



Honesty, stigma, and cooperation in an overlapping-generations game

David Li¹ · Georgy Lukyanov² 

Received: 11 November 2025 / Accepted: 4 August 2026

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2026

Abstract

We study a two-period overlapping-generations game in which a young agent interacts with an old agent and carries a binary record into old age. Honest agents cooperate with clear partners, whereas strategic agents have a privately known loss from being exploited. A clear record is informative: honest agents remain clear, while some strategic agents acquire stigma. We incorporate this selection into Bayes-consistent stationary equilibrium beliefs. With uniformly distributed losses, strategic young agents use a cutoff rule and equilibrium reduces to a quadratic equation. A simple parameter region has a unique equilibrium; when the temptation to defect is high, low- and high-cooperation corner equilibria coexist with an interior tipping equilibrium. Along that interior branch, a larger honest share lowers both the strategic cutoff and the young cooperation rate. The branch is locally unstable, however, so the result describes a moving coordination threshold rather than a deterioration of stable equilibrium selection. Under a maximin decision rule, strategic cooperation disappears. Probabilistic record clearing weakens incentives at every fixed conjecture, reduces cooperation at locally stable interior equilibria, and under full clearing leaves only defection by strategic agents.

Keywords Overlapping generations · Reputation · Community enforcement · Prisoner's Dilemma · Stigma · Honest types

JEL classification: C72 · C73 · D83.

We thank Ronaldo Carpio, the editor, and an anonymous referee for helpful comments and suggestions. Georgy Lukyanov acknowledges support from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements d'Avenir programme).

✉ Georgy Lukyanov
georgy.lukyanov@tse-fr.eu

David Li
dsli@pku.edu.cn

¹ School of Economics, Peking University, Beijing, China

² Toulouse School of Economics, Toulouse, France

1 Introduction

Cooperation often has to survive a change of partners: a junior supplier deals with an established buyer, a new professional works with a senior member of another organization, or a first-time platform participant is matched with an incumbent.¹ The relationship may be short, but conduct within it can affect the terms on which the junior is treated once she becomes the incumbent. This creates an intergenerational reputational incentive even though no pair ever meets twice. How much discipline can a single directed bit of memory extract from agents who would otherwise have every reason to exploit a partner they will never see again—and is a larger share of honest commitment types always good news for cooperation?

We study this question in a deliberately sparse environment. Each agent acts once when young and once when old. A young agent who defects against a cooperating old opponent receives a binary stigma that is observed by the agent's next partner. Some agents are honest commitment types: they cooperate with a clear opponent and defect against a stigmatized one. The remaining agents are strategic and differ privately in the loss they suffer when they cooperate against a defector. Strategic old agents defect because they have no future interaction. Strategic young agents may nevertheless cooperate in order to enter old age with a clear record.

The label changes not only behavior but also beliefs. Starting from clear records, honest agents never become stigmatized, whereas strategic agents can. Consequently, an old agent with a clear record is more likely to be honest than a randomly selected member of the population. Our equilibrium analysis accounts for this selection explicitly: a Perfect Bayesian treatment of the model requires it, and doing so changes the fixed-point equation relative to one that simply keeps the honest probability at its unconditional value.

We show that stationary symmetric cutoff equilibria admit a full characterization once losses are uniformly distributed. Bayes' rule turns the indifference condition into a quadratic equation, with exact conditions for the zero- and unit-cutoff corners and a closed-form expression for every interior cutoff. If the temptation payoff is no larger than the golden ratio, the equilibrium is unique. If it is larger than two, there is a transparent parameter region with three equilibria: two strict corners and an interior cutoff between them.

The interior branch is where intuition needs the most care. Common sense offers two guesses about what a larger honest share should do to cooperation, and neither survives intact. The first guess is that more honesty can only help: if opponents are more often trustworthy, cooperation should rise. The second guess, once the interior comparative static is in view, overcorrects in the other direction and treats honesty as actively harmful. What is shown instead is narrower, and more interesting than either. Within the three-equilibrium region, a larger share of honest agents does lower the interior cutoff, the probability of cooperation with a clear opponent, and the unconditional cooperation rate *along that branch*. But the branch is the tipping point separating the

¹ Throughout, the record is *directed* in a specific sense: it is created by one match and consulted by exactly one later match, never broadcast to the community at large. This is more restrictive than the community-wide or stationary record structures usual in the literature on bounded and public memory; Sect. 2 situates the comparison.

two strict corner equilibria, not a stable outcome in its own right, and a lower tipping point enlarges the set of conjectures from which the high-cooperation corner is reached. The comparative static is a fact about where the boundary sits, not an unconditional verdict on whether honesty helps or hurts.

We then examine two robustness questions. First, we compare Bayesian expected-payoff choice with a maximin rule. A strategic young agent who evaluates each action at the worst possible current opponent type always defects: the worst-case payoff from cooperation is lower by the private exploitation loss, for every strictly positive loss. Second, we allow a newly created stigma to be cleared before the next match. Clearing makes defection more attractive at every fixed conjecture: it lowers cooperation at locally stable interior equilibria and raises an unstable tipping cutoff, and if records are always cleared, every strategic young agent defects. A short extension shows that discounting weakens reputational incentives through the same channel, provided the current temptation payoff is not mistakenly discounted along with the future one.

The paper is related to, but distinct from, our earlier work (Lukyanov & Li, 2025). Both papers use honest and strategic types and private losses to obtain cutoff behavior. The earlier paper is a static two-player Prisoner's Dilemma and asks how common versus heterogeneous beliefs about a partner's honesty affect cooperation. The present paper instead has common knowledge of the primitive honest share, two-period lives, an endogenous record distribution, Bayes updating from records, and a reputational link between successive partners. Neither overlapping generations nor record-based community enforcement appears in the earlier paper: the question there is what agents believe about a fixed partner, while the question here is what a population can sustain once beliefs are pinned down by an endogenous, one-bit history. These differences define the contribution of the present paper.

The remainder of the paper proceeds as follows. Section 2 relates the model to the literature. Section 3 presents the environment, terminology, and applications. Section 4 characterizes stationary equilibrium. Section 5 studies cooperation and stability. Section 6 gives the maximin benchmark. Section 7 studies record clearing and discounting. Section 8 concludes.

2 Related literature

Community-enforcement models show how information about past conduct can support cooperation when partners are randomly matched. Labels and social norms can transmit punishments beyond a bilateral relationship (Ellison, 1994; Kandori, 1992). More recent work establishes both folk-theorem and anti-folk-theorem results in anonymous matching environments (Deb et al., 2020; Sugaya & Wolitzky, 2020). Our agents are not long-lived participants in an indefinitely repeated game. They have only one young and one old interaction, and a one-bit record links those interactions.

The record structure connects the paper to work on bounded memory and the design of reputational histories. Coarse or finite records can be essential for community enforcement (Bhaskar & Thomas, 2019), while the updating rule and the amount of information retained determine which cooperative outcomes can be sustained (Clark et al., 2021). Record deletion can also shape experimentation and trade (Kovbasyuk

& Spagnolo, 2024), and agents' ability to manipulate records introduces additional incentive constraints (Pei, 2026). Our contribution is narrower: we solve a two-period model in which a directed record is seen only by the next partner, and we make explicit the selection of types induced by that record.

Overlapping-generations games study how incentives and information survive cohort turnover (Lagunoff & Matsui, 2004). The alternating two-group structure also recalls models of intergroup interaction such as Acemoglu and Wolitzky (2014), although our stage game, individual labels, and source of dynamics are different. Related reputation models show how an individual or firm can carry incentives across career stages (Board & Meyer-ter Vehn, 2013; Tadelis, 2002). Here the record is not traded and has value for only one subsequent interaction.

Finally, honest agents are commitment types in the tradition of Kreps et al. (1982) and resemble the intrinsically rule-following agents discussed by Frank (1987). Evolutionary models provide one foundation for moral preferences (Alger & Weibull, 2013); we instead take the honest share as a primitive. Section 6 also relates the Bayesian solution to maximin choice under ambiguity (Gilboa & Schmeidler, 1989).

3 Model, terminology, and applications

3.1 Demography, matching, and records

Time is discrete. There are two populations, A and B , each containing a continuum of agents of unit mass in every cohort. An agent belongs to the same population throughout life, is young at date t , and is old at date $t + 1$. At each date, every young agent from A is matched uniformly and anonymously with an old agent from B , and every young agent from B is matched with an old agent from A . Equal cohort masses ensure that every agent has exactly two interactions.

Every entrant begins with a clear record. Before actions are chosen, a young agent observes whether the old opponent is clear or stigmatized. Actions are simultaneous. If a player chooses D while the opponent chooses C , the defector receives a stigma. For a young defector, that label is observed by the next-period partner. A label acquired by an old agent has no strategic consequence because the old agent exits after the match. No identity or richer history is revealed (Fig. 1).

3.2 Types, information, and payoffs

Each agent is honest with probability $\pi \in [0, 1]$ and strategic with probability $1 - \pi$. Types are independent across agents and the common prior π is the same in both populations. An honest agent follows a commitment rule: choose C against a clear opponent and D against a stigmatized opponent.

A strategic agent has a private exploitation loss ℓ_i . It is drawn at entry from the uniform distribution on $[0, 1]$ and is observed by agent i before the young-period action is chosen. Thus the decision is made after ℓ_i is known. The material payoff to

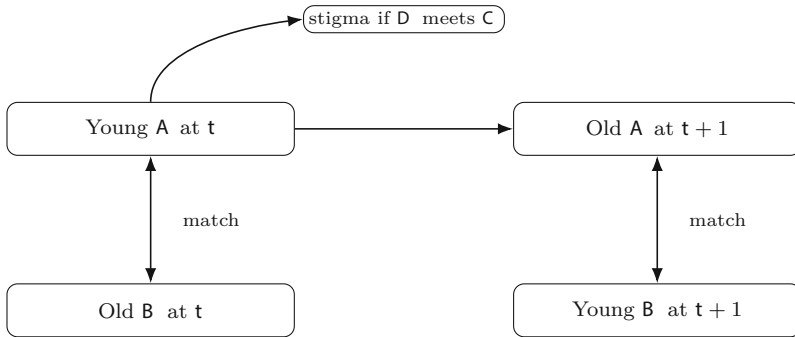


Fig. 1 The directed record. Conduct when young can affect only the action of the agent's next partner

a strategic player is

	opponent D	opponent C	
D	0	b	$b > 1.$
C	$-\ell_i$	1	

In the last interaction, D weakly dominates C and is strictly better for every strategic type with $\ell_i > 0$. We select D for the measure-zero zero-loss type when that type is indifferent. Hence every strategic old agent defects. A strategic young agent also defects against a stigmatized old opponent: in equilibrium a stigmatized old agent is known to be strategic and therefore chooses D . The same tie-breaking convention applies when $\ell_i = 0$. The only nontrivial decision is made by a strategic young agent facing a clear old opponent.

Payoffs from the old interaction are initially undiscounted. Section 7.2 introduces a discount factor. We focus on stationary symmetric equilibria: the two groups use the same strategy and induce the same record distribution at every date (Table 1).

3.3 Institutional interpretations

The two dates should be read as two roles or career stages, not necessarily as biological generations. Three interpretations fit the model particularly well.

First, in apprenticeship and professional networks, a junior may work once with a senior from another organization and later become the senior member in a new cross-organizational project. A verified incident of free riding, misappropriation, or nonpayment can be communicated to the next collaborator. The private loss ℓ_i captures how exposed the junior would be if trust were not reciprocated.

Second, in family-firm or local supplier networks, a new manager first deals with an incumbent counterpart and later represents the firm in a new match. The record belongs to the manager or firm and affects only the next transaction. The model abstracts from richer repeated relationships in order to isolate that single reputational link.

Third, a recommendation or matching system can implement the directed record. A platform or professional association verifies whether one party exploited a cooperating

Table 1 Terminology

Term	Definition and role
Young agent	An entrant in the first of two interactions; the only agent whose current action can affect a future payoff
Old agent	An agent in the final interaction; a strategic old agent always defects
Honest agent	A commitment type who cooperates with a clear opponent and defects against a stigmatized opponent
Strategic agent	A payoff-maximizing type with private exploitation loss ℓ_i
Clear record	No recorded defection against a cooperator. A clear old agent may be honest or strategic
Stigma	A one-bit record created by choosing D against C and observed by the defector's next partner
Cutoff x	A strategic young agent facing a clear old opponent cooperates if and only if $\ell_i \leq x$

partner and displays a binary warning to the next recommended counterparty. A clear participant may be recommended for a trust-intensive interaction, whereas a flagged participant receives a defensive response. This is not a literal model of a platform with indefinitely lived sellers and long review histories; it is a benchmark for systems that retain one verified incident for one subsequent match.

4 Stationary equilibrium

Suppose strategic young agents use cutoff $x \in [0, 1]$ against a clear old opponent. Because losses are uniform, a strategic young agent cooperates with probability x in such a match.

Lemma 1 *Under a stationary cutoff x , honest old agents are clear. A strategic old agent is stigmatized with probability*

$$z(x) = \pi(1 - x)$$

and is clear with probability

$$a(x) = 1 - z(x) = 1 - \pi + \pi x.$$

The total mass of clear old agents and the posterior probability that a clear old agent is honest are, respectively,

$$m(x) = \pi + (1 - \pi)a(x), \quad h(x) = \frac{\pi}{m(x)}. \quad (1)$$

A clear old agent is therefore more likely to be honest than an unconditioned agent whenever $\pi \in (0, 1)$ and $x < 1$.

If the focal strategic agent reaches old age with a clear record, the next young opponent cooperates with probability

$$q(x) = \pi + (1 - \pi)x. \quad (2)$$

Honest young agents cooperate with a clear old opponent, and a fraction x of strategic young agents do so. A stigmatized strategic old agent instead faces defection with probability one.

Consider a strategic young agent with loss ℓ who observes a clear old opponent. Bayes' rule assigns probability $h(x)$ to the event that the old opponent is honest. Cooperation yields

$$U_C(\ell; x) = h(x) - (1 - h(x))\ell + bq(x), \quad (3)$$

whereas defection yields

$$U_D(\ell; x) = h(x)b + (1 - h(x))bq(x). \quad (4)$$

If the old opponent is honest, defection gives b now but creates stigma. If the old opponent is strategic, both agents defect, no stigma is created, and the young agent retains the continuation value $bq(x)$.

The payoff difference is

$$\Delta(\ell; x) = h(x)[1 - b + bq(x)] - (1 - h(x))\ell. \quad (5)$$

It is strictly decreasing in ℓ . The best-response cutoff to conjecture x is consequently

$$T(x) = \left[\frac{\pi[1 - b + bq(x)]}{(1 - \pi)a(x)} \right]_0^1, \quad (6)$$

where $[y]_0^1 = \min\{1, \max\{0, y\}\}$. A stationary symmetric Perfect Bayesian equilibrium is a fixed point $x = T(x)$ together with the strategies described above and the Bayes-consistent beliefs in (1).

For an interior fixed point, define

$$\begin{aligned} g(x; \pi, b) &= \pi[1 - b + bq(x)] - (1 - \pi)a(x)x \\ &= A + Bx - Cx^2, \end{aligned} \quad (7)$$

where

$$A = \pi[1 - b(1 - \pi)], \quad B = (1 - \pi)[b\pi - (1 - \pi)], \quad C = \pi(1 - \pi). \quad (8)$$

Proposition 1 Fix $b > 1$ and $\pi \in (0, 1)$. A stationary symmetric cutoff equilibrium exists. Moreover:

(i) $x = 0$ is an equilibrium if and only if

$$\pi \leq 1 - \frac{1}{b}.$$

(ii) $x = 1$ is an equilibrium if and only if $\pi \geq \frac{1}{2}$.

(iii) Every interior equilibrium is a root of $g(x; \pi, b) = 0$. Let

$$\mathcal{D} = B^2 + 4CA.$$

Whenever $\mathcal{D} \geq 0$, the candidate roots are

$$x_{\pm} = \frac{B \pm \sqrt{\mathcal{D}}}{2C}; \quad (9)$$

an interior root is an equilibrium precisely when it lies in $(0, 1)$.

The root formula provides a complete characterization for all $b > 1$ and $\pi \in (0, 1)$. Two regions are especially transparent.

Proposition 2 Let $\varphi = (1 + \sqrt{5})/2$.²

- (i) If $1 < b \leq \varphi$, the equilibrium is unique. It is $x = 0$ when $\pi \leq 1 - 1/b$, the root x_+ when $1 - 1/b < \pi < 1/2$, and $x = 1$ when $\pi \geq 1/2$.
 (ii) If $b > 2$ and

$$\frac{1}{2} < \pi < 1 - \frac{1}{b},$$

there are exactly three equilibria: $x = 0$, $x = 1$, and the interior root $x_I = x_-$.

The bound φ is a sufficient global uniqueness condition, not a claim that multiplicity begins sharply at φ for every honest share. The root formula shows that additional multiplicity configurations can arise for some $b \in (\varphi, 2]$. Proposition 2(ii) isolates the high-temptation region used for the main comparative static.

5 Cooperation, comparative statics, and stability

We distinguish conditional cooperation from its unconditional frequency. Given cutoff x , the probability that a young agent cooperates conditional on meeting a clear old opponent is $q(x)$ in (2). The probability that a random old opponent is clear is

$$m(x) = 1 - \pi + \pi q(x).$$

The unconditional rate at which young agents choose C is therefore

$$R(x) = m(x)q(x). \quad (10)$$

This measure is the endogenous cooperation rate studied below. It does not count the mechanically specified actions of honest old agents.

Proposition 3 Suppose $b > 2$ and $\pi \in (1/2, 1 - 1/b)$. Along the interior equilibrium $x_I(\pi)$:

$$\frac{dx_I}{d\pi} < 0, \quad \frac{dq(x_I)}{d\pi} < 0, \quad \frac{dR(x_I)}{d\pi} < 0.$$

Under iterated best responses $x_{n+1} = T(x_n)$, the interior equilibrium is locally unstable. The two strict corner equilibria are locally stable, and x_I is the tipping cutoff separating their basins.

The proposition gives a precise sense in which a larger honest share can be associated with less cooperation in equilibrium: both conditional and unconditional young cooperation fall along the interior equilibrium branch. Stability changes the interpretation. Because x_I itself falls, a smaller initial conjecture about strategic cooperation is

² The bound is exactly the positive root of $b^2 - b - 1 = 0$: the sufficiency argument in the proof needs $b^2 - b - 1 \leq 0$, and φ is where that quadratic first turns positive. The appearance of the golden ratio is this algebraic coincidence and carries no independent economic content.

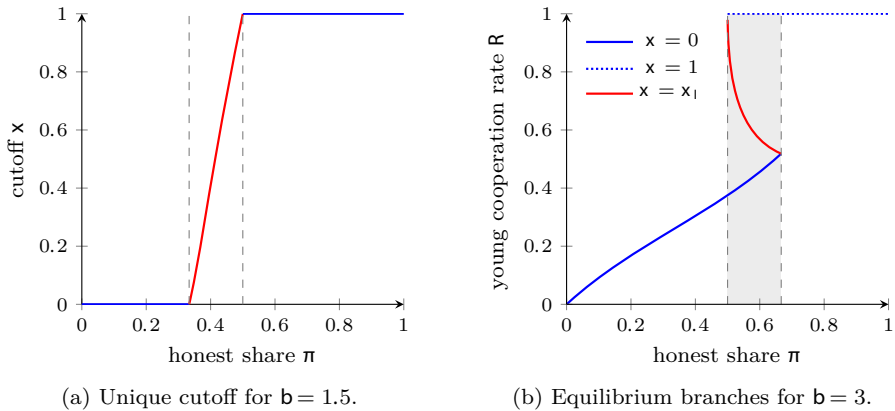


Fig. 2 Uniform-loss benchmark. Panel (a) shows the unique equilibrium when temptation is moderate. Panel (b) shades the three-equilibrium region. The interior young-cooperation rate decreases with π , while the strict corner branches do not

sufficient to move best responses toward $x = 1$. More honesty lowers the coordination hurdle even though cooperation measured *at* the hurdle declines.³

Figure 2 illustrates both regimes requested in the interpretation. For $b = 1.5$, the unique cutoff moves from complete strategic defection through an interior range to complete strategic cooperation. For $b = 3$, the shaded region contains the two corner equilibria and the declining interior tipping branch.

6 A robust worst-case benchmark

The baseline uses expected payoffs and the Bayes posterior $h(x)$. As a robust alternative, suppose a strategic young agent evaluates each action at the worst possible current old-opponent type, while taking the future cooperation probability $q(x)$ as given. If the agent defects, the payoff is b against an honest old opponent and $bq(x)$ against a strategic old opponent. Because $q(x) \leq 1$, the worst-case payoff is $bq(x)$. If the agent cooperates, the two payoffs are $1 + bq(x)$ and $-\ell + bq(x)$, so the worst-case payoff is $-\ell + bq(x)$.

Proposition 4 *Under the worst-case rule described above, every strategic young agent with $\ell > 0$ defects against a clear old opponent. The robust cutoff is $x = 0$.*

The result is deliberately stark. Bayesian cooperation relies on assigning positive probabilities to the benefits and reputational consequences of each opponent type. A player who instead protects against the worst type action by action regards cooperation as yielding exactly ℓ less in the worst case. If the future cooperation probability is itself treated as ambiguous, taking its worst value only reinforces the conclusion.

³ The interior cutoff plays the same role here as the threshold separating basins of attraction in models of equilibrium selection under noisy learning (Kandori et al., 1993). We do not pursue a full stochastic-stability analysis—there is no mutation process here—but that literature likewise highlights the role of an unstable equilibrium as a coordination boundary.

7 Record clearing and discounting

7.1 Probabilistic record clearing

Suppose a stigma created when young is cleared before the old-period match with probability $\rho \in [0, 1]$. Under a conjectured cutoff x , a strategic old agent is clear with probability

$$a_\rho(x) = 1 - (1 - \rho)\pi(1 - x).$$

Conditional on observing a clear old opponent, the honest posterior becomes

$$h_\rho(x) = \frac{\pi}{\pi + (1 - \pi)a_\rho(x)}.$$

The mass of clear old agents and the resulting unconditional young-cooperation rate are

$$m_\rho(x) = \pi + (1 - \pi)a_\rho(x), \quad R_\rho(x) = m_\rho(x)q(x). \quad (11)$$

Cooperation still guarantees a clear record. Following defection against an honest old opponent, the continuation value $bq(x)$ is restored with probability ρ . The cooperation-defection payoff difference is therefore

$$\Delta_\rho(\ell; x) = h_\rho(x)[1 - b + (1 - \rho)bq(x)] - (1 - h_\rho(x))\ell. \quad (12)$$

The associated best-response cutoff is

$$T_\rho(x) = \left[\frac{\pi[1 - b + (1 - \rho)bq(x)]}{(1 - \pi)a_\rho(x)} \right]_0^1. \quad (13)$$

For every fixed x , $T_\rho(x)$ is weakly decreasing in ρ : clearing reduces the reputational cost of defection and also enlarges the pool of clear strategic old agents.

For an interior equilibrium, define

$$g_\rho(x) = \pi[1 - b + (1 - \rho)bq(x)] - (1 - \pi)a_\rho(x)x. \quad (14)$$

The corner conditions are

$$x = 0 \text{ is an equilibrium} \iff 1 - b + (1 - \rho)b\pi \leq 0, \quad (15)$$

$$x = 1 \text{ is an equilibrium} \iff 2\pi - 1 - b\pi\rho \geq 0. \quad (16)$$

Proposition 5 *At a locally stable interior equilibrium, the cutoff and both cooperation measures q and R_ρ strictly decrease with ρ . At a locally unstable interior tipping equilibrium, the cutoff increases with ρ , so the high-cooperation basin becomes harder to reach. If $\rho = 1$, the unique strategic cutoff is $x = 0$.*

Thus the pointwise incentive effect of forgiveness is unambiguous, but a blanket comparison between arbitrarily selected equilibria would not be. Distinguishing stable outcomes from tipping equilibria resolves the apparent ambiguity.

7.2 Discounting

Let $\delta \in [0, 1]$ discount the old-period payoff. Current payoffs are not discounted, since the temptation to exploit an honest opponent is realized immediately and only the reputational return to staying clear lies in the future. The payoff difference becomes

$$\Delta_{\delta}(\ell; x) = h(x)[1 - b + \delta bq(x)] - (1 - h(x))\ell,$$

and an interior equilibrium solves

$$g_{\delta}(x) = \pi[1 - b + \delta bq(x)] - (1 - \pi)a(x)x = 0.$$

Proposition 6 *If $\delta = 0$, the unique stationary equilibrium is $x = 0$. At every locally stable interior equilibrium, the cutoff is strictly increasing in δ .*

Lower patience therefore weakens cooperation through the same channel already used for record clearing.⁴ The result cannot be recovered by replacing b everywhere with δb : doing so would also discount the immediate payoff from exploiting an honest old opponent, which is realized before the agent has any old age left to discount.

8 Conclusion

We studied a minimal intergenerational reputation mechanism, deliberately stripped down: agents interact once when young and once when old, a single bit of stigma links the two matches, and no identity or deeper history is ever recorded. Honest agents follow a commitment rule; strategic young agents trade off current exploitation against the value of entering old age with a clear record.

Accounting for the information contained in a clear record is essential. Since honest agents remain clear while some strategic agents do not, the honest probability among clear old agents differs from the population share; ignoring this selection produces beliefs that are inconsistent with the intended Bayesian equilibrium. With that selection incorporated, equilibrium reduces to a quadratic cutoff equation: a simple region has a unique equilibrium, and a high-temptation region has two strict corners coexisting with an interior tipping equilibrium.

What is maintained, and qualified, is the comparative static that motivated the exercise. A larger honest share does lower cooperation along the interior branch—but that branch is unstable, and its cutoff separates the basins of the two corner equilibria rather than describing a locally stable outcome. The right reading is a shift in a coordination threshold, not a verdict that honesty is counterproductive. The robust maximin benchmark is more clear-cut: it eliminates strategic cooperation altogether, since an agent who insures against the worst case treats the exploitation loss as a sure cost with

⁴ The comparative-static proof parallels the record-clearing argument, but discounting leaves $a(x)$ unchanged and scales only the continuation payoff. Unlike clearing, it does not change which old agents count as clear.

no offsetting sure benefit. Record clearing weakens the incentive to preserve a clear record at every fixed conjecture, and under full clearing strategic agents always defect.

Taken together, the results suggest a qualified design lesson rather than a simple one: even a one-bit, directed record can discipline behavior on its own, without repeated bilateral interaction or community-wide monitoring, but its informational content, its persistence, and its role in equilibrium selection have to be analyzed jointly.⁵ A system that clears records to appear forgiving, or that fails to update beliefs about who reaches old age clear, may therefore dissipate the incentive the record was meant to create.

Funding Georgy Lukyanov acknowledges support from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements d’Avenir programme).

Data availability No datasets were generated or analysed during the current study.

Code availability Not applicable.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethics approval Not applicable.

A Proofs

A.1 Proof of Lemma 1

Proof Every entrant is clear. An honest young agent cooperates with a clear old opponent. Against a stigmatized old opponent, the honest young agent defects; in equilibrium the stigmatized old agent is strategic and also defects. Hence an honest young agent never defects against a cooperator and remains clear when old.

A strategic young agent becomes stigmatized precisely when two events occur: the old opponent is honest, which has unconditional probability π , and the strategic loss exceeds x , which has probability $1 - x$. Thus $z(x) = \pi(1 - x)$ and $a(x) = 1 - z(x)$. The mass of clear old agents is the mass π of honest agents plus the mass $(1 - \pi)a(x)$ of clear strategic agents. Bayes’ rule then gives (1). $\square$

A.2 Proof of Proposition 1

Proof Equation (5) is strictly decreasing in ℓ , so the best reply has a cutoff. Substituting (1) into the zero of (5) yields the unclipped expression in (6). The clipped map $T : [0, 1] \rightarrow [0, 1]$ is continuous, so Brouwer’s theorem gives a fixed point.

At $x = 0$, the fixed-point condition is $T(0) = 0$, or

$$g(0; \pi, b) = A = \pi[1 - b(1 - \pi)] \leq 0.$$

⁵ Minimal records are attractive precisely because they are cheap to build and hard to game, but that same minimalism means small design choices—what gets recorded, for how long, and how beliefs are updated from it—do a disproportionate share of the disciplining work.

For $\pi > 0$, this is equivalent to $\pi \leq 1 - 1/b$. At $x = 1$, the condition $T(1) = 1$ is equivalent to

$$g(1; \pi, b) = 2\pi - 1 \geq 0.$$

For an interior fixed point clipping is inactive, so $g(x; \pi, b) = 0$. Expanding gives (7)–(8); the quadratic formula gives (9). Conversely, every retained root or valid corner satisfies the corresponding best-response condition. $\square$

A.3 Proof of Proposition 2

Proof The function $g(x) = A + Bx - Cx^2$ is strictly concave because $C = \pi(1 - \pi) > 0$.

Suppose first that $1 < b \leq \varphi$. If $\pi < 1/2$ and $A > 0$, then $g(0) > 0 > g(1)$, so strict concavity gives a unique root in $(0, 1)$ and neither corner is an equilibrium. If $A \leq 0$, then $\pi \leq 1 - 1/b$ and

$$B = (1 - \pi)[(b + 1)\pi - 1] \leq \frac{1 - \pi}{b}(b^2 - b - 1) \leq 0.$$

Thus g decreases from a nonpositive value and $x = 0$ is the only equilibrium. If $\pi \geq 1/2$, then $g(1) \geq 0$. Moreover, $1 - 1/b < 1/2$ for $b < 2$, so $A > 0$. Concavity implies that g is positive between its endpoint values (apart from a possible equality at $x = 1$), and $x = 1$ is the unique equilibrium. This proves part (i).

For part (ii), $b > 2$ and $1/2 < \pi < 1 - 1/b$ imply $g(0) < 0 < g(1)$. Hence both corners satisfy their strict best-response inequalities. Strict concavity gives one root in $(0, 1)$ and a second root above one. The interior root is the smaller root x_- in (9); no other fixed point exists. $\square$

A.4 Proof of Proposition 3

Proof At $x_I = x_-$,

$$g_x(x_I) = B - 2Cx_I = \sqrt{D} > 0.$$

Direct differentiation gives

$$g_\pi(x) = 1 + 2(1 - \pi)x + (2\pi - 1)x^2 + b(2\pi - 1)(1 - x) > 0$$

throughout the three-equilibrium region. Hence

$$\frac{dx_I}{d\pi} = -\frac{g_\pi(x_I)}{g_x(x_I)} < 0.$$

Since $q = \pi + (1 - \pi)x_I$,

$$\frac{dq}{d\pi} = \frac{(1 - x_I)g_x - (1 - \pi)g_\pi}{g_x}.$$

Using $g(x_I) = 0$, the numerator simplifies to

$$\frac{1-\pi}{\pi} \left[\pi^2 x_I^2 - (1-\pi+2\pi^2)x_I - \pi(1-\pi) \right].$$

The expression in square brackets is a convex quadratic in x_I . Its values at zero and one are $-\pi(1-\pi)$ and -1 , respectively, so it is negative on $[0, 1]$. Therefore $dq/d\pi < 0$.

The clear mass can be written as $m = 1 - \pi + \pi q$. Because $q < 1$ and $dq/d\pi < 0$,

$$\frac{dm}{d\pi} = -1 + q + \pi \frac{dq}{d\pi} < 0.$$

It follows from $R = mq$ that $dR/d\pi < 0$.

Finally, at an interior fixed point,

$$g_x = (1-\pi)a(x)[T'(x) - 1].$$

Thus $g_x(x_I) > 0$ implies $T'(x_I) > 1$, so the interior equilibrium is locally unstable. At each strict corner, clipping makes the best-response map constant in a sufficiently small one-sided neighborhood, establishing local stability. The sign of g changes from negative to positive at x_I , so iterated best responses move toward zero below the threshold and toward one above it. $\square$

A.5 Proof of Proposition 4

Proof The worst-case payoff from D is $\min\{b, bq(x)\} = bq(x)$. The worst-case payoff from C is $\min\{1 + bq(x), -\ell + bq(x)\} = -\ell + bq(x)$. Defection is strictly preferred for every $\ell > 0$; the zero-loss type is indifferent and has measure zero. $\square$

A.6 Proof of Proposition 5

Proof Let $\widehat{T}_\rho(x)$ denote the unclipped expression in (13). Direct differentiation gives

$$\frac{\partial \widehat{T}_\rho(x)}{\partial \rho} = -\frac{\pi[\pi + x(b-\pi)]}{(1-\pi)a_\rho(x)^2} < 0.$$

Clipping therefore makes $T_\rho(x)$ weakly decreasing in ρ .

At an interior equilibrium,

$$\frac{\partial g_\rho}{\partial \rho} = -\pi bq(x) - \pi(1-\pi)x(1-x) < 0.$$

Implicit differentiation gives

$$\frac{dx}{d\rho} = -\frac{\partial g_\rho / \partial \rho}{\partial g_\rho / \partial x}.$$

At a locally stable interior fixed point, $\partial g_\rho / \partial x < 0$, so $dx/d\rho < 0$. At a locally unstable tipping fixed point, $\partial g_\rho / \partial x > 0$, so $dx/d\rho > 0$. Since q is strictly increasing in x , it decreases with ρ at a stable interior equilibrium.

It remains to account for the direct increase in the clear-record mass in (11). Put $k = 1 - \rho$, $s = 1 - \pi$, and

$$H = 1 - k\pi(1 + b - 2x).$$

At a locally stable interior fixed point, $\partial g_\rho / \partial x = -sH < 0$, so $H > 0$; moreover $H < 1$. Implicit differentiation and (11) yield

$$\frac{dR_\rho}{d\rho} = \frac{\pi}{H} \{s(1-x)qH - [m_\rho + k\pi q][bq + sx(1-x)]\}.$$

Now $m_\rho > s(1-x)$, because $m_\rho = 1 - \pi sk(1-x)$ and $s(1-x)(1 + \pi k) < 1$. Consequently,

$$b[m_\rho + k\pi q] > s(1-x)H.$$

The expression in braces is therefore strictly negative, proving that R_ρ decreases with ρ along a stable interior branch.

If $\rho = 1$, then

$$g_1(x) = \pi(1-b) - (1-\pi)x < 0$$

for all $x \in [0, 1]$. The only fixed point is therefore $x = 0$. $\square$

A.7 Proof of Proposition 6

Proof At $\delta = 0$,

$$g_\delta(x) = \pi(1-b) - (1-\pi)a(x)x.$$

Since $b > 1$, the first term is negative, and $(1-\pi)a(x)x \geq 0$ throughout $[0, 1]$. Equivalently, the unclipped best-response cutoff is negative for every conjecture x , so the clipped best response satisfies $T_\delta(x) = 0$ throughout. Its unique fixed point is therefore $x = 0$.

For the comparative static, $\partial g_\delta / \partial \delta = \pi bq(x) > 0$. As in the proof of Proposition 3, $\partial g_\delta / \partial x = (1-\pi)a(x)[T'_\delta(x) - 1]$ at an interior fixed point. Local stability therefore implies $\partial g_\delta / \partial x < 0$. Implicit differentiation then gives

$$\frac{dx}{d\delta} = -\frac{\partial g_\delta / \partial \delta}{\partial g_\delta / \partial x} > 0$$

at every locally stable interior equilibrium. $\square$

References

- Acemoglu, D., & Wolitzky, A. (2014). Cycles of conflict: An economic model. *American Economic Review*, 104(4), 1350–1367. <https://doi.org/10.1257/aer.104.4.1350>
- Alger, I., & Weibull, J. W. (2013). Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica*, 81(6), 2269–2302. <https://doi.org/10.3982/ECTA10637>
- Bhaskar, V., & Thomas, C. (2019). Community enforcement of trust with bounded memory. *The Review of Economic Studies*, 86(3), 1010–1032. <https://doi.org/10.1093/restud/rdy048>
- Board, S., & Meyer-ter Vehn, M. (2013). Reputation for quality. *Econometrica*, 81(6), 2381–2462. <https://doi.org/10.3982/ECTA9039>
- Clark, D., Fudenberg, D., & Wolitzky, A. (2021). Record-keeping and cooperation in large societies. *The Review of Economic Studies*, 88(5), 2179–2209. <https://doi.org/10.1093/restud/rdab016>
- Deb, J., Sugaya, T., & Wolitzky, A. (2020). The folk theorem in repeated games with anonymous random matching. *Econometrica*, 88(3), 917–964. <https://doi.org/10.3982/ECTA16680>
- Ellison, G. (1994). Cooperation in the prisoner's dilemma with anonymous random matching. *The Review of Economic Studies*, 61(3), 567–588. <https://doi.org/10.2307/2297904>
- Frank, R. H. (1987). If homo economicus could choose his own utility function, would he want one with a conscience? *The American Economic Review*, 77(4), 593–604.
- Gilboa, I., & Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2), 141–153. [https://doi.org/10.1016/0304-4068\(89\)90018-9](https://doi.org/10.1016/0304-4068(89)90018-9)
- Kandori, M. (1992). Social norms and community enforcement. *The Review of Economic Studies*, 59(1), 63–80. <https://doi.org/10.2307/2297925>
- Kandori, M., Mailath, G. J., & Rob, R. (1993). Learning, mutation, and long run equilibria in games. *Econometrica*, 61(1), 29–56. <https://doi.org/10.2307/2951777>
- Kovbasyuk, S., & Spagnolo, G. (2024). Memory and markets. *The Review of Economic Studies*, 91(3), 1775–1806. <https://doi.org/10.1093/restud/rdad067>
- Kreps, D. M., Milgrom, P., Roberts, J., & Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, 27(2), 245–252. [https://doi.org/10.1016/0022-0531\(82\)90029-1](https://doi.org/10.1016/0022-0531(82)90029-1)
- Lagunoff, R., & Matsui, A. (2004). Organizations and overlapping generations games: Memory, communication, and altruism. *Review of Economic Design*, 8(4), 383–411. <https://doi.org/10.1007/s10058-003-0111-1>
- Lukyanov, G., & Li, D. (2025). Belief diversity and cooperation. *Journal of Economic Behavior & Organization*, 229, 106815. <https://doi.org/10.1016/j.jebo.2024.106815>
- Pei, H. (2026). Community enforcement with endogenous records. *The Review of Economic Studies*, 93(2), 1241–1260. <https://doi.org/10.1093/restud/rdaf049>
- Sugaya, T., & Wolitzky, A. (2020). A few bad apples spoil the barrel: An anti-folk theorem for anonymous repeated games with incomplete information. *American Economic Review*, 110(12), 3817–3835. <https://doi.org/10.1257/aer.20200068>
- Tadelis, S. (2002). The market for reputations as an incentive mechanism. *Journal of Political Economy*, 110(4), 854–882. <https://doi.org/10.1086/340781>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.