When is Reputation Bad?

Jeffrey Ely

Drew Fudenberg

David K. Levine

11/13/02

traditional reputation theory

- Kreps and Wilson [1982], Milgrom and Roberts [1982], Fudenberg and Levine [1992]
- gang-of-four type model with long run versus short-run player
- reputation is good for the long-run player through imitating commitment type

“bad reputation”

- Ely and Valimaki [2001] give example in which reputation is unambiguously bad
- this paper tries to determine in what class of games reputation is bad
  - participation is optional for the short-run players
  - 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”
- broaden the set of commitment types, allowing many types, including the “Stackelberg type”

The Dynamic Game

$N+1$ players, long run-player 1, N short-run players $2 \ldots N+1$

game begins at $t=1$ and is infinitely repeated

each period, each player i chooses from finite action space A^i

use a^{-i} to denote the play of all players except player i

long-run player discounts future with discount factor δ

each short-run player plays only in one period - is replaced by an identical short-run player next period

set Θ of types of long-run player

type $0 \in \Theta$ "rational type"

for each pure action a^1, type $\theta(a^1)$ is a "committed type"

no other types in Θ

stage game utility functions are $u^i(a)$, where $u^1(a)$ corresponds to the long-run player of type $\theta = 0$

common prior distribution over long-run player types is denoted $\mu(0)$.

a finite public signal space Y with signal probabilities $\rho(y \mid a)$

all players observe the history of the public signals

short-run players observe only the history of the public signals

- observe neither the past actions of the long-run player, nor of previous short-run players

do not assume payoffs depend on actions only through signals, so the short-run players at date *t* need not know the realized payoffs of the previous generations of short-run players

let $h_t = (y_1, y_2 \ldots, y_t)$ denote public history through end of period t

null history is 0

h_t^1 denote private history known only to long-run player; includes own actions, and may or may not include the actions of the short-run players he has faced in the past

strategy for the long-run player is sequence of maps
$\sigma^1(h_t, h_t^1, \theta) \in \text{conhull}\, A^1 \equiv \mathcal{A}^1$

strategy profile for short-run players is a sequence of maps
$\sigma^j(h_t) \in \text{conhull}\, A^j \equiv \mathcal{A}^j$.

short-run profile α^{-1} is Nash response to α^1 if
$u^i(\alpha^1, \alpha^i, \alpha^{-1-i}) \geq u^i(\alpha^1, a^i, \alpha^{-1-i})$ for all $a^i \in A^i$

set of short-run Nash responses to α^1 is $B(\alpha^1)$.

given strategy profiles σ, the prior distribution over types $\mu(0)$ and a public history h_t that has positive probability under σ, we can calculate from σ^1 the conditional probability of long-run player actions $\overline{\alpha}^1(h_t)$ given the public history

Nash Equilibrium is a strategy profile σ such that for each positive probability history

1)$\sigma^{-1}(h_t) \in B(\overline{\alpha}^1(h_t))$ [short-run players optimize]

2)$\sigma^1(h_t, h_t^1, \theta(a^1)) = a^1$ [committed types play accordingly]

3)$\sigma^1(\cdot, \cdot, 0)$ is a best-response to σ^{-1} [rational type optimizes].

The Ely-Valimaki Example

long-run player a mechanic

action a map from the privately observed state of the customer's car $\omega \in \{E, T\}$ to announcements $\{e, t\}$

E means the car needs a new engine, *T* means it needs at tune-up

the announcements, which are what the mechanic says the car needs, determine what is actually done to the car

$A^1 = \{ee, et, te, tt\}$,first component announcement in response to signal *E*

one short-run player each period chooses $A^2 = \{In, Out\}$

public signal $Y = \{e, t, Out\}$

short-run player chooses Out the signal is Out

otherwise the signal is the announcement of the long-run player

two states of the car i.i.d. and equally likely

short-run player chooses Out, everyone gets 0

short-run plays In and long-run player's announcement is truthful
short-run player receives u; untruthful receives $-w$

$w > u > 0$

"rational type" of long-run player has *exactly* the same stage-game payoff function as the short run players

	In	*Out*
ee	$(u-w)/2, (u-w)/2$	$0, 0$
et	u, u	$0, 0$
te	$-w, -w$	$0, 0$
tt	$(u-w)/2, (u-w)/2$	$0, 0$

rational type the only type in the model

an equilibrium where he chooses the action that matches the state, all short-run players enter, and the rational type's payoff is u

EV example

there is a probability that long-run player is a “bad type” who always plays ee

long-run player's payoff is bounded by an amount that converges to 0 as the discount factor goes to 1

Participation Games and Bad Reputation Games

"participation games" short-run players may choose not to participate

crucial aspect of non-participation is that it conceals the action taken by the long-run player from subsequent short-run players

certain public signals $y^e \in Y^e$ are *exit signals*

associated with these exit signals are *exit profiles*, which are pure action profiles $e^{-1} \in E^{-1} \subseteq A^{-1}$ for the short run players.

for each exit profile e, $\rho(y^e \mid a^1, e^{-1}) = \rho(y^e \mid e^{-1})$ for all a^1, and $\rho(Y^e \mid e^{-1}) = 1$

moreover, if $a^{-1} \notin E^{-1}$ then $\rho(y^e \mid a^1, a^{-1}) = 0$ for all $a^1 \in A^1, y^e \in Y^e$

participation game is a game in which $E^{-1} \neq \varnothing$

Definition 1: A non-empty finite set of pure actions for the long-run player N^1 is *unfriendly* if there is a number $\psi < 1$ such that $\alpha^1(N^1) \geq \psi$ implies $B(\alpha^1) \subseteq \text{conhull } E^{-1}$.

unfriendly actions induce exit

in EV example the set $\{ee, tt, te\}$ is unfriendly, and so is any subset.

Definition 2: A non-empty finite set of mixed actions F^1 for the long run player is friendly if there is a number $\gamma > 0$ such that $B(\alpha^1) \cap \left[\mathcal{A}^{-1} - \text{conhull}(E^{-1})\right] \neq \varnothing$ implies $\alpha^1 \geq \gamma f^1$ for some $f^1 \in F^1$. The number γ is called the size of the friendly set

actions that induce entry must put weight on a friendly action

may be many different friendly sets

in EV example, the action et is friendly, with

$$\underline{\alpha} = \frac{w - u}{w + u/2}.$$

Definition 3: The *support* $A^1(F^1)$ of a friendly set F^1 are the actions that are played with positive probability:

$$A^1(F^1) \equiv \{a^1 \in A^1 \mid f^1(a^1) > 0, f^1 \in F^1\}$$

We say that a friendly set F^1 is *orthogonal* to an unfriendly set N^1 if $N^1 \cap A^1(F^1) = \varnothing$

Definition 4: We say that a set of signals $\widehat{Y}$ is *unambiguous* for a set of actions N^1 if for all $a^{-1} \notin E^{-1}, \widehat{y} \in \widehat{Y}, n^1 \in N^1, a^1 \notin N^1$ we have $\rho(\widehat{y} \mid n^1, a^{-1}) > \rho(\widehat{y} \mid a^1, a^{-1})$.

every action in N^1 must assign a higher probability to each signal in $\widehat{Y}$ than any action not in N^1

a given set of actions may not have signals that are unambiguous

in the EV example, E is an unambiguous signal for the unfriendly set $\{ee\}$

Definition 5: An action a^1 is *vulnerable to temptation relative to a set of signals* $\widehat{Y}$ if there exist numbers $\underline{\rho}, \tilde{\rho} > 0$ and an action b^i such that

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

2. If $a^{-1} \notin E^{-1}$ and $y \notin \widehat{Y} \cup Y^e$ then $\rho(y \mid b^1, a^{-1}) \geq (1 + \tilde{\rho})\rho(y \mid a^1, a^{-1})$.

3. For all $e^{-1} \in E^{-1}$, $u^1(b^1, e^{-1}) \geq u^1(a^1, e^{-1})$.

The action b^1 is called a temptation, and the parameters $\underline{\rho}, \tilde{\rho}$ are the temptation bounds.

in EV example, the action et is vulnerable relative to $\{E\}$: the temptation b^i is *tt*, which sends the probability of the signal E to zero. (Since there is one other signal, condition 2 of the definition is immediate.)

Definition 6: A mixed action α^1 for the long run player is *enforceable* if there does not exist another action $\tilde{\alpha}^1$ such that for all $a^{-1} \in E^{-1}$, $u^1(\tilde{\alpha}^1, a^{-1}) \geq u^1(\alpha^1, a^{-1})$ and for all $a^{-1} \in A^{-1} - E^{-1}$, $u^1(\tilde{\alpha}^1, a^{-1}) > u^1(\alpha^1, a^{-1})$ and $\rho(\cdot \mid \tilde{\alpha}^1, a^{-1}) = \rho(\cdot \mid \alpha^1, a^{-1})$. When α^1 is not enforceable, we say that the action $\tilde{\alpha}^1$ defeats α^1.

Definition 7: A participation game has an *exit minmax* if

$$\max_{\alpha^{-1} \in E^{-1} \cap range(B)} \max_{\alpha^1} u^1(\alpha^1, \alpha^{-1}) =$$
$$\min_{\alpha^{-1} \in range(B)} \max_{\alpha^1} u^1(\alpha^1, \alpha^{-1})$$

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 range of *B.* It is convenient in this case to normalize the minmax payoff to 0

Definition 8: A participation game is a *bad reputation game* if it has an exit minmax, there is an unfriendly set N^1, a friendly set F^1 that is orthogonal to N^1, and a set of signals $\widehat{Y}$ that are unambiguous for N^1, and such that every enforceable $f^1 \in F^1$ is vulnerable to temptation relative to $\widehat{Y}$.

The signals $\widehat{Y}$ are called the *bad signals*.

the EV game is a bad reputation game

the friendly set $\{et\}$

the unfriendly set $\{ee\}$

the unfriendly signals $\{E\}$

constants describing a bad reputation game

ψ is the probability in the definition of an unfriendly set

γ is the scale factor in the definition of a friendly set

since the friendly set is finite, define $\varphi > 0$ to be the minimum, taken over elements of the friendly set, of the values $\underline{\rho}$ in the definition of temptation

$$r = \min_{n^1 \in N^1, a^1 \notin N^1, \alpha^{-1} \notin \mathrm{conhull}(E^{-1}), \widehat{y} \in \widehat{Y}} \frac{\rho(\widehat{y} \mid n^1, \alpha^{-1})}{\rho(\widehat{y} \mid a^1, \alpha^{-1})}.$$

since friendly set non-empty and orthogonal to the unfriendly set denominatoris well defined

since $\widehat{Y}$ is unambiguous for the unfriendly set, $r > 1$

$$\eta = -\log(\gamma\varphi) / \log r, \qquad k_0 = -\frac{\log(\psi)}{\log\left(\psi + (1-\psi)\frac{1}{r}\right)}.$$

The Theorem

what it means for unfriendly types to be likely "enough"

$\Theta(F^1)$ be the commitment types corresponding to actions in the support of F^1; we will call these the *friendly commitment types.* Let $\hat{\Theta}$ be the *unfriendly commitment types* corresponding to the unfriendly set N^1.

Definition 9: A bad reputation game has *commitment size* ε, ϕ if

$$\mu(0)[\Theta(F^1)] \leq \varepsilon \left(\frac{\mu(0)[\widehat{\Theta}]}{\mu(0)[\Theta(F^1)]} \right)^{\phi}$$

where $\phi > 0$.

places a bound on the prior probability of friendly commitment types that depends on the prior probability of the unfriendly types

since ϕ is positive, the larger the prior probability of $\widehat{\Theta}$, the larger the probability of the friendly commitment types is allowed to be

assumption of a given commitment size does not place any restrictions on the *relative* probabilities of commitment types

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

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 $\lambda\tilde{\mu}$.

for sufficiently small λ the assumption of commitment size ε, η is satisfied

$$U^1 = \max_a u^1(a) - \min\{0, u^1\}$$

$$\tilde{\rho}_{\min} = \min_{f^1 \in F^1} \tilde{\rho}$$

$$\underline{F}^1 \equiv \min\{f(a) > 0 \mid f \in F^1\}$$

Theorem 1: In a bad reputation game of commitment size $\left((\gamma \underline{F}^1 / 2)^{(1+\eta)}, \eta\right)$ let $\overline{v}^1$ be the supremum of the payoff of the rational type in any Nash equilibrium. Then

$$\overline{v}^1 \leq (1-\delta) k^* \left(\frac{1}{\tilde{\rho}_{\min}}\right)^{k^*} \left(1 + \frac{1}{\tilde{\rho}_{\min}}\right) U^1,$$

where $k^* = k_0 + \log\left(\mu(0)[\widehat{\Theta}]\right) / l \log\left(\psi + (1-\psi)\frac{1}{r}\right)$. In particular, $\lim_{\delta \to 1} \overline{v}^1 \leq 0$.

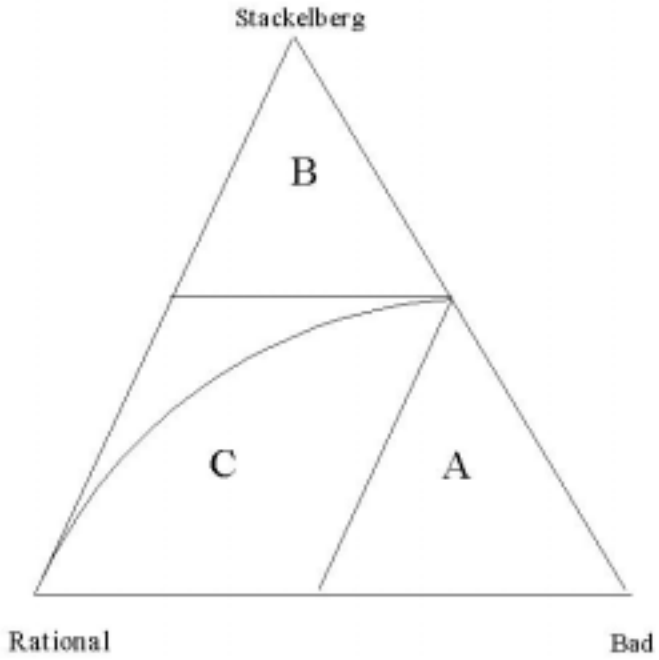
Examples

EV With Stackelberg Type

relaxed original assumptions of EV in a number of ways

allow for positive probabilities of all commitment types including "Stackelberg type" committed to the honest strategy et, which is the optimal commitment

suppose in particular that there are 3 types, rational, bad, and Stackelberg


Stackelberg
B
C
A
Rational
Bad

region A probability of the bad type is too high, and short run players refuse to enter regardless of the behavior of the rational type; long-run player obtains the minmax payoff of zero

EV prior assigned probability zero to the Stackelberg type; prior and all posteriors on the equilibrium path belong to the lower boundary of the simplex

region B sufficiently high probability of the Stackelberg type, the short-run players will enter regardless of the behavior of the rational type; long-run player gets nearly u as discount factor approaches 1

region C where our theorem applies 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

Adding an Observed Action to EV

add new observable action "*g*" for the long-run player called “give away money.”

induces the short-run players to participate so it must be in every friendly set

action is observable, so not vulnerable to temptation with respect to any signals that are unambiguous for the unfriendly actions, so this is not a bad reputation game

an equilibrium where the rational type plays *g* in the first period

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

the assumption that every friendly action is vulnerable to temptation is seen to be both important and economically restrictive.

Principal-Agent Entry Games

single short-run player (the principal) whose only choice is whether to enter or to exit

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

$A^2 = \{exit, enter\}$

$u^2(a^1, exit) = 0$ for each $a^1 \in A^1$

write $u^2(a^1, enter) = u^2(a^1)$

an action $a^1 \in A^1$ for which $u^1(a^1) \geq 0$, so that the exit minmax assumption is satisfied

will hold whenever the principal has the option to refuse to participate

Games with Hidden Information

each period, nature draws a state $\omega \in \Omega$ independently from a probability distribution that we denote by *p*

agent privately observes the state and selects a decision $d \in D$

signal $z \in Z$ is drawn from $m(z \mid \omega, d) > 0$

future short run players observe both z and the decision d

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 $m(z \mid \omega, d)$.

Proposition 5: The hidden information game is a bad reputation game if there exists a decision d such that $a^1 \in F^1$ implies $\varnothing \neq \{\omega : a^1(\omega) = d\} \neq \Omega$.

a decision taken sometimes but not always by all friendly actions

example of extension to EV

if the correct repair is chosen, then the car works, otherwise it does not, and this outcome is observed by future motorists

Proposition 5 implies that as long as the mechanic's diagnosis is not perfect, the game is a bad reputation game

so why are advisors employed?

costly signal, or is hired not for the advice but for its implementation. a country might know the advice that the IMF will recommend, but find it useful to delegate the implementation of the advice so that it can avoid taking full responsibility for the resulting hardships.

some advice games aren't participation games because the advisor can make "speeches" even without any "customers;" this may describe political advisors and investment columnists.

some of the short-run players are "naïve" and enter even when entry is not a best response to equilibrium play. These "noise players" ensure that the long player can build a track record

discount factors may not be close to 1

Rules Rather than Discretion

college admissions

university (the long-run agent) receives an application

applicant is described by a set of characteristics $\omega \in \Omega = \Omega^o \times \Omega^n$. Some ($\Omega^o$) of these characteristics are publicly observable (for example race and SAT scores) and others (Ω^n) are observed only by the university

pure strategy for the university is a map from characteristics to the decision space *D*= (admit, deny)

probability of drawing characteristics ω is $p(\omega) > 0$

university's preferences over applicants are summarized by the payoff function $\pi^1(\omega)$ if the student is admitted, and R if the student is denied

short-run principal is the state governor who chooses between allowing the university discretion in admissions, or imposing a rigid admission rule based on observable characteristics

many possible rules that the principal might use, but since she is a short-run player we can restrict attention to the rule that maximizes the principal's expected short-run payoff. This rule is a mapping $g : \Omega^o \rightarrow D$ that mandates admission if and only if $\omega^o \in g^{-1}(admit)$

imposition of a rigid admission rule represents "exit"

governor shares the same preferences as the university, receiving a utility of $\pi^1(\omega)$ for admits and R for rejects.

university can always implement *g* its own so exit minmax condition is satisfied

for discretion to improve upon *g* for some set of verifiable characteristics, the admission decision should depend on the unverifiable characteristics

by essentially the same argument as in Proposition 5, the game is a bad reputation game with unfriendly set

$$\{a^1 : a^1(\omega^o, \cdot) = deny\}$$

for example, ω^o may be racial characteristics, and the type associated with this unfriendly set represents the governor's fear that the university admissions are biased against members of the race in question

Mulilateral Entry Games

multiple principals

short-run players choose only whether to participate or exit

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

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; denote this "entry payoff" as $u^j(a^1)$

all principals exit, the long-run player's payoff is 0

m of them choose to enter, the long-run player's payoff is $u^1(a^1, m)$

agent cannot be forced to participate, so that there exists an action a^1 such that for all m, $u^1(a^1, m) \geq 0$

do not require that $u^1(a^1, m)$ 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

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

- EV example of car repairs, where the long-run player is able to determine the type of repair the car needs
- stockbrokers advising clients on portfolio choices
- doctors advising patients on treatments
- IMF advising countries on economic policies

EV example private information emerges as a consequence of the decision of the short-run player to consult the long-run player, so the advice is specific to the short-run player

generally, at least some part of the information is not specific to the short-run player

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

the advisor may receive 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.

costs incurred on exit are consistent with a bad reputation game provided that conditional on exit the temptations are less costly than the friendly actions

long-run player might be a stockbroker, and the general non-client specific information might be something about general economic conditions, acquired in advance in the form of economic reports that will be presented to the client

friendly actions in this case are to report truthfully; the bad action might be to always claim that times are good. In this case the temptation is to announce that times are bad when they are actually good, to avoid being mistaken for the type that always announces good times. If it is costly to put together a persuasive package of economic data indicating that times are bad when in fact they are good this would not be a bad reputation game. If it is more costly to put together an honest report, then it would be a candidate for a bad reputation game.

following obvious extension of Proposition 5.

Proposition 8: Suppose in a multilateral entry game that $Y - Y^e = \{y^L, y^H\}$ and that $\widehat{a}^1$ strictly maximizes the probability of y^L with $u^j(\widehat{a}^1) < 0$. If for every friendly enforceable a^1 there is a b^1 such that $\rho(y^L \mid b^1) < \rho(y^L \mid a^1)$ the game is a bad reputation game.

another example:

short-run players are students, long-run player a teacher, and the signals are teaching evaluations

could apply equally well to the decision to attend a particular college, graduate school, or take a particular job

short-run player decides whether to enter - that is, take the class, or not

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

public signals are whether the evaluations (averaged over respondents) are good, y^H or poor, y^L

evaluations are administered honestly and the class is taught well, there is probability .9 of a good evaluation

evaluations are administered honestly and the class is taught poorly, the probability of good evaluations is only .1

Manipulating the evaluations is certain to lead to a good evaluation

all players get 0 if no students decide not to take the class

short-run player who enters, the short run player's payoffs are +1 for good teaching and -1 for bad

m denote the number of students who take the class

rational type of long-run player pays a cost of m to teach well; good evaluations are worth $2m$, while manipulating evaluations costs $3m$

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 .8.

when there is a small probability that the instructor is a bad type, and the instructor faces a sequence of short-run students, Proposition 8 applies

the action “teach well, manipulate” is unenforceable: teach poorly and manipulate yields a higher stage game payoff and the same distribution over signals

the only enforceable action in the friendly set is “teach well, administer honestly”

admits the temptation “teach poorly, manipulate”

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