also reveal the boundary between games with good and bad reputations.

EV with a Stackelberg type

EV assumed that there were just two types of mechanic: rational and bad. The bad mechanic always replaces the engine, whether or not a replacement is required. The bad reputation effect arises because the rational type distorts his choice of repairs in order to separate himself from the bad type. On the other hand, traditional reputation arguments are built upon the assumption that the long-run player is potentially a Stackelberg type (the type committed to providing the best service.) It is therefore natural to investigate a model with a variety of types, including Stackelberg and other potentially good types. Our more general framework can extend EV to allow for general type distributions and our results imply that the bad reputation effect persists if the overall distribution of types satisfies the condition illustrated in Fig. 1.

In Fig. 1, the simplex represents the set of probability distributions over the three types, rational, bad, and Stackelberg. In region A, the probability of the bad type is too high and the mechanic is never hired in any equilibrium. On the other hand, we show that in region B, the probability of the Stackelberg type is high enough to ensure existence of a good equilibrium in

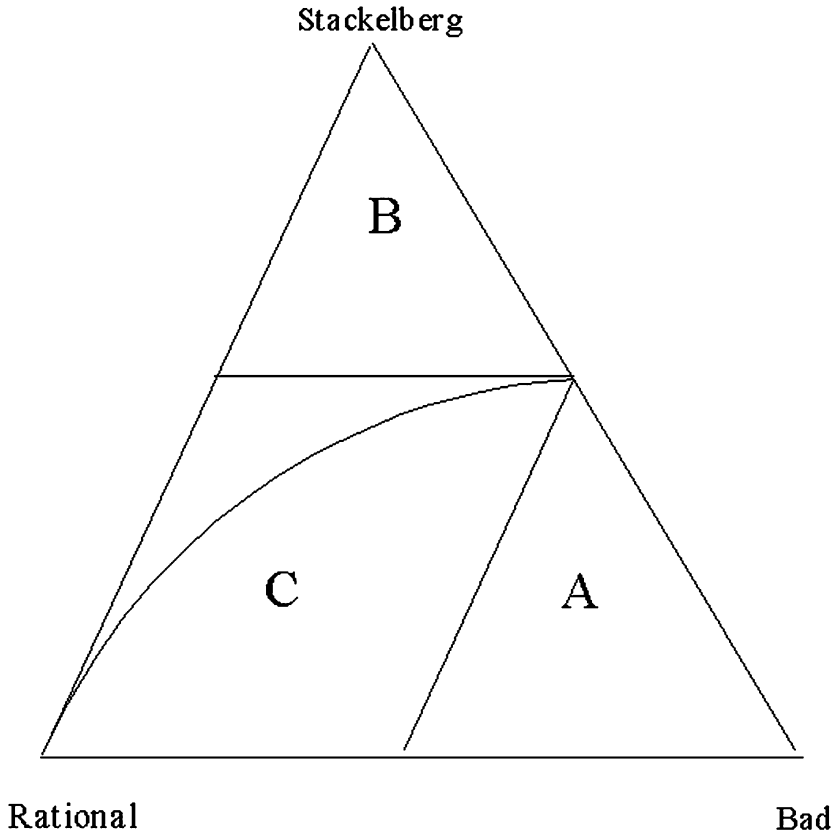

Fig. 1. Distribution of types.

which the rational type obtains his Stackelberg payoff. The analysis is more interesting in the remaining region. We show that there is a curve whose shape is represented in the figure, such that the bad reputation effect occurs when the prior falls below this curve (region C). As illustrated in the figure, the left boundary of the simplex is tangent to this curve as it approaches the lower left corner, where the long-run player is certain to be rational. Thus, bad reputation holds for any generic perturbation of the complete information game. We discuss this issue at more length in Section 4.1.

Observing payoffs

EV assumed that the short-run players observe only the past repairs chosen by the mechanic, and receive no information about the success of the repair. Our results apply to a more general class of information structures: in particular we can allow motorists to observe not only past repairs but also information about their success. We show in Section 6 that the EV conclusion goes through unchanged when motorists observe an arbitrarily precise but imperfect signal of the realized payoffs of all previous motorists; the key is that observing the realized payoffs of the short-run players conveys only partial information about the stage-game action used by the long-run player.

Observing actions

Our bad-reputation result rules out games where the long-run player's action is perfectly observed whenever entry occurs. This is immediate when the long-run player has a pure action which makes participation a best reply. In that case, a good equilibrium exists in which the long run player plays this action in the first period and separates from any bad types. The more interesting case arises when only completely mixed actions induce entry. The following is a typical example of such a game (Fig. 2).

In this stage game the long-run player is an auditor who choose either to audit the East or West division of the short-run corporation, or not to audit ($\boldsymbol{N}$). The short-run player chooses whether to hire the auditor ($\boldsymbol{H}$) or take one of the exit actions ($\boldsymbol{L}$, $\boldsymbol{R}$). Notice also that hiring the auditor is a best response only if the auditor places strictly positive probability on both $\boldsymbol{E}$ and $\boldsymbol{W}$.

This game is strategically similar to EV and other bad reputation games: the auditor would like to be hired, but the corporation would not hire "bad" auditor who always audited one division, say $\boldsymbol{E}$. In order to avoid such a bad reputation, the long-run player has an incentive to increase the probability of $\boldsymbol{W}$, but this would also cause the short-run player to exit. Despite this similarity, we show that this is not a bad-reputation game, and more strongly that the game does have an equilibrium where the long-run player obtains a good payoff. This shows that a key property of EV and other bad reputation games is that the long-run player's action is not perfectly observed even when participation occurs.

	L	H	R
E	0,4	1,3	0,0
W	0,0	1,3	0,4
N	0,1	0,0	0,1

Fig. 2. Corporate fraud game.

We should also point out that adding an observed action to a bad reputation game can eliminate the bad-reputation effect. For example, in the EV game we might allow the mechanic give away money in addition to any repair performed. If the gift is large enough to induce participation regardless of the repair, the game no longer satisfies our conditions for a bad reputation game. In fact there will be a good equilibrium in which in early stages the rational type pays to establish a good reputation and is eventually able to separate from any bad type.

The teaching-evaluation game

The following game illustrates both the extension to several short-run players in each period, and the way that two actions by the long-run player can lead to the same observed outcome even when entry occurs. Our definition of bad reputation games will accommodate both of these features.

In this game the short-run players are students, the long-run player is a teacher, and the signals are teaching evaluations. Each period, each short-run player decides whether to enter—that is, take the class—or not, and the class is taught regardless of how many students enter. The long run player has a pair of binary choices: he can either teach well or teach poorly, and he can either administer teaching evaluations honestly or manipulate them. The public signals are the evaluations. If the evaluations are administered honestly and the class is taught well, there is a high probability of a good evaluation, while if evaluations are administered honestly and the class is taught poorly, the probability of good evaluations is low. Manipulating the evaluations is certain to lead to a good evaluation, irrespective of the quality of teaching. Students only want to take the class if it will be taught well, and the teacher would rather teach well and have students than face an empty class, and it is too costly to manipulate the ratings to be worthwhile. So in the one-shot game with only the rational type, the rational type teaches well and does not manipulate the ratings. However, when there is a small probability that the instructor is a bad type who prefers to teach poorly and manipulate the ratings, this is a bad reputation game, so when the teacher is patient his payoff corresponds to no one taking the class.

2. The model

2.1. The dynamic game

There are $J + 1$ player roles, filled by a long run-player L, and J dynasties of short-run players $i = 1, \ldots, J$. The game begins at time $t = 1$ and is infinitely repeated. Each period, players simultaneously choose actions from their action spaces. We denote the action space of the long-run player by A and the action space of short-run player i by B^i, and assume that all action spaces are finite. A profile of actions for the short-run players is denoted $b \in B$.

The long-run player discounts the future with discount factor δ. Each short-run player plays only in one period, and is replaced by an identical short-run player in the next period. There is a set Θ of types of long-run player. There are two sorts of types: type $0 \in \Theta$ is called the "rational type," and is the focus of our interest, with utility described below. For each pure action $a \in A$, type $\theta(a)$ is a "committed type" that is constrained to play a. These are the only possible types in Θ. The stage-game utility functions of the short-run players are u^i, and u^L denotes the utility function of the long-run player of type $\theta = 0$. The common prior distribution over types of the long-run player at time 0 is a probability measure denoted μ_0 We will not assume that every pure action commitment type necessarily has positive probability.

There is a finite public signal space Y with signal probabilities $\rho(y \mid a, b)$, given action profile (a, b). All players observe the history of the public signals. Short-run players observe only the history of the public signals, and in particular observe neither the past actions of the long-run player, nor of previous short-run players.[4] We let $h_t = (y_1, y_2, \ldots, y_t)$ denote the public history through the end of period t. We denote the null history by 0. We let h_t^L denote the private history known only to the long-run player. This includes his own actions, and may or may not include the actions of the short-run players he has faced in the past.

A strategy for the long-run player is a sequence of maps $\sigma^L(h_t, h_t^L, \theta) \in \mathcal{A}$, where $\mathcal{A} = \Delta(A)$ is the set of mixed strategies over A; a strategy profile for the short-run players is a sequence of maps $\sigma^j(h_t) \in \Delta(B^j) \equiv \mathcal{B}^j$; $\mathcal{B} = \times_j \mathcal{B}^j$ denotes the product of the $\mathcal{B}^j$'s (and not the convex hull of the product of the $\mathcal{B}^j$'s). Note that we write σ^{-L} for the profile of short-run player strategies. A short-run profile β is a *Nash response* to $\alpha \in \mathcal{A}$ if $u^i(\alpha, \beta^i, \beta^{-i}) \geqslant u^i(\alpha, \beta^i, \beta^{-i})$ for all short-run players i and $\beta^i \in \mathcal{B}^i$. We denote the set of short-run Nash responses to α by $R(\alpha)$.

Given strategy profiles σ, the prior distribution over types μ_0 and a public history h_t that has positive probability under σ, we can calculate from σ^L the conditional probability of long-run player actions, denoted $\bar{\alpha}(h_t)$, given the public history. A *Nash Equilibrium* is a strategy profile σ such that for each positive probability history

(1) $\sigma^{-L}(h_t) \in R(\bar{\alpha}(h_t))$ [short-run players optimize],
(2) $\sigma^L(h_t, h_t^L, \theta(a)) = a$ [committed types play accordingly],
(3) $\sigma^L(h_t, h_t^L, 0)$ is a best-response to σ^{-L} [rational type optimizes].

2.2. The Ely–Valimaki example

In EV, the long-run player is a mechanic, her action is a map from the privately observed state of the customer's car $\omega \in \{\boldsymbol{E}, \boldsymbol{T}\}$ to a choice of repair $\{\boldsymbol{e}, \boldsymbol{t}\}$. Here $\boldsymbol{E}$ means the car needs a new engine, $\boldsymbol{T}$ means it needs a tune-up, $\boldsymbol{e}$ is the decision to replace the engine, and $\boldsymbol{t}$ is the decision to give the car a tune-up. Thus the long-run player's action space is the set $A = \{\boldsymbol{ee}, \boldsymbol{et}, \boldsymbol{te}, \boldsymbol{tt}\}$, where the first component is the repair chosen in the state $\boldsymbol{E}$. The one short-run player chooses an element of $B^1 = \{\boldsymbol{In}, \boldsymbol{Out}\}$. The public signal takes on the values $Y = \{\boldsymbol{e}, \boldsymbol{t}, \boldsymbol{Out}\}$. If the short-run player chooses $\boldsymbol{Out}$ the signal is $\boldsymbol{Out}$ regardless of the action of the long-run player, that is $\rho(\boldsymbol{Out} \mid \cdot, \boldsymbol{Out}) = 1$. Otherwise the signal is the repair chosen by the long-run player. The two states of the car are assumed to be i.i.d. and equally likely, so $\rho(\boldsymbol{e} \mid (\boldsymbol{et}, \boldsymbol{In})) = \rho(\boldsymbol{e} \mid (\boldsymbol{te}, \boldsymbol{In})) = 1/2$, $\rho(\boldsymbol{e} \mid (\boldsymbol{ee}, \boldsymbol{In})) = 1$, and $\rho(\boldsymbol{e} \mid (\boldsymbol{tt}, \boldsymbol{In})) = 0$.

If the short-run player chooses $\boldsymbol{Out}$, each player gets utility 0. If he plays $\boldsymbol{In}$ and the long-run player's repair is the correct one (that is, matches the state), the short-run player receives u; otherwise, he receives $-w$, where $w > u > 0$. The "rational type" of long-run player has exactly the same stage-game payoff function as the short run players. When the rational type is the only type in the model, there is an equilibrium where he chooses the correct repair, all short-run players enter, and the rational type's payoff is u. When there is also a probability that the long-

[4] We do not assume that the payoffs depend on the actions only through the signals, so the short-run players at date t are not necessarily able to infer the realized payoffs of the previous generations of short-run players. Fudenberg and Levine (1992) assumed that a player's payoff was determined by his own action and the realized signal, but that assumption was not used in the analysis. The assumption is used in models with more than one long-run player to justify the restriction to public equilibria, but it is not needed here.

run player is a "bad type" who always plays ***ee***, the long-run player's payoff is bounded by an amount that converges to 0 as the discount factor goes to 1.

2.3. Participation games and bad reputation games

As we indicated, we will study *participation games* in which the short-run players may choose not to participate. *Bad reputation games* are a subclass of participation games that have the additional features needed for the bad reputation result; the following is a brief summary of the key features of these games. First, there is a set of "friendly" actions that must receive sufficiently high probability to induce the short-run players to participate, such as ***et*** in the EV example. Next, there are "bad signals." These are signals that are most likely to occur when "unfriendly" actions are played but also occur with positive probability when friendly actions are played. In EV the bad signal *is* ***e***. Finally, there are some actions that are not friendly, but reduce the probability of the bad signals, such as ***tt*** in EV; we call these actions "temptations." If there is a positive prior probability that the long-run player is a "bad type" that is committed to one of the unfriendly actions, then after histories with many bad signals the short-run players will become sufficiently convinced they are facing such a bad type and exit. In order to avoid these histories the rational type of long-run player may choose to play one of the temptations, and foreseeing this, the short-run players will chose not to enter. Our main result shows that this leads to a "bad reputation" result whenever the prior does not assign too much probability to types that are committed to play friendly actions.

To model the option to not participate, we assume that certain public signals $y \in Y^E$ are *exit signals*. Associated with these exit signals are *exit profiles*, which are pure action profiles $e \in E \subseteq B$ for the short run players. We refer to $B - E$ as the entry profiles.[5]

For each such e, $\rho(y \mid a, e) = \rho(y \mid e)$ for all a, and $\rho(Y^E \mid e) = 1$. In other words, if an exit profile is chosen, the distribution of signals is independent of the long-run player's action, and only exit signals can be observed. Moreover, if $b \notin E$ then $\rho(Y^E \mid a, b) = 0$ for all $a \in A$, so that an entry profile cannot give rise to an exit signal. Formally, a *participation game* is a game in which $E \neq \emptyset$ and moreover there is some $\alpha \in A$ with $R(\alpha) \cap E \neq \emptyset$. The remainder of the paper specializes to participation games.

We begin by distinguishing actions by the long-run player that encourage the short-run players to enter (friendly actions) and those that cause them to exit (unfriendly actions). In the EV example, the "honest" strategy ***et*** induces entry, and each of the other pure strategies induces exit. The appropriate definitions of friendly and unfriendly are more complex in our more general setting for several reasons. First of all, in EV the friendly action is pure, but there are games with a single short-run player in which only mixed actions by the long-run player induce entry. Second, in games with several short-run players, the set of exit profiles need not have a product structure, as for example when any player can unilaterally "veto" the participation of all of them. Similarly, the set of Nash responses typically does not have a product structure, as when (***In***, ***In***) and (***Out***, ***Out***) are both in $R(a)$ but (***In***, ***Out***) is not.

Let $\beta\{E\}$ be the probability assigned to the set $E \subseteq B$ by β.

[5] We allow for the possibility that there are several exit signals; this could correspond for example to the case where any short-run player can veto participation by all of them, and the identity of the vetoing player is observed.

Definition 1. A non-empty finite set of pure actions F for the long-run player is friendly if there is a number $\gamma > 0$ such that, for all α if $\beta \in R(\alpha)$ and $\beta\{E\} < 1$ then $\alpha(f) \geqslant \gamma$ for some $f \in F$. An *unfriendly* set N corresponding to F is any non-empty subset of $A\backslash F$.

This definition says that if the Nash response of the short-run players assigns positive probability to a non-exit profile, the probability given to some friendly action must be bounded below by $\gamma > 0$. Note that this is a necessary condition to induce entry, but it need not be sufficient. Conversely, if F is a friendly set, then any pure action not in F must cause the short-run players to exit, hence the definition of an unfriendly set. Note that the definition requires that N and F be disjoint; we discuss the reason for this in Section 4.3.[6] In the EV example the only friendly action is $\boldsymbol{et}$, so the maximal unfriendly set is $\{\boldsymbol{ee}, \boldsymbol{tt}, \boldsymbol{te}\}$.

For any action to be played in equilibrium, it must at least be possible to design continuation payoffs that deter deviation to some action which improves the current payoff. The following is therefore a necessary condition for a friendly action to induce entry in equilibrium. It is related to the notion of an action being identified, as in Fudenberg et al. (1994).

Definition 2. A mixed action α for the long run player is *enforceable* (*using actions that permit entry*) if there does not exist another action $\tilde{\alpha}$ such that for all β such that $\beta \in R(\alpha)$ and $\beta\{E\} < 1$, $u^L(\tilde{\alpha}, \beta) > u^L(\alpha, \beta)$ and $\rho(\cdot \mid \tilde{\alpha}, \beta) = \rho(\cdot \mid \alpha, \beta)$. When α is not enforceable, we say that the action $\tilde{\alpha}$ *undermines* α.

Only enforceable friendly actions can induce entry in equilibrium. When entry occurs, rather than play an unenforceable action, the long run player would switch to the undermining action, thereby strictly increasing his current payoff while maintaining the same distribution over signals, and hence future payoff. In the teaching-evaluations game described earlier, the action "teach well, manipulate" is unenforceable: "teach poorly and manipulate" yields a higher stage game payoff and the same distribution over signals. Hence the only enforceable friendly action in that game is "teach well, administer honestly."

Next we consider what signals may reveal about actions.

Definition 3. A set of signals $\widehat{Y}$ is *evidence* for a set of actions N if N is non-empty and $\rho(\hat{y} \mid n, b) > \rho(\hat{y} \mid a, b)$ for all $b \notin E$, $\hat{y} \in \widehat{Y}$, $n \in N$, $a \notin N$.

This is a strong condition: Every action in N must imply a higher probability for each signal in $\widehat{Y}$ than any action not in N. A given set of actions may not have signals that are evidence; in the case of the EV example, $\boldsymbol{e}$ is evidence for the unfriendly set $\{\boldsymbol{ee}\}$.

Definition 4. An action a is *vulnerable to temptation relative to a set of signals* $\widehat{Y}$ if there exist numbers $\underline{\rho}$, $\tilde{\rho} > 0$ and an action d such that

(1) If $b \notin E$, $\hat{y} \in \widehat{Y}$, then $\rho(\hat{y} \mid d, b) \leqslant \rho(\hat{y} \mid a, b) - \underline{\rho}$.
(2) If $b \notin E$ and $y \notin \widehat{Y} \cup Y^E$ then $\rho(y \mid d, b) \geqslant (1 + \tilde{\rho})\rho(y \mid a, b)$.
(3) For all $b \in E$, $u^L(d, b) \geqslant u^L(a, b)$.

[6] The maximal unfriendly set corresponding to a fixed F is $A\backslash F$. The reason that we do not simply define N to equal $A\backslash F$ is that this would make the other conditions for the bad-reputation result harder to satisfy.

The action d is called a *temptation*. The largest parameters $\underline{\rho}$, $\tilde{\rho}$ for which there is a d that satisfy (1) and (2) are the *temptation bounds* for action a.

In other words, an action is vulnerable if it is possible to lower the probability of all of the signals in $\widehat{Y}$ by at least $\underline{\rho}$ while increasing the probability of each other signal by at least the multiple $(1+\tilde{\rho})$. Notice that for an action to be vulnerable to a temptation, it must place at least weight $\underline{\rho}$ on each signal in $\widehat{Y}$. Notice also that the definition does not restrict the payoff to the vulnerable action conditional on participation—the temptation here is not to increase the current payoff, but rather to decrease the probability of the signals in $\widehat{Y}$.

In the EV example, the friendly action $\boldsymbol{et}$ is enforceable but vulnerable relative to $\{\boldsymbol{E}\}$. The temptation is $\boldsymbol{tt}$, which sends the probability of the signal $\boldsymbol{E}$ to zero. (Since there is only one other signal, condition (2) of the definition is immediate.) Section 5 considers games which satisfy the stronger assumption that the temptation leaves the relative probabilities of all signals in $\widehat{Y} \cup Y^E$ unchanged. The latter assumption would be satisfied in any game with only two non-exit signals.

Definition 5. A participation game has *exit minmax* if

$$\max_{b\in E\cap \text{image}(R)} \max_a u^L(a,b) = \min_{\beta\in \text{image}(R)} \max_a u^L(a,\beta).$$

In other words, any exit strategy forces the long-run player to the minmax payoff, where the relevant notion of minmax incorporates the restriction that the action profile chosen by the short-run players must lie in the image of R.[7] It is convenient in this case to normalize the minmax payoff to 0. In participation games without exit minmax, there are outcomes that are even worse for the long-run player than obtaining a bad reputation. In this case, not only is exit "not so bad" for the long-run player, but as we show in Section 4.4, there can be equilibria where the long-run player does better than the exit payoff. Loosely speaking, in such games the long-run player can be deterred from his temptation to avoid exit by the threat of a stronger punishment.

We are now in a position to define *bad reputation games*.

Definition 6. 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 $\widehat{Y}$ that are evidence for N, such that every enforceable $f \in F$ is vulnerable to temptation relative to $\widehat{Y}$. The signals $\widehat{Y}$ are called the *bad signals*.

In particular, the EV game is a bad reputation game. We take the friendly set to be $\{\boldsymbol{et}\}$, the unfriendly set to be $\{\boldsymbol{ee}\}$ and the unfriendly signals to be $\{\boldsymbol{E}\}$. We have already observed that $\{\boldsymbol{et}\}$ is a friendly set and $\{\boldsymbol{ee}\}$ unfriendly. Moreover, $\{\boldsymbol{E}\}$ is evidence for $\{\boldsymbol{ee}\}$.

[7] By the image of correspondence R we mean the union over all α of the sets $R(\alpha)$. In a participation game, the set $E \cap \text{image}(R)$ is non-empty. When there is a single short-run player, the restriction $\beta \in \text{image}(R)$ collapses to the constraint of not playing strictly dominated strategies, but when there are multiple short-run players it can involve additional restrictions. It is clear that no sequential equilibrium could give the rational type a lower payoff than the minmax level defined in Definition 5. Conversely, in complete-information games, any long-run player payoff above this level can be supported by a perfect public equilibrium if actions are identified and the public observations have a "product structure" (Fudenberg and Levine, 1994). This is true in particular when actions are publicly observed as shown in Fudenberg et al. (1990).

In a bad reputation game, the relevant temptations are those relative to $\widehat{Y}$. For the remainder of the paper when we examine a bad reputation game and refer to a temptation, we will always mean relative to the set $\widehat{Y}$.

It is useful to define several constants describing bad reputation games. Recall that γ is the bound in the definition of a friendly set. Since the friendly set is finite, we may define $\varphi > 0$ to be the minimum, taken over elements of the friendly set, of the temptation bounds $\underline{\rho}$, and let ζ be the minimum over the friendly set of the temptation bounds $\tilde{\rho}$. Define

$$r = \min_{n\in N,\, a\notin N,\, \beta\{E\}<1,\, \widehat{y}\in\widehat{Y}} \frac{\rho(\widehat{y}\mid n,\beta)}{\rho(\widehat{y}\mid a,\beta)},$$

where we take $1/0 = +\infty$. Since N and F are non-empty and disjoint, r is finite, and since $\widehat{Y}$ is evidence for N, $r > 1$.

In other words, r measures how revealing the evidence is. Note that γ is the minimum probability a friendly action must be played, while φ measures the amount by which a tempting action lowers the probability of bad signals. This leads us to define the *signal lag*

$$\eta = -\log(\gamma\varphi)/\log r,$$

which is positive. To interpret it, suppose that a friendly action is supposed to be played with probability γ; the signal lag is a measure of how long it would take to learn that a temptation is actually being played instead. It is also convenient to define

$$k_0 = -\frac{\log(1-\gamma)}{\log(1-\gamma+\frac{\gamma}{r})}.$$

3. The theorem

We now prove our main result: In a bad reputation game with a sufficiently patient long-run player and likely enough unfriendly types, in any Nash equilibrium, the long-run player gets approximately the exit payoff. The proof uses several lemmas proven in Appendix A.

We begin by describing what it means for unfriendly types to be likely "enough." Let $\Theta(F)$ be the commitment types corresponding to actions in F. We will call these the *friendly commitment types*. Let $\Theta(N)$ be the *unfriendly commitment types* corresponding to the unfriendly set N. Note that the sets $\Theta(F)$ and $\Theta(N)$ are disjoint.

Definition 7. A bad reputation game with friendly set F and unfriendly set N has *commitment size* ε if

$$\mu_0\big[\Theta(F)\big] \leqslant \varepsilon^{1+\eta}\left(\frac{\mu_0[\Theta(N)]}{\mu_0[\Theta(F)]}\right)^{\eta}.$$

This notion of commitment size places a bound on the prior probability of friendly commitment types that depends on the prior probability of the unfriendly types. Since η is positive, the larger the prior probability of $\Theta(N)$, the larger the probability of the friendly commitment types is allowed to be. The hypothesis that the priors have commitment size ε for sufficiently small ε is a key assumption driving our main results. Notice that this bound can be rewritten as $\mu_0[\Theta(F)] \leqslant \varepsilon(\mu_0[\Theta(N)])^{\eta/(1+\eta)}$, so as η grows the probability of friendly commitment types must become lower.

Note that the assumption of a given commitment size does not place any restrictions on the relative probabilities of commitment types. In particular, let $\tilde{\mu}$ be a fixed prior distribution over the commitment types, and consider priors of the form $\lambda\tilde{\mu}$, where the remaining probability is assigned to the rational type. Then the right-hand side of the inequality defining commitment size depends only on $\tilde{\mu}$, and not on λ, while the left-hand side has the form $\lambda\tilde{\mu}$. Hence for sufficiently small λ the assumption of commitment size ε, ϕ is satisfied for any ϕ. Note that the EV example has commitment size 0 since the only types are the rational type and the commitment type who plays *ee*.

Define $U^L = (\max_{a,b} u^L(a,b)) - \min\{0, \min_{a,b} u^L(a,b)\}$. Let v^L be the maximum of the payoff of the rational type in any Nash equilibrium.

Theorem 1. *In a bad reputation game if the commitment size is $\gamma/2$, $\lim_{\delta\to 1} v^L = 0$.*

Moreover, for any δ

$$v^L \leqslant (1-\delta)k^*\left(\frac{2}{\zeta}\right)^{k^*}\left(1+\frac{1}{\zeta}\right)U^L,$$

where

$$k^* = k_0 + \log\bigl(\mu_0\bigl[\Theta(N)\bigr]\bigr)\Big/\log\Bigl(1-\gamma+\frac{\gamma}{r}\Bigr)$$

is an upper bound on the number of consecutive bad signals that can be observed before all subsequent short-run players choose exit.

For the rest of this section, we fix an arbitrary Nash equilibrium. Given this equilibrium, for each positive probability public history h_t, let $v(h_t)$ denote the expected continuation value to the rational long-run player, let $\mu(h_t)$, $\beta(h_t)$ be the posterior beliefs and actions of the short-run players, and $\alpha_0(h_t)$ be the action of the rational type of long-run player. If a has positive probability under $\bar{\alpha}(h_t)$, and b positive probability under $\beta(h_t)$, then we define

$$v(h_t, a, b) \equiv (1-\delta)u^L(a,b) + \delta\sum_y \rho(y \mid a, b)v(h_t, y).$$

When mixed actions α and β put weight only on such a, b, it is convenient also to define $v(h_t, \alpha, \beta)$ in the natural way.

The proof uses a series of lemmas whose proofs are in the appendix. The first lemma says that when the prior on friendly types is sufficiently low, entry (and hence the realization of a bad signal $\hat{y} \in \widehat{Y}$) can occur only if the rational type is playing a friendly strategy with appreciable probability.

Lemma 1. *In a participation game, if h_t is a positive probability history in which $\hat{y} \in \widehat{Y}$ occurs in period t and $\mu(h_{t-1})[\Theta(F)] \leqslant \gamma/2$ then $\alpha_0(h_t)(f) \geqslant \gamma/2$ for some friendly f.*

This is a consequence of the definition of friendly strategies: entry requires that the overall strategy assigns some minimum probability to a friendly action, and if the friendly types are unlikely, then a non-negligible part of this probability must come from the play of the rational type.

Lemma 2. *In a bad reputation game, if h_t is a positive probability history, and the signals in h_t all lie in $Y^E \cup \widehat{Y}$, then*

(a) *At most*

$$k^* = k_0 + \log\bigl(\mu_0\bigl[\Theta(N)\bigr]\bigr)/\log\Bigl(1-\gamma+\frac{\gamma}{r}\Bigr)$$

of the signals are in $\widehat{Y}$.

(b) *If the commitment size is* $\gamma/2$ *then* $\mu(h_t)[\Theta(F)] \leqslant \gamma/2$.

Remark. The intuition for part (a) is simple, and closely related to the argument about the deterministic evolution of beliefs in FL: The short-run players exit if they think it is likely that entry will lead to the observation of a bad signal. Hence each observation of a bad signal is a "surprise" that increases the posterior probability of the bad type by (at least) a fixed ratio greater than 1, so along a history that consists of only bad signals and exit signals, the posterior probability of the bad type eventually gets high enough that all subsequent short-run players exit. This argument holds no matter what other types have positive probability, and it is the only part of this lemma that would be needed when there are only two types, one rational and one bad, as in EV.

However, as we will show by example below (see Section 4.1), we cannot expect the "bad reputation" result to hold when there is sufficiently high probability of the Stackelberg type. Therefore, part (b) of the lemma provides a condition on the prior (expressed in terms of commitment size) which ensures that the probability of the Stackelberg type remains low along any history which consists only of exit outcomes and bad signals. This follows because the friendly types are disjoint from the unfriendly types, and the bad signals are evidence for the unfriendly actions, so every observation a bad signal increases the relative probability of unfriendly types compared to any other commitment type.

Define

$$\bar{u}(y,\tilde{\rho}) = \begin{cases} (1+\frac{1}{\tilde{\rho}})U^L & y \in \widehat{Y}, \\ 0 & \text{otherwise}, \end{cases}$$

$$\bar{\delta}(y,\tilde{\rho}) = \begin{cases} \frac{\delta}{\tilde{\rho}}+1 & y \in \widehat{Y}, \\ \delta & \text{otherwise} \end{cases}$$

and $Y(h_t) = \{y \in Y^E \cup \widehat{Y} \mid \rho(y \mid \bar{\alpha}(h_t), \beta(h_t)) > 0\}$.

Lemma 3. *In a participation game if* $\beta(h_t)\{E\} = 1$, *or* $\beta(h_t)\{E\} < 1$ *and* $\alpha_0(h_t)(f) > 0$ *for some vulnerable friendly action* f *with temptation bounds* $\underline{\rho}$, $\tilde{\rho}$, *then*

$$v(h_t) \leqslant \max_{y \in Y(h_t)} (1-\delta)\bar{u}(y,\tilde{\rho}) + \bar{\delta}(y,\tilde{\rho})v(h_t,y).$$

Remark. This lemma says that if the rational type is playing a friendly strategy, his payoff is bounded by a one-period gain and the continuation payoff conditional on a bad signal. This follows from the assumption that for every entry-inducing strategy it is possible to lower the probability of all of the signals in $\widehat{Y}$ by at least φ while increasing the probability of each other signal by at least the multiple $\tilde{\rho}$. Because there is an exit minmax, the fact that the rational type chooses not to reduce the probability of the bad signal means that the continuation payoff after the bad signal cannot be much worse than the overall continuation payoff.

Proof of Theorem 1. The idea is to construct a particular positive probability sequence of histories and show that at most k^* of the signals are in $\widehat{Y}$, the unfriendly set. In addition because of

the commitment size assumption, the probability of friendly types is not too large. This enables us to give a bound on $v(0)$ leading to the desired conclusion.

Given an equilibrium, we begin by constructing a positive probability sequence of histories beginning with an initial history at date 0. The construction is recursive. Once we have constructed h_t, we define $h_{t+1} = (h_t, y_{t+1})$. Recall that ζ is the minimum over the friendly set of the temptation bounds $\tilde{\rho}$. We choose $y_{t+1} \in \arg\max_{y \in Y(h_t)} (1-\delta)\bar{u}(y,\zeta) + \bar{\delta}(y,\zeta)v(h_t,y)$, where we know that $Y(h_t)$ is not empty because either $\beta(h_t)\{E\} = 1$ or $\beta(h_t)\{E\} < 1$. This latter case implies that $\bar{\alpha}(h_t)(f) \geqslant \gamma$ for some friendly f, and since only enforceable actions can induce entry in equilibrium, this f must be vulnerable to temptation, so $\rho(\widehat{Y} \mid \bar{\alpha}(h_t), \beta(h_t)) \geqslant \gamma\rho(\widehat{Y} \mid f, \beta(h_t)) > 0$.

Since the history is a sequence of signals all of which lie in $Y^E \cup \widehat{Y}$, we can apply Lemma 2 to conclude that for each h_t at most k^* of the signals are in $\widehat{Y}$ and $\mu(h_t)[\Theta(F)] \leqslant \gamma/2$. Consider an h_t such that $\beta(h_t)\{E\} < 1$. As argued above, we know that $\bar{\alpha}(h_t)(f) \geqslant \gamma$ for some vulnerable friendly f, so $\mu(h_t)[\Theta(F)] \leqslant \gamma/2$ and Lemma 1 implies that $\alpha_0(h_t)(f) \geqslant \gamma/2 > 0$. Now apply Lemma 3 to conclude that for each h_t

$$v(h_t) \leqslant (1-\delta)\bar{u}(y_{t+1},\zeta) + \bar{\delta}(y_{t+1},\zeta)v(h_{t+1}).$$

Since $v(h_t) \leqslant U^L$, it follows that for any history h_t the associated y_t are such that

$$v(0) \leqslant (1-\delta)\sum_{t=1}^{\infty}\prod_{\tau=2}^{t}\bar{\delta}(y_\tau,\zeta)\bar{u}(y_t,\zeta),$$

when $t = 1$ the product term is defined to equal 1. Since $Y^E \cap \widehat{Y} = \emptyset$, $\bar{u}(y,\zeta) = 0$ for $y \in Y^E$. Since $y_t \in \widehat{Y}$ at most k^* times, the right hand is largest in when all k^* of these times occurs at the start of the history. Substituting in the definitions of $\bar{\delta}$ and $\bar{u}$ we see that

$$v(0) \leqslant (1-\delta)\sum_{t=1}^{k^*}\left(\prod_{\tau=2}^{t}\left(\frac{\delta}{\zeta}+1\right)\right)\left(1+\frac{1}{\zeta}\right)\bar{U}.$$

The fact that $\delta/\tilde{\rho} + 1 \leqslant 2/\tilde{\rho}$ and that $y_t \in \widehat{Y}$ at most k^* times now gives the desired bound. □

4. Examples

To illustrate the scope of Theorem 1, and also the extent to which the assumptions are necessary as well as sufficient, we now formally examine the examples we presented in the introduction. To begin, Example 4.1 illustrates what happens when the commitment size hypothesis is not satisfied. Example 4.2 shows that the EV conclusion is not robust to the addition of an observed action that makes the short-run players want to enter. Example 4.3 examines participation games that are not bad reputation games, and Example 4.4 illustrates the role of the exit-minmax assumption. In all of the examples but Example 4.1, we assume that the commitment size hypothesis is satisfied, and investigate whether the game is a bad reputation game.

4.1. EV with Stackelberg type

One way that we relax the original assumptions of EV is to allow for positive probabilities of all commitment types. In particular, we allow a positive probability of a "Stackelberg type" committed to the honest strategy *et*, which is the optimal commitment. However, a hypothesis of the theorem is that the prior satisfies the commitment size assumption.

It is immediate that the short run players refuse to enter regardless of the behavior of the rational type when the probability of the bad type is sufficiently high. This is the region labeled A in Fig. 1. Bad reputation arises because the long-run player tries to prevent the posterior from moving into this region.

When there is a sufficiently high probability of the Stackelberg type, the short-run players will enter regardless of the behavior of the rational type; this is region B in Fig. 1. Moreover, in this region, there is a Nash (and indeed sequential) equilibrium in which the long run player receives the best commitment payoff, which is "u" in the notation of EV. The equilibrium is constructed as follows: Consider the game in which the posterior probability of the bad type is zero. In this game there exists a sequential equilibrium in which the long-run player gets u. Suppose that we assume that this is the continuation payoff in the original game in any subform in which the long-run player played t at least once in the past. A sequential equilibrium of this modified game is clearly a sequential equilibrium of the original game, and by standard arguments, this modified game has a sequential equilibrium. How much does the rational long-run player get in this sequential equilibrium? One option is to play tt in the first period. Since the short-run player is entering regardless, this means that beginning in period 2 the rational type gets u. In the first period he gets $(u - w)/2$. Hence in equilibrium he gets at least $(1 - \delta)(u - w)/2 + \delta u$, which converges to u as $\delta \to 1$.

Our theorem is about the set of equilibrium payoffs for priors outside of these two regions. The theorem states that there is a curve, whose shape is represented in Fig. 1, such that when the prior falls below this curve (region C), the set of equilibrium payoffs for the long-run player is bounded above by a value that approaches the minmax value as the discount factor converges to 1.

4.2. Adding an observed action to EV

We now modify the EV game by giving the long-run player the option of "giving away money" in addition to performing a repair. Giving money simply transfers a fixed amount of utility from the mechanic to the customer, independent of the outcome of the repair. We denote giving money by $\boldsymbol{g}$, and not giving by $\boldsymbol{n}$, so that the long-run player's action set is now $A = \{\boldsymbol{nee}, \boldsymbol{net}, \boldsymbol{nte}, \boldsymbol{ntt}, \boldsymbol{gee}, \boldsymbol{get}, \boldsymbol{gte}, \boldsymbol{gtt}\}$, and the set of observed outcomes is $Y = \{\boldsymbol{ge}, \boldsymbol{gt}, \boldsymbol{ne}, \boldsymbol{nt}, \boldsymbol{Out}\}$.

Suppose first that the gift is large enough that $\boldsymbol{gtt}$ induces participation; this implies that gtt is in every friendly set. Moreover, since the gift gives rise to the signal ge for sure whenever the short-run player participates, it is not vulnerable to temptation with respect to any signals that are evidence for any other action, so this is not a bad reputation game. Moreover, even without a Stackelberg type the EV conclusion fails in this game: there is an equilibrium where the rational type plays gee in the first period. This reveals that he is the rational type, and there is entry in all subsequent periods, while playing anything else reveals him to be the bad type so that all subsequent short-run players exit.

On the other hand, if the gift is small enough that the only friendly actions are net and get, then the possibility of the gift does not matter, and this remains a bad reputation game. As this shows, the assumption that every friendly action is vulnerable to temptation is important for the results and economically restrictive.

Our definition of a bad reputation game requires that the friendly and corresponding unfriendly sets are disjoint. There is a tension between this requirement and the requirement that friendly actions be vulnerable to temptation. It is because of this tension that games in which

	L	H	R
E	0,4	1,3	0,0
W	0,0	1,3	0,4
N	0,1	0,0	0,1

Fig. 3.

the long-run player's action is perfectly observed (conditional on entry) are never bad reputation games. To illustrate this point, recall the auditing game (see Fig. 3).

This game is a participation game where $\boldsymbol{L}$ and $\boldsymbol{R}$ by player 2 are exit actions and $\boldsymbol{H}$ is entry. Note that both $\boldsymbol{E}$ and $\boldsymbol{W}$ for player 1 cause exit, while mixing between the two actions with probability of $\boldsymbol{E}$ between $1/4$ and $3/4$ can induce entry Thus $(\boldsymbol{E}, \boldsymbol{W})$ is a friendly set, and $\boldsymbol{N}$ is unfriendly.

Suppose first that when the short-run player enters, the action of the long-run player is perfectly observed. In this case, the bad signals are simply the unfriendly actions themselves. However, the game is not a bad reputation game for any choice of friendly and unfriendly sets.

Proposition 1. *Participation games in which, conditional on entry, the action of the long-run player is perfectly observed, are never bad reputation games.*

Proof. If the game is a bad reputation game, it must have a friendly set $F$and non-empty unfriendly set N, and a set of signals that is evidence for N, such that the enforceable friendly actions are vulnerable to temptation relative to $\widehat{Y}$. Since the actions of the long-run player are observed, the only signals that are evidence for N are the signals corresponding to actions in the set N. With observed actions, no action in F gives rise to a bad signal with positive probability. But then no element of F is vulnerable to temptation. □

Proposition 1 shows that games with observed actions do not satisfy the hypotheses of Theorem 1. In fact, the conclusion typically fails as well, and even when the only commitment types are unfriendly types, there are equilibria (for δ close enough to 1) in which the long-run player obtains his best payoff in the game without commitment types.

To see this, note that there are two possibilities when the action of the long run player is observed. The first possibility is that there is at least one pure action a of the long-run player that induces entry. In this case, consider the following strategy profile. The rational type of long-run player plays a in the first period and thereafter plays according to his best equilibrium of the complete information game. Since a is friendly, the short-run players will enter and the long-run player separates from the unfriendly types after one period. Thereafter, the probability of commitment types is zero and hence the specified continuation profile is a continuation equilibrium. Finally, as δ approaches 1, the payoff of the long-run player approaches the best equilibrium payoff of the game without commitment types since he gets that amount beginning in period 2.

The second possibility is that the only entry-inducing actions are non-degenerate probability distributions. In this case, in order to induce entry, the long-run player must play a mixed action that assigns positive probability to pure actions that are unfriendly. We do not have a general result to offer here, but the game in Fig. 3 is the prototypical example of this case, and the

following argument shows that the conclusion of the theorem fails for this game. Suppose that the only commitment type with positive probability is $\boldsymbol{N}$, and that the probability of the bad type is less than $1/4$. Consider the following strategies: For any current probability $\mu(h_t)[\boldsymbol{N}]$ less than $1/4$ the rational type mixes so that the overall probability of $\boldsymbol{N}$ is exactly $3/4$. (In particular, this is true when the long-run player has been revealed to be rational, so that $\mu(h_t)[\boldsymbol{N}] = 0$.) The short-run player always enters. If $\boldsymbol{E}$ is observed, the type is revealed to be rational. If $\boldsymbol{N}$ is observed, the probability of the bad type increases by a factor of $4/3$, so when it first exceeds $1/4$ it is at most equal to $1/3$. At this point, the rational type may reveal himself by playing $\boldsymbol{E}$ with probability 1, while preserving the incentive of the short-run player to enter. It is easy to see that this is a Nash equilibrium for any discount factor of the long-run player, yet in this equilibrium, the long-run player's payoff is 1.

Next we consider games of almost-perfect monitoring, where all signals have positive probability under any action, but where if entry occurs the probability of the signal corresponding to the long-run player's action is at least $1 - \varepsilon$. In contrast to the argument above, these can be bad reputation games, but there is a sense in which "good reputation" is nonetheless the right prediction here. To see this, let F be a friendly set, and N an associated unfriendly set; for small ε the only signals that are evidence for N are those corresponding to actions in N. Since the friendly actions do generate bad signals, it is possible that there are $\underline{\rho}, \tilde{\rho} > 0$ such that one of the friendly actions is vulnerable to a temptation, and in this case we have a bad reputation game. However, as ε goes to 0, $\tilde{\rho}$ goes to 0 as well, so for any fixed discount factor δ the payoff bound in Theorem 1 become vacuous.[8]

4.3. *Exit minmax*

In participation games, reputation plays a role because the short run players will guard against unfriendly types by exiting. This is "bad" for the long-run player only if exit is worse than the payoff he otherwise would receive, and the exit minmax assumption ensures that this is the case.

In participation games without exit minmax, there are outcomes that are even worse for the long-run player than obtaining a bad reputation. In this case it is possible that there exist equilibria in which the long-run player is deterred from his temptation to avoid exit by the even stronger threat of a minmaxing punishment. For example consider the game in Fig. 4, where the first matrix represents the payoffs, and the second represents the distribution of signals conditional on entry.

	In	*Out*$_1$	*Out*$_2$
F	1,1	0,0	-2,0
U	1,0	0,1	-2,0
T	1,0	0,0	-2,1

	g	*b*	*r*
F	½	½	0
U	0	1	0
T	½	0	½

Fig. 4.

[8] We thank the referee for asking us about the impact of imperfect observability.

This game is a participation game with exit actions $\boldsymbol{Out}_1$ and $\boldsymbol{Out}_2$, unfriendly action $\boldsymbol{U}$ and friendly action $\boldsymbol{F}$ vulnerable to temptation $\boldsymbol{T}$. There are only two types, the rational type and a bad type that plays $\boldsymbol{U}$. Exit minmax fails because the maximum exit payoff exceeds the minmax payoff, and we claim that there are good equilibria in this game because the threat of exiting with $\boldsymbol{Out}_2$ is worse than the fear of obtaining a reputation for playing $\boldsymbol{U}$ which would only lead to exit with $\boldsymbol{Out}_1$.

To see this, consider the following strategy profile. The rational type plays $\boldsymbol{F}$ at every history unless the signal r has appeared at least once; in that case the rational type plays $\boldsymbol{T}$. The short run player plays $\boldsymbol{Out}_2$ if a signal of r has ever appeared. Otherwise, the short run player plays $\boldsymbol{Out}_1$ if the posterior probability of the bad type exceeds $1/2$ and In if this probability is less than $1/2$. Observations of r are interpreted as signals that the long-run player is rational.

Since $(\boldsymbol{T}, \boldsymbol{Out}_2)$ is a Nash equilibrium of the stage game, the continuation play after a signal of r is a sequential equilibrium. When r has not appeared, the long run player optimally plays $\boldsymbol{F}$. Playing $\boldsymbol{U}$ gives no short-run gain and hastens the onset of $\boldsymbol{Out}_1$, and playing $\boldsymbol{T}$ shifts probability from the bad signal b to the signal r, which is even worse.[9] The short-run players are playing short-run best responses. In this equilibrium, the long run player does not give in to the temptation to play $\boldsymbol{T}$. As a result, with positive probability, the short-run players never become sufficiently pessimistic to begin exiting, and so the long run player achieves his best payoff.

In the above example there were two exit actions. The next proposition states that when there is only one exit action and the long-run player's exit payoff is independent of his own action, the worst Nash equilibrium payoff in the dynamic game for the long run player is (not much worse than) his exit payoff. Note that this condition is satisfied in the principal-agent applications discussed in Section 5. The proposition is a consequence of FL.

Proposition 2. *Consider a participation game with a single short-run player and a unique exit action. If*

(i) *there exists a pure action*[10] $\hat{a}$*, such that* $R(\hat{a}) = \{exit\}$*,*
(ii) *the prior distribution assigns positive probability to a type that is committed to* $\hat{a}$*, and*
(iii) *the long-run player's action is identified conditional on entry*

then there is a lower bound on the payoffs to the rational type which converges to the exit payoff, as the discount factor approaches 1.

Proof. FL established[11] that for any game there exists a lower bound $l(\delta)$ on the set of Nash equilibrium payoffs for the rational type, and that as $\delta \to 1$, $l(\delta)$ converges to a limit that is at least

$$\max_{a \in C} \min_{\beta \in \tilde{B}(a)} u^L(\alpha, \beta)$$

[9] Playing T gives probability $1/2$ of shifting to the absorbing state where payoffs are -2. Playing the equilibrium action of F has probability at most $1/2$ of switching to the state where payoffs are 0.

[10] The assumption that this is a pure action is not necessary here; we state the result this way for consistency with the rest of the paper.

[11] The statement of the FL theorem requires that commitment types including mixing types have full support, in which case the set C is the space of all (mixed) actions, but the proof given there also shows that the version of the lower bound given here is correct.

where $\tilde{B}(a)$ is the set of self-confirmed best-responses to for the short-run player to a, and C is the set of actions corresponding to the support of the prior distribution over commitment types. Because the long-run player's action is identified conditional on entry and $R(\hat{a}) = \{exit\}$, we have $\tilde{B}(\hat{a}) = \{exit\}$, and because the type that plays $\hat{a}$ has positive prior probability, the FL bound is at least $u(\hat{a}, exit)$. □

For games satisfying the conditions of the proposition, the exit minmax condition is not necessary for bad reputation. The worst equilibrium continuation value that the short-run players could inflict is arbitrarily close to the exit payoff and hence a patient long run player could not be deterred from his temptation to avoid a bad reputation.

5. Poor reputation games and strong temptations

Recall that an action is vulnerable to a temptation if conditional on participation by the short-run players, the temptation lowers the probability of all bad signals, and increases the probability of all others. In this case the bad reputation result requires the exit minmax condition, as demonstrated by the example in Section 4.4. Notice, however, that in the example the relative probability of g and r is changed by the temptation. If the temptation satisfies the stronger property that the relative probability of the other signals remains constant, then we can weaken the assumption of exit minmax. In this section we develop this result, and give an application to games with two actions.

Define

$$\hat{u}^L = \max_{a,\beta \in \text{conhull}(E) \cap \text{image}(R)} u^L(a, \beta).$$

This is a bound on the long-run player's payoff when the short-run players play exit actions that are a best response to some (possibly incorrect) conjectures; it is this payoff (and not the possibly lower minmax payoff) that will provide a bound on the long-run player's equilibrium payoff under the assumptions of this section.

Definition 3S. An action a is *vulnerable to a strong temptation relative to a set of signals* $\widehat{Y}$ if there exists a number $\underline{\rho} > 0$ and an action d such that

(1) If $b \notin E$, $\hat{y} \in \widehat{Y}$ then $\rho(\hat{y} \mid d, b) \leqslant \rho(\hat{y} \mid a, b) - \underline{\rho}$.

(2) If $b \notin E$ and $y, y' \notin \widehat{Y} \cup Y^E$ then

$$\frac{\rho(y \mid d, b)}{\rho(y' \mid d, b)} = \frac{\rho(y \mid a, b)}{\rho(y' \mid a, b)}.$$

(3) For all $e \in E$, $u^L(d, e) \geqslant u^L(a, e)$.

The action d is called a *strong temptation*.

The first and third parts of this definition are the same as in the definition of a temptation; the additional strength comes from part (2), which requires that the temptation not merely increase the probability of all of the good signals, but leave their relative probabilities unchanged. Note that strong temptation is equivalent to temptation in games in which the set $Y \setminus (\widehat{Y} \cup Y^E)$ has a single element, for example games such as the EV example in which there are only two entry

signals. In particular applies when the game of Section 4.4 is modified so that the only signals when entry occurs are g and r.

This condition lets us prove an analog of Lemma 3:

Lemma 3S. *In a participation game, if* $\beta(h_t) \in \text{conhull}E$ *or* $\beta(h_t) \notin \text{conhull}E$ *and* $\alpha_0(h_t)(f) > 0$ *for some friendly action* f *that is vulnerable to a strong temptation size* $\underline{\rho}$*, then*

$$v(h_t) \leqslant \max_{y \in Y(h_t)} (1-\delta)\bar{u}(y, \underline{\rho}) + \bar{\delta}(y, \underline{\rho})\delta v(h_t, y),$$

where[12]

$$\bar{u}(y, \rho) = \begin{cases} (1 + \frac{1}{|\widehat{Y}|\rho})U & \text{if } y \in \widehat{Y}, \\ \hat{u}^L & \text{otherwise}, \end{cases}$$

and

$$\bar{\delta}(y, \rho) = \begin{cases} \delta(1 + \frac{1}{|\widehat{Y}|\rho}) & y \in \widehat{Y}, \\ \delta & \text{otherwise}. \end{cases}$$

The proof, which is in the online supplemental materials (Appendix B), follows that of Lemma 3, but takes advantage of the fact that the long-run player's continuation expected value, conditional on a friendly action, a non-exit profile, and a signal not in $\widehat{Y} \cup Y^E$, is the same for the equilibrium action and the strong temptation b.

Definition 6S. A participation game is a *poor reputation game* if it has exit minmax, and there are friendly and unfriendly sets F and N and a set of signals $\widehat{Y}$ that are evidence for N, such that every enforceable $f \in F$ is vulnerable to strong temptation relative to $\widehat{Y}$.

The next result says that poor reputation games have much the same consequences as bad reputation games.

Theorem 2. *In a poor reputation game of commitment size* $\gamma/2$*,* η*,*

$$\lim_{\delta \to 1} v^L \leqslant \hat{u}^L.$$

With Lemma 3S in hand, the proof of Theorem 2 is very close to that of Theorem 1, and is omitted. Notice that it is possible for a game to be both a bad reputation game and a poor reputation game, and, since strong and ordinary temptation are equivalent when $Y \setminus (\widehat{Y} \cup Y^E)$ is a singleton, the two are necessarily equivalent in this case. The original EV game is such an example. Notice also that Example 4.4 in which we construct a non-bad equilibrium has three signals rather than two. With two signals, the game would still fail the exit minmax condition and fail to be a bad reputation game, but it would still be a poor reputation game, and would not admit a good equilibrium. Finally, observe the proofs of both Lemmas 3 and 3S can be generalized, so that the difference between the best equilibrium payoff (in the limit as $\delta \to 1$) and the most favorable outcome with exit is bounded by a scale factor times the product of two terms, namely (i) the change in relative probabilities induced by a temptation and (ii) the excess of the best

[12] Note the abuse of notation: $\bar{u}, \bar{\delta}$ are different functions in Lemma 3S than in Lemma 3.

result given exit over the minmax. In particular, the bound on the difference is continuous in the each of these terms, so that if either is small the best equilibrium payoff for a patient long-run player can only exceed the best exit payoff by a small amount.

We turn now to the special case of two-player participation games where there is only one signal in $\widehat{Y}$ and short-run player payoffs depend only on the signal. We focus on the case where there is also one signal in $Y\backslash(\widehat{Y} \cup Y^E)$, so that bad reputation implies poor reputation. We show that these games are not poor reputation games (and by implication not bad reputation games either).

Proposition 4. *If a two-player participation game has only two "entry signals" (that is two elements of $Y - Y^E$), the short-run player has only two actions, and the short-run player's realized payoff is determined by the signal, then the game is not a poor reputation game.*

Proof. Notice that since the short-run player has only two actions, they correspond to "entry" and "exit" respectively. Consequently, the short-run player payoff conditional on entry depends only on the distribution over signals induced by the long-run player action. If we normalize the short-run player's payoff function so that his exit payoff is 0, and suppose that both the friendly and unfriendly sets are non-empty, then one signal yields a negative payoff and the other signal's payoff is positive; call these the "bad" and "good" signals respectively. If the game has no non-empty unfriendly set, it is not a poor reputation game; so we can suppose there is at least one non-empty unfriendly set. Any unfriendly set N consists of actions with a sufficiently high probability of sending the bad signal, and the bad signal (as a singleton set) is the only set $\widehat{Y}$ that can be evidence for N. Let f be the friendly action in the (finite) friendly set that maximizes the short-run player's payoff. The payoff to this action, conditional on it not generating the bad signal with the negative payoff, is positive, and since any temptation relative to $\widehat{Y}$ must reduce the probability of the bad signal, a temptation must give the short-run player a higher payoff than this "friendliest" friendly action. For this to be true, there must be a pure strategy d' that gives the short-run player at least this same utility. Clearly d' induces entry, and since it is a pure strategy, it must be in the friendly set. This contradicts the fact that f was assumed to maximize short-run player utility in the friendly set. □

We believe that the assumptions of this proposition imply that there is an equilibrium where the rational type's payoff is bounded below by a positive number as $\delta \to 1$ but we have not been able to show this.

6. Principal–agent entry games

In this class of games the only choice of the short-run player (the principal) is whether to enter or to exit. We denote his utility by u^S. If the principal enters, then the long-run player (the agent) chooses a payoff-relevant action; otherwise both players receive a reservation value which is normalized to zero. We assume there is an action $a \in A$ for which $u^L(a) \geqslant 0$, so that the exit minmax assumption is satisfied.[13]

For these games we can immediately identify the relevant friendly set, which is $F^* = \{a \in A\colon u^S(a) \geqslant 0\}$. To show that these are bad reputation games, it suffices to find an unfriendly

[13] This assumption will hold whenever the principal has the option to refuse to participate.

set that is orthogonal to F^* and associated evidence, such that every enforceable point in F is vulnerable to a temptation.

We begin with the *hidden information* case. Each period, nature draws a state $\omega \in \Omega$ independently from a probability distribution that we denote by p.[14] We assume Ω is a finite set and that p is strictly positive. The agent privately observes the state and then selects a decision $d \in D$. Conditional on the realized state and the decision of the agent, a signal $z \in Z$ is drawn from the distribution $m(z \mid \omega, d)$ where we assume that $m(z \mid \omega, d) > 0$ for each z, ω, and d. Future short run players observe both z and the decision d. Each player j has state-dependent utility function $\pi^j(\omega, d, z)$ and evaluates stage payoffs according to expected utility with respect to the distributions $p(\omega)$ and $m(z \mid \omega, d)$.

To apply Theorem 1, we find conditions under which this is a bad reputation game. The set of actions for the long-run player is the set of maps $a : \Omega \to D$. The stage-game utility function is

$$u^j(a) = \sum_{\omega \in \Omega} p(\omega) \sum_{z \in Z} m(z \mid \omega, d) \pi^j\big(\omega, a(\omega), z\big).$$

Finally $Y - Y^E = Z \times D$ and

$$\rho\big((z, d) \mid a, \mathit{entry}\big) = \sum_{\{\omega:\ a(\omega) = d\}} p(\omega) m(z \mid \omega, d).$$

Proposition 5. *The hidden information game is a bad reputation game if there exists a decision d such that $a \in F^*$ implies $\emptyset \neq \{\omega:\ a(\omega) = d\} \neq \Omega$.*

Proof. Suppose there exists a decision d such that $a \in F^*$ implies $\emptyset \neq \{\omega:\ a(\omega) = d\} \neq \Omega$.

Let $a(d)$ denote the constant action that chooses d regardless of the state ω, and take $N = \{a(d)\}$. Clearly $a(d) \notin F^*$, and hence N is an unfriendly set. Because $m(z \mid \omega, d) > 0$ for all z, ω, the set of signals $Y^d = Z \times \{d\}$ is evidence for N.

Let a be any pure friendly action, then $\Omega_d = \{\omega:\ a(\omega) = d\} \neq \emptyset$, and also $\Omega \setminus \Omega_d \neq \emptyset$. Let α' be the action that, regardless of ω, assigns probability 0 to d and equal probability to all other all decisions, and let c be the action defined by

$c(\omega) = \alpha'(\omega), \omega \in \Omega_d$ and $c(\omega) = a(\omega), \omega \notin \Omega_d$.

Since $\Omega \setminus \Omega_d \neq \emptyset$, decisions are observable, and $m(z \mid \omega, d) > 0$ for all z, ω, the friendly action a produces each signal in Y^d with positive probability. On the other hand, c reduces the probability of signals in Y^d to 0, and again since $m(z \mid \omega, d)$ is bounded away from 0, c increases the probability of every signal not in Y^d. Hence a is vulnerable to the temptation to play c. This establishes that the game is a bad reputation game. □

For illustration, consider the following extension of the EV example. If the correct repair is chosen, then the car works, otherwise it does not, and future motorists observe the car's performance. Formally, let $Z = \{\textbf{\textit{work}}, \textbf{\textit{not}}\}$. Suppose that for each motorist (independently and with equal probability), nature selects a necessary repair from the set $\{\textbf{\textit{tuneup}}, \textbf{\textit{engine}}\}$. Conditional on this, the mechanic observes state $\omega \in \{\textbf{\textit{tuneup}}, \textbf{\textit{engine}}\}$ representing the mechanic's diagnosis, and selects a repair $d \in \{\textbf{\textit{tuneup}}, \textbf{\textit{engine}}\}$. Suppose with some small probability $\varepsilon > 0$ that ω is incorrect, that is, not equal to the needed repair, so that $m(z \mid \omega, d) > 0$ for each z, ω, and d.

[14] This is a slight abuse of notation, as p also denotes the probability distribution over types in the incomplete-information games, but no ambiguity should result.

We can now define the motorists' payoff as a function of z, ω, and d to yield expected payoff function $u^S(a)$ identical to the original EV stage-game payoffs. Thus the friendly set (which is defined only in terms of the short-run player's payoffs) from EV remains so, and we can apply Proposition 5 using the decision $d = \textbf{\textit{engine}}$ to show that the modified game is a bad reputation game. However, as in the games of almost-perfect monitoring discussed in Section 4.2, there is an order of limits issue here: As $\varepsilon \to 0$, $\zeta \to 0$ as well, and for any fixed δ the payoff bound in the theorem becomes vacuous.[15]

In contrast to the hidden-action case, agency games with hidden actions tend not to be susceptible to bad reputation effects. The problem is that the second part of the definition of temptation typically fails, because deviations will generally lower the probability of some good signals. However, a special case in which a hidden-action game is a bad reputation game occurs when there is only one short-run player and only two non-exit signals, as in EV. The following proposition is an immediate application of the definition of a bad reputation game in this setting.

Proposition 6. *Suppose in a principal-agent entry game that $Y - Y^E = \{\underline{y}, \bar{y}\}$ and that $\hat{a}$ strictly maximizes the probability of $\underline{y}$ with $u^S(\hat{a}) < 0$. If for every friendly enforceable a there is a d such that $\rho(\underline{y} \mid d) < \rho(\underline{y} \mid a)$, then the game is a bad reputation game.*

To apply this result, suppose that the agent chooses an action from a one-dimensional set ordered so that higher actions are more likely to give rise to the high signal. Specifically, we let $A = \{\underline{a}, \ldots, \bar{a}\} \subset \Re$, assume that $\rho(\bar{y} \mid a)$ is increasing function in a, and assume that $u^L(a)$ is concave, so that F^* is an interval. Whether or not the game is a bad reputation game then depends on whether the principal prefers extreme or interior actions.

To show how the bad-reputation logic can extend to game with several short-run players, we now consider games with multiple principals. In these "multilateral entry" games, the short-run principals choose only whether to participate or exit. If any short-run player chooses to exit, that player receives the reservation payoff of 0, but play between the agent and other principals is unaffected. That is, $B^j = \{\textbf{\textit{exit}}, \textbf{\textit{enter}}\}$, and the unique exit profile is $e \equiv (\textbf{\textit{exit}}, \ldots, \textbf{\textit{exit}})$. The payoff of the short-run players who enter depends only on the action of the principal, and not on how many other short-run players chose to enter; to simplify notation we denote this "entry payoff" as $u^j(a)$. If all principals exit, the long-run player's payoff is 0; if m of them choose to enter, the long-run player's payoff is $u^L(a, m)$. We assume that the agent cannot be forced to participate, so that there exists an action a^1 such that for all m, $u^L(a^1, m) \geqslant 0$.

We do not require that $u^L(a, m)$ is linear in m, so this class of games includes those in which the agent has the opportunity to take a costly action prior to the entry decision of the short-run players. Consider for example, a game in which the long-run player is an expert advisor, and the decision of the short-run player is whether or not to pay the long-run player for advice. The advisor receives a report about the general desirability of various actions, and then meets with each of his n short-run customers, possibly learning about their individual needs. Here the advisor receives the signal regardless of whether or not he is consulted by any particular short-run player, and he may incur costs ahead of time for doing so. That is, the long-run player's payoff may depend on his action even if the short-run players decline to participate.

We have the following obvious extension of Proposition 5.

[15] This is not a game of almost-perfect monitoring because the signal does not reveal any information about what the mechanic would have done if the diagnosis had been different.

Proposition 7. *Suppose in a multilateral entry game that* $Y - Y^E = \{y^L, y^H\}$ *and that* $\hat{a}$ *strictly maximizes the probability of* y^L *with* $u^j(\hat{a}) < 0$. *If for every friendly enforceable* a *there is a* d *such that* $\rho(y^L \mid d) < \rho(y^L \mid a)$, *the game is a bad reputation game.*

As an illustration, return to the example in which the short-run players are students, the long-run player a teacher, and the signals are teaching evaluations. We denote the public signals representing good and bad evaluations by y^H or poor, y^L. We assume specific payoffs: If the evaluations are administered honestly and the class is taught well, there is probability 0.9 of a good evaluation. If evaluations are administered honestly and the class is taught poorly, the probability of good evaluations is only 0.1. Manipulating the evaluations is certain to lead to a good evaluation. All players get 0 if no students decide to take the class. For a short-run player who enters, the short run player's payoffs are $+1$ for good teaching and -1 for bad. Let m denote the number of students who take the class. The rational type of long-run player pays a cost of m to teach well; good evaluations are worth $2m$, while manipulating evaluations costs $3m$. Hence in the one-shot game with only the rational type, the unique sequential equilibrium is for the rational type to teach well and not manipulate the evaluations, for an expected payoff of 0.8.

However, when there is a small probability that the instructor is a bad type, and the instructor faces a sequence of short-run students, Proposition 7 applies. To see this, we see that teaching poorly and administering the evaluations honestly is the unfriendly action $\hat{a}^1$. The friendly set consists of the pure actions "teach well, administer honest evaluations" and "teach well, manipulate." As we pointed out earlier, the action "teach well, manipulate" is unenforceable, so the only enforceable action in the friendly set is "teach well, administer honestly." This admits the temptation "teach poorly, manipulate." Here the short-run player recognizes that if the long-run player chooses not to send the signal honestly, he loses his incentive to teach well, and so there is no reason to enter.

7. Concluding remarks

We have extended the EV bad reputation example to a broader class of games and prior distributions, and complemented these with cases where bad reputation does not take hold. Throughout the analysis, we have maintained the structure of Fudenberg and Levine (1989, 1992), with a single long-run player, whose type is fixed once and for all, building a reputation against a sequence of short-run opponents. It would be interesting to explore bad reputation in other sorts of set-ups, such as the games with multiple long-run players studied by Schmidt (1993) and Celentani et al. (1996); Morris (2001) analyzes an example of this without commitment types. It would also be interesting to study the extension of bad reputation to models where the long-run player's preferences can evolve over time, as in Mailath and Samuelson (2001).

Acknowledgments

We are grateful to Adam Szeidl and Maria Goltsman for careful proofreading, to Juuso Valimaki for helpful conversations, and to National Science Foundation Grants SES-9730181, SES-9986170, SES-9985462, SES-0112018, SES-0314713, and SES-0426199 for financial support.

Appendix A. Proofs

Lemma 1. *If h_t is a positive probability history in which $\hat{y} \in \hat{Y}$ occurs in period t and $\mu(h_{t-1})[\Theta(F)] \leqslant \gamma/2$ then $\alpha_0(h_t)(f) \geqslant \gamma/2$ for some friendly f.*

Proof. Given h_{t-1} the short-run players' profile has positive probability on a profile that does not exit. At such profiles $\bar{\alpha}(h_t)(f) \geqslant \gamma$ for some friendly f. Since $\mu(h_{t-1})[\Theta(F)] \leqslant \gamma/2$ we see that $\alpha_0(h_t)(f) \geqslant \gamma - \gamma/2 \geqslant \gamma/2$. □

We will let $p(\cdot \mid h_{t-1}) = \rho(\cdot \mid \bar{\alpha}(h_{t-1}), \beta(h_{t-1}))$ denote probability distributions over signals induced by the equilibrium strategies at history h_{t-1}; similarly

$$p\big(\cdot \mid \Theta(N), h_{t-1}\big) = \sum_{\theta \in \hat{\Theta}} \mu(h_t)[\theta] \rho\big(\cdot \mid a(\theta), \beta(h_{t-1})\big)$$

denotes the equilibrium distribution on signals conditional on θ being in the set $\hat{\Theta}$ of types that are committed to actions in N. The probability distribution on B induced by a mixed profile β can be written as a convex combination of the conditional distributions β_{-E}, which has support entirely in $B - E$, and β_E, which has support entirely in E. Then $\lambda\beta_{-E} + (1-\lambda)\beta$ induces the same distribution over B as $\beta(h_t)$, where $\beta(h_t) \notin \operatorname{conhull} E$, $\lambda > 0$. Although, β_{-E} and β_E need not correspond to mixed strategy profiles (since they can have correlation), we may still write $u^L(\alpha, \beta_E)$, $v(h_t, \alpha, \beta_E)$, $\rho(\cdot \mid \alpha, \beta_E)$ and so forth for the expected values of $u^L(\alpha, b)$, $v(h_t, \alpha, b)$, $\rho(\cdot|\alpha, b)$ with respect to the weights β_E, and similarly for β_{-E}. With that in mind, let $p(\cdot \mid \textit{entry}, h_{t-1}) = \rho(\cdot \mid \bar{\alpha}(h_{t-1}), \beta_{-E}(h_{t-1}))$ be the distribution of signals after history h_{t-1}, given that the realization of the short-run players' (equilibrium) action is an entry profile, and let $\rho(\cdot \mid a, \textit{entry}, h_{t-1}) = \rho(\cdot \mid a, \beta_{-E}(h_{t-1}))$.

Lemma 2. *In a bad reputation game, if h_t is a positive probability history with respect to a Nash equilibrium, and the signals in h_t all lie in $Y^E \cup \hat{Y}$*

(a) *At most $k^* = k_0 + \log(\mu_0[\Theta(N)])/\log(1 - \gamma + \gamma/r)$ of the signals are in $\hat{Y}$.*
(b) *If the commitment size is $\gamma/2$ then $\mu(h_t)[\Theta(F)] \leqslant \gamma/2$.*

Proof. First observe that if $\mu(h_{t-1})[\Theta(N)] \geqslant 1 - \gamma$, then the short-run players must exit in period t, so $\mu(h_t) = \mu(h_{t-1})$. Thus if h_t is a positive probability history in which $\hat{y}$ occurs in period t, $\mu(h_{t-1})[\Theta(N)] < 1 - \gamma$, and $\beta(h_{t-1})$ has positive probability of entry. Recall that

$$r = \min_{n \in N, a \notin N, \beta\{E\} < 1, \hat{y} \in \hat{Y}} \frac{\rho(\hat{y} \mid n, \beta)}{\rho(\hat{y} \mid a, \beta)}.$$

From Bayes' rule

$$\begin{aligned}
&\mu(h_t)\big[\Theta(N)\big] \\
&\quad = \frac{p(\hat{y} \mid \Theta(N), h_{t-1})}{p(\hat{y} \mid h_{t-1})} \mu(h_{t-1})\big[\Theta(N)\big] \\
&\quad = \frac{p(\hat{y} \mid \Theta(N), h_{t-1})}{p(\hat{y} \mid \Theta(N), h_{t-1})\mu(h_{t-1})[\Theta(N)] + p(\hat{y} \mid \neg\Theta(N), h_{t-1})(1 - \mu(h_{t-1})[\Theta(N)])} \mu(h_{t-1})\big[\Theta(N)\big] \\
&\quad = \frac{1}{\mu(h_{t-1})[\Theta(N)] + \frac{p(\hat{y}|\neg\Theta(N), h_{t-1})}{p(\hat{y}|\Theta(N), h_{t-1})}(1 - \mu(h_{t-1})[\Theta(N)])} \mu(h_{t-1})\big[\Theta(N)\big].
\end{aligned}$$

Substituting $\mu(h_{t-1})[\Theta(N)] < 1-\gamma$ and

$$\frac{p(\hat{y} \mid \neg\Theta(N), h_{t-1})}{p(\hat{y} \mid \Theta(N), h_{t-1})} < r$$

shows that

$$\mu(h_t)\big[\Theta(N)\big] \geqslant \frac{\mu(h_t)[\Theta(N)]}{1-\gamma+\frac{\gamma}{r}}.$$

Since signals in Y^E convey no information about the long-run player's type, it follows that if all signals lie in $Y^E \cup \widehat{Y}$, and signals in $\widehat{Y}$ occur k times, then

$$\mu(h_t)\big[\Theta(N)\big] \geqslant \left(\frac{1}{1-\gamma+\frac{\gamma}{r}}\right)^k \mu_0\big[\Theta(N)\big].$$

Hence if

$$k \geqslant -\frac{\log(1-\gamma) - \log(\mu_0[\Theta(N)])}{\log(1-\gamma+\frac{\gamma}{r})}$$

then $\mu(h_t)[\Theta(N)] \geqslant 1-\gamma$, so in all subsequent periods the signal must be an exit signal. (Recall that in a bad reputation game, $r > 1$; this implies that the denominator above is not zero.) Again because $r > 1$, it is sufficient that

$$k \geqslant \frac{-\log(\psi) + \log(\mu(0)[\Theta(N)])}{\log(1-\gamma+\frac{\gamma}{r})}$$

which is the condition in part (a).

We now turn to part (b). For any history h on the equilibrium path at which entry occurs with positive probability, we must have $\bar{\alpha}(h)(f) \geqslant \gamma$ for some friendly f. By assumption every enforceable friendly action is vulnerable to temptation, so that conditional on entry, the total probability of each bad signal $\hat{y} \in \widehat{Y}$ is at least $\varphi\gamma$ at such a history h. In particular, consider any history $h_{\tau-1}$ after which a bad signal $\hat{y}$ occurs. Since bad signals are entry signals, entry must have had positive probability at $h_{\tau-1}$, and hence conditional on entry, $\hat{y}$ had total probability at least $\varphi\gamma$. Bayes' rule then implies $\mu(h_\tau)[\theta] \leqslant (1/\varphi\gamma)\mu(h_{\tau-1})[\theta]$ for any type θ. Thus $\mu(h_\tau)[\Theta(F)] \leqslant (1/\varphi\gamma)\mu(h_{\tau-1})[\Theta(F)]$ for each τ in which a bad signal occurs. In particular

$$\mu(h_t)\big[\Theta(F)\big] \leqslant (1/\varphi\gamma)^k \mu(0)\big[\Theta(F)\big] \tag{1.1}$$

where k is the number of bad signals in h_t.

Bayes' rule gives us the following inequalities

$$\mu(h_\tau)\big[\Theta(N)\big] \geqslant \frac{\mu(h_{\tau-1})[\Theta(N)] \min_{n\in N} p(\hat{y} \mid n, \mathit{entry}, h_{\tau-1})}{p(\hat{y} \mid \mathit{entry}, h_{\tau-1})}$$

and

$$\begin{aligned}\mu(h_\tau)\big[\Theta(F)\big] &\leqslant \frac{\mu(h_{\tau-1})[\Theta(F)] \max_{a\in F} p(\hat{y} \mid a, \mathit{entry}, h_{\tau-1})}{p(\hat{y} \mid \mathit{entry}, h_{\tau-1})} \\ &\leqslant \frac{\mu(h_{\tau-1})[\Theta(F)] \max_{a\notin N} p(\hat{y} \mid a, \mathit{entry}, h_{\tau-1})}{p(\hat{y} \mid \mathit{entry}, h_{\tau-1})}\end{aligned}$$

where the last inequality comes from the assumption that F and N are disjoint.

Divide the first inequality by the second and apply the definition of r to get

$$\frac{\mu(h_\tau)[\Theta(N)]}{\mu(h_\tau)[\Theta(F)]} \geqslant r \frac{\mu(h_{\tau-1})[\Theta(N)]}{\mu(h_{\tau-1})[\Theta(F)]}$$

for each $\tau \in \{1, \ldots, t\}$ such that $y_\tau \in \widehat{Y}$. Since there are k such τ, we obtain

$$\frac{\mu(h_t)[\Theta(N)]}{\mu(h_t)[\Theta(F)]} \geqslant r^k \frac{\mu_0[\Theta(N)]}{\mu_0[\Theta(F)]}. \tag{1.2}$$

Finally, we define κ by the following equation

$$\frac{\mu_0[\Theta(N)]}{\mu_0[\Theta(F)]} r^\kappa = \left(\frac{\gamma}{2}\right)^{-1}. \tag{1.3}$$

Our commitment size assumption is that

$$\begin{aligned}
\mu_0\big[\Theta(F)\big] &\leqslant \left(\frac{\gamma}{2}\right)^{(1-\log(\gamma\varphi)/\log r)} \left[\frac{\mu_0[\Theta(N)]}{\mu_0[\Theta(F)]}\right]^{(\log(\frac{1}{\gamma\varphi})/\log(r))} \\
&= \left(\frac{\gamma}{2}\right)\left(\frac{1}{\gamma\varphi}\right)^{\frac{\log(\frac{\mu_0[\Theta(N)]}{\mu_0[\Theta(F)]})}{\log r}} \\
&= \left(\frac{\gamma}{2}\right)\left(\frac{1}{\gamma\varphi}\right)^{-\kappa}.
\end{aligned} \tag{1.4}$$

With these preliminaries in hand, we can conclude the proof. First suppose that

$$\frac{\mu(h_t)[\Theta(N)]}{\mu(h_t)[\Theta(F)]} > \left(\frac{\gamma}{2}\right)^{-1}.$$

Then it follows immediately that

$$\mu(h_t)\big[\Theta(F)\big] < \frac{\gamma}{2}$$

and we are done. On the other hand, the opposite inequality implies by (1.2) and (1.3) that $k \leqslant \kappa$. Consequently, by (1.1) and (1.4) we have

$$\mu(h_t)\big[\Theta(F)\big] \leqslant \frac{\gamma}{2}. \qquad \square$$

Recall that

$$\bar{u}(y, \tilde{\rho}) = \begin{cases} \left(1 + \frac{1}{\tilde{\rho}}\right) U^L & y \in \widehat{Y}, \\ 0 & \text{otherwise}, \end{cases}$$

$$\bar{\delta}(y, \tilde{\rho}) = \begin{cases} \frac{\delta}{\tilde{\rho}} + 1 & y \in \widehat{Y}, \\ \delta & \text{otherwise} \end{cases}$$

and $Y(h_t) = \{y \in Y^E \cup \widehat{Y} \mid \rho(y \mid \bar{\alpha}(h_t), \beta(h_t)) > 0\}$.

Lemma 3. *In a participation game if $\beta(h_t)\{E\} = 1$ or $\beta(h_t)\{E\} < 1$ and $\alpha_0(h_t)(f) > 0$ for some vulnerable friendly action f with temptation bounds $\underline{\rho}$, $\tilde{\rho}$ then*

$$v(h_t) \leqslant \max_{y \in Y(h_t)} (1-\delta)\bar{u}(y, \tilde{\rho}) + \bar{\delta}(y, \tilde{\rho}) v(h_t, y).$$

Proof. We need to calculate the long-run player payoff separately as a function of whether the short-run players exit or not. Using the decomposition of the short-run players' profile that we introduced before the proof of Lemma 2, we write

$$v(h_t) = \lambda v(h_t, f, \beta_E) + (1-\lambda)v(h_t, f, \beta_{-E}).$$

First assume $\lambda \geqslant 0$ and consider the value $v(h_t, f, \beta_E)$ conditional on exit. By exit minmax, this value is no more than $\max_{y\in Y(h_t)} \delta v(h_t, y)$ and thus from the definition of $\bar{u}$ we can conclude that

$$v(h_t, f, \beta_E) \leqslant \max_{y\in Y(h_t)} (1-\delta)\bar{u}(y, \tilde{\rho}) + \bar{\delta}(y, \tilde{\rho})v(h_t, y).$$

If $\beta(h_t)\{E\} = 1$ then $\lambda = 1$ and we are done with the first case in the statement.

Now consider the second case in the claim of the lemma $\beta(h_t)\{E\} < 1$ and $\alpha_0(h_t)(f) > 0$ for some vulnerable friendly action f with temptation bounds $\underline{\rho}, \tilde{\rho}$. In this case, $\lambda < 1$. Let d be a temptation for f. Since f is played in equilibrium, it earns at least as much as d, so that

$$\begin{aligned} &v\big(h_t, f, \beta(h_t)\big) - v\big(h_t, d, \beta(h_t)\big) \\ &\quad = \lambda\big[v(h_t, f, \beta_E) - v(h_t, d, \beta_E)\big] + (1-\lambda)\big[v(h_t, f, \beta_{-E}) - v(h_t, d, \beta_{-E})\big] \\ &\quad \geqslant 0. \end{aligned}$$

The expression $\lambda[v(h_t, f, \beta_E) - v(h_t, d, \beta_E)]$ is non-positive because given exit, d and f induce the same distribution over signals, and hence earn identical continuation values, and by the definition of a temptation, $u^L(d, \beta_E) \geqslant u^L(f, \beta_E)$, so that d does at least as well in the current period. Thus, $v(h_t, f, \beta_{-E}) - v(h_t, d, \beta_{-E}) \geqslant 0$. Expanding this inequality and using the fact that $u^L(f, \beta_{-E}) - u^L(d, \beta_{-E}) \leqslant U^L$,

$$\begin{aligned} &(1-\delta)U^L + \delta \sum_{\hat{y}\in\hat{Y}} \big[\rho\big(\hat{y} \mid f, \beta_{-E}\big) - \rho\big(\hat{y} \mid d, \beta_{-E}\big)\big]v(h_t, \hat{y}) \\ &\quad \geqslant \delta\Bigg[\sum_{y\in Y\setminus\hat{Y}} \big[\rho(y \mid d, \beta_{-E}) - \rho(y \mid f, \beta_{-E})\big]v(h_t, y)\Bigg]. \end{aligned} \tag{1.5}$$

Define

$$v\big(h_t, \hat{Y}\big) = \max_{y\in\hat{Y}} v(h_t, y) \geqslant 0.$$

The inequality holds because continuation values for histories on the equilibrium path of a Nash equilibrium must exceed the minmax value, which we have normalized to zero. We will use this fact repeatedly in the remainder of the proof. By the definition of a temptation, $\rho(\hat{y} \mid b, \beta_{-E}) < \rho(\hat{y} \mid f, \beta_{-E})$ for each $\hat{y}$. Thus, inequality (1.5) can be reduced to

$$\begin{aligned} (1-\delta)U^L + \delta v\big(h_t, \hat{Y}\big) &\geqslant \delta\Bigg\{\sum_{y\in Y\setminus\hat{Y}} \big[\rho(y \mid d, \beta_{-E}) - \rho(y \mid f, \beta_{-E})\big]v(h_t, y)\Bigg\} \\ &\geqslant \delta\tilde{\rho} \sum_{y\in Y\setminus\hat{Y}} \rho(y \mid f, \beta_{-E})v(h_t, y) \end{aligned}$$

where the second inequality uses part (2) of the definition of a temptation. We can now expand the definition of $v(h_t, f, \beta_{-E})$ and bound it as follows.

$$\begin{aligned}
&(1-\delta)u^L(f,\beta_{-E}) + \delta\left[\sum_{\widehat{y}\in\widehat{Y}} \rho\big(\widehat{y} \mid f,\beta_{-E}\big)v\big(h_t,\widehat{y}\big) + \sum_{y\notin\widehat{Y}} \rho(y \mid f,\beta_{-E})v(h_t,y)\right] \\
&\quad\leqslant (1-\delta)U^L + \delta v\big(h_t,\widehat{Y}\big) + \frac{(1-\delta)U^L}{\tilde{\rho}} + \frac{\delta}{\tilde{\rho}}v\big(h_t,\widehat{Y}\big) \\
&\quad= \max_{\widehat{y}\in\widehat{Y}} (1-\delta)\bar{u}\big(\widehat{y},\tilde{\rho}\big) + \left(\frac{\delta}{\tilde{\rho}}+1\right)v\big(h_t,\widehat{y}\big) \\
&\quad\leqslant \max_{y\in Y(h_t)} (1-\delta)\bar{u}(y,\tilde{\rho}) + \bar{\delta}(y,\tilde{\rho})v(h_t,y)
\end{aligned}$$

where the last inequality follows because when $\lambda < 1$, $\widehat{Y} \subseteq Y(h_t)$. (Recall that the vulnerability of f implies $\rho(\widehat{y} \mid f, b) > 0$ for each $\widehat{y} \in \widehat{Y}$ and each entry profile b.)

This concludes the proof because if $\lambda = 0$ then $v(h_t) = v(h_t, f, \beta_{-E})$ and if $\lambda \in (0,1)$ then $v(h_t) \leqslant \max\{v(h_t, f, \beta_E), v(h_t, f, \beta_{-E})\}$. □

Appendix B

Supplementary material associated with this article can be found, in the online version, at doi: 10.1016/j.geb.2006.08.007.

References

Celentani, M., Fudenberg, D., Levine, D.K., Pesendorfer, W., 1996. Maintaining a reputation against a long-lived opponent. Econometrica 64, 691–704.

Ely, J., Valimaki, J., 2003. Bad reputation. NAJ Econ. 4, 2; http://www.najecon.org/v4.htm. Quart. J. Econ. 118 (2003) 785–814.

Fudenberg, D., Kreps, D., 1987. Reputation in the simultaneous play of multiple opponents. Rev. Econ. Stud. 54, 541–568.

Fudenberg, D., Kreps, D., Maskin, E., 1990. Repeated games with long-run and short-run players. Rev. Econ. Stud. 57, 555–573.

Fudenberg, D., Levine, D.K., 1994. Efficiency and observability with long-run and short-run players. J. Econ. Theory 62, 103–135.

Fudenberg, D., Levine, D.K., 1992. Maintaining a reputation when strategies are imperfectly observed. Rev. Econ. Stud. 59, 561–579.

Fudenberg, D., Levine, D.K., 1989. Reputation and equilibrium selection in games with a patient player. Econometrica 57, 759–778.

Fudenberg, D., Maskin, E., Levine, D.K., 1994. The folk theorem with imperfect public information. Econometrica 62, 997–1039.

Kreps, D., Wilson, R., 1982. Reputation and imperfect information. J. Econ. Theory 27, 253–279.

Mailath, G., Samuelson, L., 2001. Who wants a good reputation. Rev. Econ. Stud. 68, 415–441.

Milgrom, P., Roberts, J., 1982. Predation, reputation, and entry deterrence. J. Econ. Theory 27, 280–312.

Morris, S., 2001. Political correctness. J. Polit. Economy 109, 311–323.

Schmidt, K., 1993. Reputation and equilibrium selection in repeated games of conflicting interests. Econometrica 61, 325–351.

Sorin, S., 1999. Merging, reputation, and repeated games with incomplete information. Games Econ. Behav. 29, 274–308.