WORKING PAPERS

N° 1687

October 2025

"Intergroup cooperation and reputation for honesty in an OLG framework"

David Li and Georgy Lukyanov

Toulouse School of Economics

Intergroup cooperation and reputation for honesty in an OLG framework*

David Li† Georgy Lukyanov‡

Abstract

This paper studies an infinite-horizon framework in which two large populations of players are randomly matched to play a Prisoner's Dilemma. Each player lives for two consecutive periods: as a young player from one group, and then as an old player in the other group. Each population has a known fraction of *honest* types—individuals who always cooperate unless paired with a player who has been observed to defect against a cooperating partner in the past. Because such defections (i.e., breakdowns of trust) are publicly observed, any defector risks carrying a stigma into future interactions. We show that when the benefits from defection are sufficiently large, there exists an equilibrium in which an increase in the fraction of honest types can *reduce* the likelihood of cooperation. Moreover, we demonstrate that introducing imperfect public memory—allowing past misdeeds to be probabilistically "cleared"—does not enhance cooperation.

Keywords: Overlapping generations, Prisoner's Dilemma, Reputation, Stigma.
JEL Codes: C72, C73, D82, D83

1 Introduction

It is well-established that mutual cooperation can be sustained if the same players interact repeatedly. Under these circumstances, the immediate gains from defection can be outweighed by the lost benefits from future cooperation. In the classic repeated Prisoner's Dilemma, this idea is formalized by grim-trigger strategies: once one player defects, the other retaliates with defection forever.

Real-world interactions between parties often occur under short horizons or sporadic contact, yet credible information about past behavior can circulate, encouraging trust or deterring exploitation. Online marketplace environments (e.g., eBay, Airbnb) provide a prominent illustration (Resnick et al. (2006), Bolton et al. (2004)). Buyers and sellers typically have only a single or short-lived exchange, but platforms maintain public ratings. Some participants are "honest" by nature, providing consistent quality and refraining from cheating even when the monetary incentives to misbehave are

*We would like to express our gratitude to Ronaldo Carpio for an invaluable feedback throughout this research. All remaining errors are ours.

†School of Economics, Peking University, 5 Yiheyuan Road, Haidian District, Beijing, P.R.China. Email address: dshli@edu.hse.ru

‡Toulouse School of Economics, 1, Esplanade de l'Université 31080 Toulouse, France. E-mail address: georgy.lukyanov@tse-fr.eu

large. Others are more opportunistic, cooperatively delivering good service only if future interaction value—or the reputational penalty for bad behavior—is high enough. In our framework, a defecting seller (e.g., one shipping substandard goods) becomes "stigmatized" through negative feedback, reducing future demand. However, if feedback or records are imperfect (akin to a "forgiveness" feature), the deterrent power of ratings dissipates, which can undermine cooperation.

Community-based finance or *rotating savings and credit associations (ROSCAs)* represent another important setting for these ideas. Members deposit money into a communal fund, each taking turns to withdraw a lump sum (Besley et al. (1993)). The group often relies on social norms and known reputations to ensure compliance. A fraction of individuals may be intrinsically trustworthy (honest types), consistently contributing their share. Others are more strategic; they contribute faithfully so long as the threat of being denied future rounds—or incurring social stigma—is real. If the benefit of cheating (e.g., reneging on contributions) is large and the fraction of honest types is moderate, the fear of "blacklisting" can sustain cooperation. But paradoxically, in some cases a higher proportion of honest players can trigger a shift in collective beliefs that leads to less overall cooperation—particularly if the short-term incentives to default are substantial.

Business-to-business (B2B) contracting across generations offers a further real-world domain where overlapping interactions and reputational concerns matter, see e.g. Baker et al. (2002); Fafchamps (2000). For instance, family-run suppliers and buyers in industrial districts pass relationships from one generation to the next. A new (young) manager may have only one overlapping period with an older manager on the other side, so the two do not expect to trade repeatedly with each other; however, reputational information about "unfair" actions (such as refusing to honor a final payment) persists in the local community. If a manager is known to have exploited a trusting partner, prospective future collaborators might refuse to deal with that firm. Nevertheless, if the temptation from one-off betrayal is large and record-keeping less than perfect, such reputational mechanisms can break down. The theoretical framework here clarifies why "doing the right thing" in one match hinges on how future partners—both honest and strategic—will respond to publicly observed misdeeds.

We focus on two key elements. First, some individuals are intrinsically *honest*: they simply "do the right thing" rather than acting out of narrowly defined self-interest. In philosophical terms, they adhere to *deontological ethics*, which are *non-consequentialist*. Second, there exists a community-wide mechanism for information transmission. Whenever a player defects against a cooperative partner, that act becomes publicly observable; hence, a defector may face a damaging stigma that jeopardizes future cooperation.

To capture these ideas, we develop an overlapping-generations (OLG) game between two groups, building on the intergroup framework initially proposed by Acemoglu and Wolitzky (2014) for studying conflict spirals. Each population is a continuum of players; every individual lives for two consecutive periods (young and old). At each date, a young player from one group is randomly matched to play a Prisoner's Dilemma with an old player from the other group. Because the match occurs only once, long-term reputational mechanisms do not apply to the *same* pair; instead, the

threat of future stigma motivates cooperation in subsequent matches with different partners.

A known fraction of players in each group are *honest*, meaning they will cooperate unless they detect (through public records) that their opponent defected in the past against a cooperative partner. Observing such a defection reveals the opponent to be opportunistic, so an honest type will never trust them. To avoid mixed-strategy complications, we assume that each strategic player has a privately observed cost of being defected against. In equilibrium, a player cooperates if this cost is below a threshold and defects otherwise. We restrict attention to symmetric equilibria, where both groups have the same fraction of honest types, and each strategic player employs the same threshold rule.

Our first main result (Proposition 3.3) characterizes equilibrium. When the net benefit from defection today is smaller than the net loss of future cooperation breakdown, equilibrium is unique. By contrast, if the net benefit is larger, there is a range of beliefs over which multiple equilibria emerge.

Next, we examine how the equilibrium threshold—and hence the likelihood of cooperation—responds to an increase in the fraction of honest types (Proposition 4.1). If the defection benefit is moderate (leading to a unique equilibrium), a higher fraction of honest types unambiguously raises the probability of cooperation. However, once the defection benefit is large enough to admit multiple equilibria, an increase in honest types can *reduce* overall cooperation.

One might conjecture that introducing partial forgiveness—where reputational "records" are cleared with some probability—could foster more cooperation. Surprisingly, we show this is not the case (Proposition 5.1): imperfect public memory does not expand the scope for cooperative behavior.

Although our model is highly stylized, it can be extended to analyze a variety of settings in applied game theory. For example, it sheds light on how intergroup cooperation or recurring cycles of conflict can evolve under rumors or partial information. The interplay between honest types, stigma, and the threat of lost cooperation thus offers a robust framework for understanding the persistence or dissolution of trust in environments marked by sporadic interaction.

1.1 Literature Review

A central theme in the literature on repeated interaction is how cooperation can arise even in environments where defection is a dominant one-shot strategy. One classical reference is Kreps et al. (1982), who illustrate how incomplete information about opponents' "types" can sustain cooperation in a repeated Prisoner's Dilemma. The present paper brings a twist to this logic by introducing short-term reputational incentives within an overlapping-generations framework similar to Acemoglu and Wolitzky (2014). In their model, conflict spirals eventually dissolve as players come to believe that ongoing discord may have stemmed from an initial misunderstanding. This higher probability of "mistaken" conflict over time incentivizes experimentation with cooperation.

In contrast, we study a setting where some conflicts originate from genuine defection, reinforced by the classic Prisoner's Dilemma payoffs. A *strategic* player's only motive to cooperate when young

is to avoid carrying a negative stigma into old age. This hinges on the presence of *honest* players—i.e., non-strategic individuals who follow deontological principles of "doing the right thing." Such moral behavior, akin to "blushers" in the sense of Frank (1987), can serve as a commitment device. Defection against an honest cooperator becomes publicly known, tarnishing the defector's reputation. Hence, young players may refrain from defecting in order to preserve future opportunities for exploitation. By including honest types, we endogenize a moral cost of defection, echoing Lukyanov and Li (2025), who examine how belief heterogeneity influences cooperation in a static framework. Our approach also resonates with the evolutionary perspective of Alger and Weibull (2013), where "homo moralis" preferences—combining selfish and moral motives—emerge as evolutionarily stable under certain assortative matching conditions.

In the spirit of repeated games with overlapping generations, our work connects to Fudenberg et al. (1998), who restrict the set of equilibrium payoffs when some players are inherently short-run while others are long-run. Similarly, Tadelis (2002) shows that trading reputations can extend incentives across life stages, rendering them effectively "ageless." Building on this, Bar-Isaac (2007) demonstrates how senior agents with established reputations deliberately hire juniors of uncertain type, leveraging the juniors' future career concerns to maintain overall effort. More recently, Acemoglu and Wolitzky (2024) develop a model where soft employment relations are sustained by community-based cooperation and mutual trust, yet can be undermined by improved monitoring, automation, or privatized services. All these analyses highlight how reputation, whether individual or firm-level, can bridge short-lived relationships and foster cooperation.

A further dimension concerns the role of cultural and psychological factors. Kets and Sandroni (2020) propose a "culture as shared cognition" framework, arguing that cultural diversity can heighten strategic uncertainty, potentially trapping societies in low-trust equilibria. This aligns with our result that an increased fraction of "honest" types does not universally guarantee more cooperation; heightened diversity in beliefs or types can, under some circumstances, make cooperation less likely. From another angle, Charness and Dufwenberg (2006) demonstrate experimentally that guilt aversion—people's desire to meet others' expectations—sustains cooperation, especially when promises shift beliefs about future actions. Similarly, López-Pérez (2012) highlights how norms of honesty and fairness can foster cooperation in social dilemma games, provided players fear the utility loss from breaking these norms. Miettinen (2013) formalizes the role of guilt in upholding informal agreements, noting that it is easier to sustain cooperation in games with strategic complements than strategic substitutes.

Empirical and experimental studies reinforce how misreporting and reputation concerns shape cooperation. Gneezy et al. (2018) show that while some participants lie maximally for monetary gain, others engage in only partial lies, reflecting heterogeneity in lying costs. This resonates with Bahel et al. (2022), who show that populations with a higher valuation for truth-telling do not necessarily achieve higher cooperation: if individuals are less willing to misrepresent private information, some coordination-enhancing "cheap talk" may fail. In large-scale settings, van den Assem et al. (2012) analyze the British game show *Golden Balls*, observing unexpectedly high cooperation for relatively

small but contextually "large" monetary sums. Their data point to reciprocal preferences, though they find little evidence that players condition their moves on opponents' likely cooperation.

In sum, our work contributes to these literatures by combining short-term reputational incentives in an overlapping-generations setup with a fraction of inherently moral (honest) individuals. We show how reputational concerns can sustain cooperation, but also illuminate a surprising result: increases in the fraction of honest players can, under certain conditions, *reduce* the likelihood of cooperation. This underscores the nuanced ways in which reputation, cultural diversity, and moral commitment can jointly shape cooperation, even in the face of dominant defection incentives.

The rest of the paper is organized as follows. Section 2 describes the model. Section 3 presents the equilibrium characterization in threshold strategies. Section 4 provides our central analysis regarding the likelihood of cooperation. Section 5 studies the role of noise in the record-keeping of past defections. Section 6 proposes several extensions. Section 7 concludes.

2 Model Description

We study an infinite-horizon game involving two large groups of players, labeled A and B. Time is discrete, indexed by $t = 1, 2, \ldots$. Each player lives for two consecutive periods: as a young individual in one group and subsequently as an old individual in the other group. At each date t, a *young* player from group A is randomly matched to interact with an *old* player from group B, and vice versa in alternating fashion. Thus, each player has exactly two interactions over his lifetime: one when young (matched with an old player from the other group), and one when old (matched with a young player from the other group).

The identity of one's partner is not publicly revealed beyond the two involved. Importantly, we assume the entire community (across both groups) may observe whether a defection occurred against a cooperating partner, but not *which* players were involved in a given interaction. Consequently, a "stigma" for having defected against a cooperator is publicly recorded.

Although the community only observes that some defection happened, we assume there is a mechanism—such as a commonly accessible record-keeping system—that flags the actual defector with a public label (a 'stigma'). Thus, the next person to interact with that defector can verify that they are indeed 'stigmatized.' This label is effectively observed by future partners but does not disclose the person's name or other personal identifiers to everyone else. In this way, a defecting player becomes publicly known as untrustworthy, while still preserving the idea that the broader community may not know all the details of the past interaction.

Each match consists of a one-shot Prisoner's Dilemma. Denoting the two actions as C (Cooperate) and D (Defect), we represent the payoffs in normal form as follows:

		Agent B: D	Agent B: C
Agent A: D		0, 0	b, $-\ell_j$
Agent A: C		$-\ell_i$, b	1, 1

Here, $b > 1$ is the benefit from defecting against a cooperator, and ℓ_i is the cost borne by player i when he cooperates but his opponent defects. Note that if both players cooperate, they each receive a payoff of 1; if both defect, they each get 0. In mixed interactions, the defector earns b while the cooperator suffers a cost ℓ_i. Throughout, we make the following natural assumption:

Assumption 2.1. *The benefit from defection satisfies* $b > 1$.

If $b \leq 1$, the game becomes more akin to a coordination problem with multiple equilibria rather than a strict Prisoner's Dilemma.

Players vary in their disutility from being exploited when cooperating. Specifically, each player i has a cost parameter ℓ_i that is *privately observed* and drawn from a common distribution $F(\cdot)$. When player i cooperates but faces a defection, his payoff in that encounter is $-\ell_i$.

For tractability, we use the following specification:

Assumption 2.2. *The cost* ℓ_i *is drawn from a uniform distribution on* $[0, 1]$*, so that* $F(\ell) = \ell$ *for* $\ell \in [0, 1]$.

In each group $i \in \{A, B\}$, a fraction π_i of players are *honest* types who always *attempt* to cooperate unless they know their opponent has previously defected against a cooperator. Any honest old player who experiences defection from his young opponent publicly signals that this opponent has defected against a cooperator. Consequently, such a young defector receives a *stigma*.

Honest Types. By assumption, an honest player's strategy is exogenously determined rather than derived from self-interested optimization: he cooperates unless his opponent is known to be stigmatized, in which case he defects. We do not model the honest player's payoffs explicitly; the Prisoner's Dilemma matrix applies primarily to strategic (non-honest) individuals.

Stigma and Defection Incentives. If a player becomes stigmatized for defecting against a cooperator, future opponents can identify him as "untrustworthy." Hence:

- An *old* strategic player will always defect in his second match, because it is his last interaction and D is a dominant action in the one-shot Prisoner's Dilemma.
- A *young* strategic player facing an opponent with a stigma knows this opponent is a strategic defector and thus expects D. Hence the young player's best response is also to play D.

The only nontrivial decision arises when a *young* strategic player faces an *old* player with a "clear" (non-stigmatized) record. In this case, defecting now yields an immediate payoff of b (provided that

the opponent cooperates), but leads to stigmatization and thus no further cooperative benefits in his second match.

Finally, we assume:

Assumption 2.3. *Players do not discount future payoffs; effectively, the discount factor is 1.*

Under no discounting, a player's total expected payoff is simply the sum of his payoffs as a young player and as an old player. This makes it straightforward to compare the immediate benefit of defection with the potential loss from being stigmatized in the next stage.

While we assume $\delta = 1$ for simplicity, real-world interactions often involve discounting, i.e., $\delta < 1$. In such cases, the essence of stigma-based cooperation remains largely intact as long as δ is not too low: a defection still imposes a reputational penalty that can deter opportunistic behavior. However, because future payoffs are now worth less in present-value terms, the immediate benefit from defecting becomes relatively more attractive. As a result, if δ decreases, the fraction of honest players (or the strength of other enforcement mechanisms) must be higher to sustain cooperation. In the extreme case where δ approaches zero, future punishments or rewards matter almost none, causing cooperation to unravel.

Given these elements, our primary focus is on characterizing equilibrium play among *strategic* (non-honest) players, taking honest types' behaviors as given. We will show that strategic players adopt threshold strategies based on their private cost ℓ_i. Specifically, a strategic player is more likely to cooperate when his cost of being exploited is relatively high, so that the threat of losing future cooperative gains outweighs the short-run benefit of defection.

We proceed next to derive equilibrium conditions and discuss how the fraction of honest types and the magnitude of b together shape the emergence (and multiplicity) of equilibria.

3 Equilibrium Characterization

We now characterize the Perfect Bayesian Equilibrium (PBE) of the game under the assumption that each *strategic* player adopts a threshold rule. Formally:

Definition 3.1. A *threshold strategy* for a strategic young player matched with an old partner of clear reputation (no stigma) takes the form:

$$\sigma_i(\ell_i) = \begin{cases} C & \text{if } \ell_i \leq \ell^*, \\ D & \text{if } \ell_i > \ell^*. \end{cases} \tag{3.1}$$

Under this rule, when young, a player cooperates if and only if his privately observed cost of being exploited, ℓ_i, does not exceed some cutoff ℓ^*. Because ℓ_i is uniformly distributed on $[0, 1]$, the probability a strategic young player cooperates in equilibrium is simply ℓ^*. By convention, $\ell^* = 0$ represents a strategy of defecting for all cost realizations, and $\ell^* = 1$ corresponds to cooperating in every possible scenario.

Throughout this section, we treat π, the fraction of honest types in each group, as fixed.[1]

To understand the equilibrium cutoff, consider the expected payoffs of a strategic young player whose next-period strategic partner also uses threshold ℓ^*. Denote by $\mathbb{E}[u_i(D, \ell_i; \ell^*)]$ the expected payoff to a young player who *defects* today (against an old partner with clear reputation). We have:

$$\mathbb{E}[u_i(D, \ell_i; \ell^*)] = [\pi(2 - \pi) + (1 - \pi)^2 \ell^*] b. \tag{3.2}$$

Intuitively, if the opponent is honest (probability π), defecting yields an immediate gain of b. However, this incurs a stigma for the next round, making future payoff zero. If the old opponent is strategic (probability $(1 - \pi)$), then the current payoff is zero when both defect; crucially, no stigma arises because defection happened in a mutual-defect profile. Thus, the player retains a "clear reputation" for the future, allowing him to exploit the next (young) opponent with probability $\pi + (1 - \pi)\ell^*$. In those cases, he again gains b.

By contrast, cooperating yields the expected payoff

$$\mathbb{E}[u_i(C, \ell_i; \ell^*)] = -(1 - \pi)\, \ell_i \; + \; \pi \; + \; [\pi + (1 - \pi)\, \ell^*]\, b. \tag{3.3}$$

When cooperating today, the young player gets $-\ell_i$ if the old partner is strategic and defects (probability $1 - \pi$); otherwise, he earns 1 from meeting an honest opponent. Moreover, since he does not acquire a stigma, the next-period payoff is $[\pi + (1 - \pi)\ell^*]\, b$.

Notice that (3.2) does not depend on the cost ℓ_i, whereas (3.3) *does* depend on ℓ_i. Consequently, for each cost realization ℓ_i, the player compares the two expressions to determine whether to cooperate or defect.

Interestingly, there can be values of π for which cooperation becomes a *dominant* action (i.e., strictly preferred for *all* $\ell_i \in [0, 1]$). The following lemma formalizes this idea:

Lemma 3.2. *Suppose $1 < b < 2$. Then there exists a threshold $\overline{\pi}(b)$ such that for every $\pi > \overline{\pi}(b)$, cooperating is a dominant strategy for all $\ell_i \in [0, 1]$.*

Proof. Consider a worst-case scenario where the future strategic opponent is certain to defect, i.e. $\ell^* = 0$. Then from (3.2) and (3.3), defection and cooperation reduce to:

$$\mathbb{E}[u_i(D, \ell_i; 0)] = \pi(2 - \pi)\, b, \quad \mathbb{E}[u_i(C, \ell_i; 0)] = -(1 - \pi)\, \ell_i + \pi(1 + b).$$

If a young player's cost is as high as $\ell_i = 1$ (the "worst-case" cost of cooperating against a defector), cooperation still yields a higher expected payoff whenever

$$\pi(1 + b) - (1 - \pi) \; > \; \pi(2 - \pi)\, b.$$

[1] In the next section, we explore how equilibrium cooperation depends on changes in π.

Rearranging gives

$$2\pi + \pi b - 1 \;>\; 2\pi b - \pi^2 b \quad\Longleftrightarrow\quad \pi \;>\; 2\Big(1 - \tfrac{1}{b}\Big) \;\equiv\; \overline{\pi}(b).$$

Hence for $\pi > \overline{\pi}(b)$, even in the worst-case scenario, cooperating strictly dominates defecting at every ℓ_i. □

When π is large enough (and b is not too large), even high-cost players prefer to cooperate rather than defect. However, if b is large (e.g. $b \geq 2$), the advantage from short-run defection can dominate, so cooperation need not be dominant. We capture this formal distinction in the next proposition.

Proposition 3.3. *1. For $b \in (1, 2)$, there exists a* unique *symmetric equilibrium with threshold*

$$\ell^*(\pi) = \begin{cases} 0 & \text{if } \pi < 1 - \frac{1}{b}, \\ \dfrac{\pi\big[\,1 - b(1-\pi)\big]}{(1-\pi)\,\big[\,1 - b\pi\,\big]} & \text{if } 1 - \frac{1}{b} \;\leq\; \pi \;\leq\; \frac{1}{2}, \\ 1 & \text{if } \pi > \frac{1}{2}. \end{cases}$$

2. For $b \geq 2$ and $\pi \in \big[\frac{1}{2}, 1 - \frac{1}{b}\big]$, there are three *equilibria: $\ell^*(\pi) = 0$, $\ell^*(\pi) = 1$, and $\ell^*(\pi) = \dfrac{\pi\big[1 - b(1-\pi)\big]}{(1-\pi)\big[1 - b\pi\big]}$.*

The first case captures the intuition that if the net gains to defection (b) are moderate and the fraction of honest types (π) is sufficiently large, the unique equilibrium can feature all players cooperating (i.e. $\ell^* = 1$), or a positive interior cutoff. If π is too low, or b is too large, the "temptation to defect" can produce multiple equilibria: everyone defects ($\ell^* = 0$), everyone cooperates ($\ell^* = 1$), or a mixed threshold equilibrium emerges where cooperation occurs only for sufficiently high cost realizations ℓ_i.

Overall, these results highlight the pivotal role of π (the fraction of honest types) and b (the payoffs to defection) in shaping equilibrium behavior. In the sections that follow, we analyze how these equilibria shift when π changes and discuss comparative statics that illustrate non-monotonic effects of increasing the honest fraction.

4 Probability of Cooperation

In this section, we examine how the equilibrium threshold $\ell^*(\pi)$ translates into an overall probability of cooperation in the population and how this probability varies with π, the fraction of honest types. Recall that in any symmetric equilibrium, a *strategic* player cooperates if and only if $\ell_i \leq \ell^*(\pi)$. By assumption, ℓ_i is uniformly distributed on $[0, 1]$, so the probability that a randomly chosen *strategic* player cooperates is precisely $\ell^*(\pi)$. Consequently, the *population-wide* probability of cooperation at

any date can be expressed as

$$\Pr(\text{cooperation}) \;=\; \pi + (1-\pi)\,\ell^*(\pi).$$

This follows because with probability π we encounter an honest type, who (by definition) cooperates unless facing a stigmatized defector. In a steady state, the share of stigmatized players is endogenous, but whenever a newly matched old opponent has a clear reputation, honest types play C. The remaining $1-\pi$ fraction of players are strategic; among them, only those whose cost realization ℓ_i falls below the cutoff $\ell^*(\pi)$ choose C.

Our next result, **Proposition 4.1**, confirms the intuition hinted by our earlier discussion. When b is not too large (so that there is a unique threshold equilibrium), the population-wide probability of cooperation is strictly increasing in π. However, once b becomes large enough to admit multiple equilibria, an increase in π may reduce the equilibrium level of cooperation.

Proposition 4.1. *Suppose $b > 1$ and consider the unique or interior-threshold equilibrium described in Proposition 3.3.*

1. *If $1 < b < 2$, $\pi + (1-\pi)\ell^*(\pi))$ is* strictly increasing *in π. In particular, higher π unambiguously raises the probability of cooperation.*

2. *If $b \geq 2$, there exist regions of π in which multiple equilibria arise. In those ranges, an increase in π can lead to an equilibrium selection that* lowers *the overall probability of cooperation.*

Proof. We divide the proof into two parts. First, we consider the case $1 < b < 2$, in which there is a *unique* threshold equilibrium for each π. We show that $\ell^*(\pi)$ is strictly increasing in π. Second, we address the case $b \geq 2$, where multiple equilibria can arise, and illustrate how increasing π can lead to lower levels of cooperation.

Recall from Proposition 3.3 that if $b \in (1,2)$ and π is large enough, there is a *unique* symmetric threshold $\ell^*(\pi)$ in $(0,1)$, or possibly one of the corner values $\{0,1\}$ if π is sufficiently small or large. We focus on the interior solution case to analyze the derivative $\frac{d\ell^*(\pi)}{d\pi}$.[2]

An interior threshold $\ell^* \in (0,1)$ is determined by the indifference condition:

$$\mathbb{E}\big[u_i(C,\ell^*);\ell^*\big] \;=\; \mathbb{E}\big[u_i(D,\ell^*);\ell^*\big] \quad \text{when} \quad \ell_i = \ell^*.$$

Recalling from (3.2) and (3.3):

$$\mathbb{E}[u_i(D,\ell_i;\ell^*)] = \big[\pi(2-\pi) + (1-\pi)^2\ell^*\big]\,b,$$

$$\mathbb{E}[u_i(C,\ell_i;\ell^*)] = -\,(1-\pi)\,\ell_i \;+\; \pi \;+\; \big[\pi + (1-\pi)\,\ell^*\big]\,b.$$

Setting $\ell_i = \ell^*$ in the cooperation payoff and equating to the defection payoff yields

$$\big[\pi(2-\pi) + (1-\pi)^2\ell^*\big]\,b = -\,(1-\pi)\,\ell^* \;+\; \pi \;+\; \big[\pi + (1-\pi)\,\ell^*\big]\,b.$$

[2] Corner solutions are trivially monotonic in π by inspection of the conditions under which they arise.

Solving this linear equation in ℓ^* gives (for an interior solution)

$$\ell^*(\pi) = \frac{\pi\big[\,1 - b(1-\pi)\big]}{(1-\pi)\big[\,1 - b\pi\,\big]} = \frac{\pi\big[\,1 - b + b\pi\,\big]}{(1-\pi)\big[\,1 - b\pi\,\big]} \quad \text{provided that } 0 < \ell^*(\pi) < 1. \tag{4.1}$$

(If this expression yields a value outside $[0,1]$, the equilibrium threshold is instead a corner solution, $\ell^* = 0$ or $\ell^* = 1$.)

Now let us establish the monotonicity in π. Focus on the domain of π values for which $\ell^*(\pi)$ in (4.1) lies strictly in $(0,1)$. We analyze

$$\frac{d}{d\pi}\,\ell^*(\pi),$$

using the explicit formula. Let us rewrite it more compactly as

$$\ell^*(\pi) = \frac{\pi\big[A + B\pi\big]}{(1-\pi)\big[C - D\pi\big]},$$

where $A = 1 - b$, $B = b$, $C = 1$, and $D = b$. That is,

$$\ell^*(\pi) = \frac{\pi\,(A + B\pi)}{(1-\pi)\,(C - D\pi)}.$$

To differentiate, we apply the quotient rule. Let

$$N(\pi) = \pi\,(A + B\pi), \quad D(\pi) = (1-\pi)\,(C - D\pi).$$

Then

$$\ell^*(\pi) = \frac{N(\pi)}{D(\pi)}.$$

Hence

$$\frac{d}{d\pi}\,\ell^*(\pi) = \frac{N'(\pi)\,D(\pi) - N(\pi)\,D'(\pi)}{\big[D(\pi)\big]^2}.$$

First compute $N'(\pi)$:

$$N(\pi) = \pi\,(A + B\pi) = A\pi + B\pi^2, \quad N'(\pi) = A + 2B\pi.$$

Next, compute $D'(\pi)$. Since

$$D(\pi) = (1-\pi)\,(C - D\pi) = C(1-\pi) - D\pi(1-\pi),$$

we get

$$D'(\pi) = -C \;-\; \Big[D(1-\pi) \;-\; D\pi\Big] = -C \;-\; D(1-2\pi).$$

Substituting $C = 1$ and $D = b$,

$$D'(\pi) = -1 - b(1-2\pi) = -1 - b + 2b\pi.$$

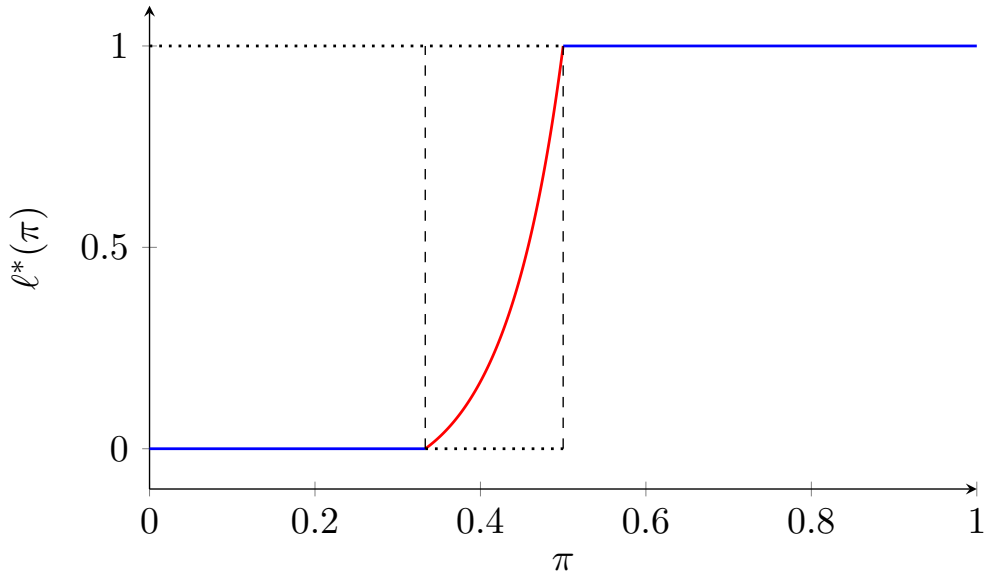

Figure 1: **Equilibrium threshold $\ell^*(\pi)$ as a function of π.** For illustration, we set $b = 1.5$. When $\pi < 1/3$, the threshold is zero (players always defect). For $\pi > 1/2$, the threshold is one (players always cooperate). In the intermediate range $\pi \in [1/3,\, 1/2]$, an interior threshold emerges from the indifference condition. As π increases, the equilibrium threshold rises, reflecting a higher likelihood of meeting an honest cooperator and thus a stronger reputational deterrent against defection.

We now compare $N'(\pi)\, D(\pi)$ with $N(\pi)\, D'(\pi)$. Observe

$$N'(\pi)\, D(\pi) = \big[A + 2B\pi\big]\big[(1-\pi)\,(C - D\pi)\big],$$

$$N(\pi)\, D'(\pi) = \big[A\pi + B\pi^2\big]\big[-1 - b + 2b\pi\big].$$

For $b \in (1, 2)$, a straightforward (though somewhat tedious) expansion and simplification shows that

$$N'(\pi)\, D(\pi) - N(\pi)\, D'(\pi) \;>\; 0 \quad \text{for all } \pi \text{ in the relevant range}$$

Thus

$$\frac{d}{d\pi}\,\ell^*(\pi) = \frac{N'(\pi)\, D(\pi) - N(\pi)\, D'(\pi)}{\big[D(\pi)\big]^2} \;>\; 0,$$

since the denominator $\big[D(\pi)\big]^2$ is positive whenever $\ell^*(\pi)$ is an admissible equilibrium threshold. Consequently, $\ell^*(\pi)$ is strictly increasing in π. Because the overall probability of cooperation is $\pi + (1-\pi)\,\ell^*(\pi)$, its derivative with respect to π is also strictly positive (in the unique interior equilibrium regime). This completes the argument that cooperation is monotonically increasing in π for $1 < b < 2$.

For $b \geq 2$, Proposition 3.3 shows there are ranges of π for which multiple equilibria exist: one corner solution with $\ell^* = 0$, another corner with $\ell^* = 1$, and (possibly) an interior threshold. In such scenarios, *equilibrium selection* matters for realized cooperation.

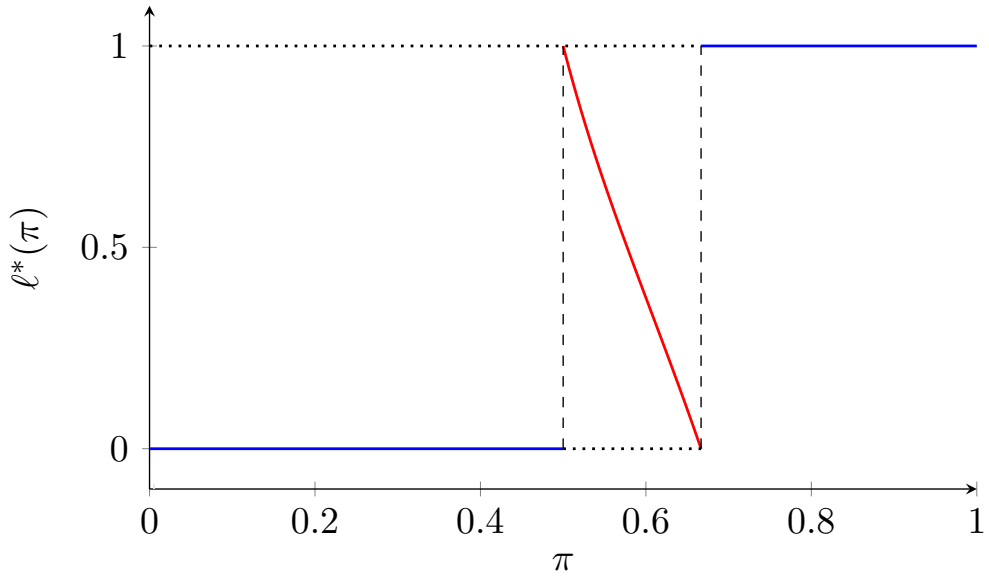

Figure 2: **Threshold $\ell^*(\pi)$ for $b = 3$.** When $\pi < 0.5$, the unique equilibrium is $\ell^*(\pi) = 0$ (everyone defects). When $\pi > 2/3$, $\ell^*(\pi) = 1$ (everyone cooperates). For the intermediate range $\pi \in [0.5, 2/3]$ (shaded), there are *three* equilibria: the corner solutions $\ell^* = 0$ and $\ell^* = 1$ *and* the interior threshold plotted in red. Hence, an increase in π within this interval can lead to a switch to a corner equilibrium with lower cooperation.

To see how increasing π can reduce cooperation, suppose we start at some $\pi = \pi_0$ with the "high-threshold" equilibrium (e.g. $\ell^* = 1$). If π rises slightly to $\pi_1 > \pi_0$, we might also have a "low-threshold" equilibrium (e.g. $\ell^* = 0$) that becomes focal because players coordinate on it. In that equilibrium, strategic types defect whenever they face a clear-reputation partner, and overall cooperation falls dramatically. Formally, the presence of a self-fulfilling expectation (where strategic players anticipate that everyone defects, so defection is indeed optimal) can supplant the cooperative equilibrium. Hence, *despite* an increase in π, the system can transition to an outcome with a lower probability of cooperation.

In other words, for large b, a small rise in π is insufficient to guarantee an unambiguous selection of the cooperative threshold; if agents collectively "expect" defection, they end up in the low-cooperation equilibrium. Accordingly, at the population level, the fraction of cooperative outcomes can decrease even though more individuals are honest.

This completes the proof. □

Figures 1 and 2 help illustrate the logic behind Proposition 4.1. In Figure 1, we depict how an individual player's choice between cooperating and defecting depends on the private cost ℓ_i. The crossing point ℓ^* arises where the payoff from cooperation equals the payoff from defection. Increasing the fraction of honest types π tends to shift this payoff comparison toward cooperation: an honest partner provides more reliable gains from mutual cooperation, while a defection against an honest cooperator entails a stigma that undermines future exploitation. Thus, when the temptation to defect (b) is moderate, the threshold ℓ^* grows monotonically in π.

Figure 2 then illustrates what happens when b is large enough to admit multiple equilibria. Over certain intervals of π, players can coordinate on a fully defective corner, a fully cooperative corner, or an intermediate threshold. An increase in π may push society away from a partially (or fully) cooperative equilibrium and toward an outcome in which strategic players universally defect. Hence, the overall rate of cooperation can decrease even as the population becomes more "honest." This graphical representation clarifies why higher π does not guarantee more cooperation when reputational incentives are overshadowed by the strong short-term gains from defection.

This result highlights that greater prevalence of honest types need not be monotonically beneficial for cooperation if the temptation to defect is sufficiently strong. In practice, this may correspond to settings where the short-run gains from opportunism dwarf long-run reputational concerns.

5 Noisy Recall of Defection History

The analysis so far assumes that stigma is *perfectly* persistent and known to all. Yet in many real situations, records of past defections can be forgotten, misreported, or only partially visible to the public. One might conjecture that such *noisy recall* or partial "forgiveness" mechanisms could restore or even enhance cooperation by giving stigmatized players a "fresh start." We now introduce a simple model of imperfect record-keeping to see whether this can occur.

Suppose that each period, with probability $\alpha \in (0, 1)$, any existing stigma is erased from public memory (and does not carry over into the next period). Concretely:

- If a young player defected against an honest cooperator at date t, he becomes stigmatized.
- With probability $1 - \alpha$, this stigma persists at $t + 1$; with probability α, it vanishes and the player is once again "clear."

Otherwise, the game structure remains the same. All honest players refuse to cooperate with a stigmatized opponent, while strategic types can observe (at the start of their second period) whether they themselves remain stigmatized.

To see whether noisy recall can increase cooperation, note that a strategic defector gains at least one period of high payoff b. In subsequent interactions, stigma may vanish with probability α, limiting the future punishment. On the other hand, *honest* players also lose the ability to impose a long-term penalty if it disappears too soon. Hence, one might expect partial forgiveness to weaken deterrence.

Proposition 5.1. *Introducing a positive probability $\alpha > 0$ that stigma is erased* never *increases the probability of cooperation in any of the equilibria described in Proposition 3.3.*

Proof. We extend the baseline model to allow for a "forgiveness" probability $\alpha > 0$ by which a stigmatized player regains a clear record at the start of the next period. We show that in any symmetric threshold equilibrium of this modified game, the cutoff ℓ^* for cooperating is weakly lower

than (or equal to) the cutoff in the no-forgiveness benchmark ($\alpha = 0$). Consequently, population-wide cooperation does not increase under any equilibrium selection.

Focus again on a *strategic* young player who is considering whether to defect against an old opponent with a clear reputation. His immediate and future payoffs under *defection* and *cooperation* now differ slightly from the original model, because a stigma acquired today might vanish with probability α in the next period.

When a young player defects against an *honest* old opponent (probability π), he obtains:

$$\underbrace{b}_{\text{current gain}} + \underbrace{(1-\alpha)\cdot 0}_{\text{remains stigmatized w.p. } (1-\alpha)} + \underbrace{\alpha \cdot \Pi(\ell^*) \cdot b}_{\text{forgiven w.p. } \alpha},$$

where

$$\Pi(\ell^*) \;=\; \pi + (1-\pi)\,\ell^*$$

is the probability of meeting a cooperating young partner in the *next* period (if one has a *clear* record). To clarify each component:

- *Immediate gain b:* he exploits the honest cooperator today.
- *Future payoff if stigma persists:* if he remains stigmatized (probability $1-\alpha$), his next opponent sees the stigma and therefore defects against him, yielding payoff 0.
- *Future payoff if forgiven:* with probability α, the stigma disappears. Then in the next period, he faces a young partner who cooperates with probability $\Pi(\ell^*)$, yielding an additional payoff of b to the defector (the same exploitation logic as the baseline).

If the old opponent is *strategic* (probability $1-\pi$), then mutual defection gives the young player a current payoff 0 but also *no stigma*. Thus in that subcase, next period is again worth $\Pi(\ell^*)\,b$. Combining, the young defector's expected payoff is

$$\underbrace{\pi}_{\substack{\text{meets honest}\\ \text{(current } b\text{)}}} \Big[b + \alpha\,\Pi(\ell^*)\,b\Big] \;+\; \underbrace{(1-\pi)}_{\substack{\text{meets strategic}\\ \text{(current 0)}}} \Big[\Pi(\ell^*)\,b\Big].$$

Hence, letting U_D denote this defection payoff,

$$U_D(\ell^*;\alpha) \;=\; \pi\Big[b + \alpha\,\Pi(\ell^*)\,b\Big] \;+\; (1-\pi)\,\Pi(\ell^*)\,b.$$

Factor out b and rearrange slightly,

$$U_D(\ell^*;\alpha) = b\Big[\pi \;+\; (1-\pi)\,\Pi(\ell^*) \;+\; \alpha\,\pi\,\Pi(\ell^*)\Big]. \tag{5.1}$$

If the young player cooperates, the analysis of immediate and future consequences is analogous to the baseline, but with an additional possibility of being *forgiven*—which does not matter here, because cooperating does not generate stigma in the first place. Specifically:

- With probability π, the old opponent is honest and thus cooperates, giving payoff 1.
- With probability $(1-\pi)$, the opponent is strategic and defects, yielding payoff $-\ell_i$.
- Next period, the player definitely has a clear reputation (since he never defected against a cooperator). Hence his future payoff is $\Pi(\ell^*)\, b$.

Thus, denoting by $U_C(\ell_i; \ell^*; \alpha)$ the cooperation payoff, we have

$$U_C(\ell_i; \ell^*; \alpha) = \pi \cdot 1 + (1-\pi)\big[-\ell_i\big] \;+\; \Pi(\ell^*)\, b,$$

which is identical to the baseline formula $\mathbb{E}[u_i(C, \ell_i; \ell^*)]$. Notably, it does *not* depend on α because no stigma arises when cooperating.

As before, a strategic young player with realization ℓ_i cooperates if and only if

$$U_C(\ell_i; \ell^*; \alpha) \;\geq\; U_D(\ell^*; \alpha).$$

When an interior threshold ℓ^* exists, it arises from the indifference condition at $\ell_i = \ell^*$:

$$\pi - (1-\pi)\,\ell^* + \Pi(\ell^*)\, b \;=\; b\Big[\pi \;+\; (1-\pi)\,\Pi(\ell^*) \;+\; \alpha\,\pi\,\Pi(\ell^*)\Big].$$

Rearranging,

$$\big[\pi(2-\pi) + (1-\pi)^2\ell^*\big]b \;+\; \alpha\,\pi\,\Pi(\ell^*)\, b \;=\; -\,(1-\pi)\,\ell^* + \pi + \Pi(\ell^*)\, b,$$

where $\Pi(\ell^*) = \pi + (1-\pi)\ell^*$. Notice that the extra term $\alpha\,\pi\,\Pi(\ell^*)\, b$ on the left side reflects the possibility of forgiveness after defecting against an honest cooperator.

We claim that for each fixed π and b, the unique (or interior) equilibrium threshold $\ell^*(\pi; \alpha)$ under $\alpha > 0$ is *weakly smaller* than $\ell^*(\pi; 0)$ under perfect memory. Concretely, when $\alpha > 0$, the expected cost of defection (i.e. the risk of future punishment) is strictly lower, because there is a positive probability of reverting to a clear record. Hence defection becomes relatively more attractive for any given ℓ_i, shrinking the set of cost realizations ℓ_i for which cooperating is worthwhile.

To see this formally, denote by $\Delta(\ell_i; \ell^*; \alpha)$ the difference:

$$\Delta(\ell_i; \ell^*; \alpha) \;=\; U_C(\ell_i; \ell^*; \alpha) \;-\; U_D(\ell^*; \alpha).$$

An interior threshold ℓ^* solves $\Delta(\ell^*; \ell^*; \alpha) = 0$. Note that $\Delta(\ell_i; \ell^*; \alpha)$ is strictly *decreasing* in ℓ_i, because the cooperation payoff is linear in $-\ell_i$ while the defection payoff is *independent* of ℓ_i. Next, observe that

$$U_D(\ell^*; \alpha) \;\text{ is strictly } \textit{decreasing} \text{ in } \;\alpha\; (\text{for } \alpha > 0).$$

Indeed, from (5.1):

$$\frac{\partial}{\partial \alpha}\, U_D(\ell^*; \alpha) = b\,\pi\,\Pi(\ell^*) \;>\; 0,$$

so the defection payoff is *increasing* in α. Consequently, for the same candidate ℓ^*, $\Delta(\ell_i; \ell^*; \alpha)$ is *strictly lower* at $\alpha > 0$ than at $\alpha = 0$. This implies that if $\ell^*(\pi; 0)$ is the indifferent cutoff under $\alpha = 0$, then for $\alpha > 0$ the same cutoff would yield $\Delta(\ell^*(\pi; 0); \ell^*(\pi; 0); \alpha) < 0$, so a lower threshold $\ell^*(\pi; \alpha) < \ell^*(\pi; 0)$ is needed to restore indifference. In corner solutions, one verifies similarly that allowing forgiveness can only push the equilibrium threshold toward defecting more frequently (i.e. from $\ell^* = 1$ to an interior or zero threshold).

Once we establish $\ell^*(\pi; \alpha) \leq \ell^*(\pi; 0)$ for any $\alpha > 0$, it follows that the fraction of strategic players who cooperate, $(1 - \pi)\, \ell^*(\pi; \alpha)$, is *weakly lower* than under perfect memory. Since honest players always cooperate unless they detect a stigma, the overall probability of cooperation cannot exceed that of the $\alpha = 0$ benchmark. Therefore, introducing forgiveness never increases cooperation in any equilibrium (and often strictly decreases it).

This completes the proof. □

Hence, while one might hope that partial forgiveness could mitigate past mistakes and encourage a return to cooperative norms, our model indicates that it can instead undermine the credibility of stigma-based deterrence. In equilibrium, this reduces cooperation incentives for those on the margin.

6 Extensions and Applications

6.1 Extensions

Our framework is intentionally stylized to highlight the interplay between honest types, reputational incentives, and overlapping generations. Nevertheless, the model can be fruitfully extended in several directions:

Heterogeneous Time Horizons. So far, each player lives exactly two periods (young and old). One could allow for random lifespans or longer lifespans (e.g. three or more interactions). As long as there is a "last period" for everyone eventually, the logic of final-period defection remains. However, the intermediate periods might feature richer dynamics if some individuals anticipate additional encounters.

Multi-Dimensional Heterogeneity. We introduced only one dimension of private information: the cost ℓ_i. Another natural extension would incorporate heterogeneity in b (the temptation payoff), or even differences in how quickly stigma is disseminated. This could capture real-world settings where certain subgroups more effectively share reputational information, leading to partial segmentation of matches.

Endogenous Monitoring Technologies. In practice, the efficacy of stigmatizing a defector depends on the ability to disseminate such information. One could extend the model to allow each

group to invest in monitoring or communication technologies. The cost of establishing reliable "blacklists" or informational networks would then factor into the equilibrium threshold strategies.

Repeated Matchings within the Same Cohort. A more complex variation would let the same cohort of young players meet each other multiple times *before* they become old. Such additional within-cohort repetition might reinforce reputational effects, since a defection in one sub-match could carry repercussions for the remainder of that cohort's interactions.

6.2 Applications

Online Marketplace Platforms. Many online platforms match participants who typically interact only once or sporadically, as in eBay auctions or Airbnb rentals. Even though buyer–seller pairs may not transact repeatedly with each other, a feedback or reputation system publicly flags sellers who fail to deliver promised quality (or guests who behave badly). This is directly analogous to a stigma in the model: once a seller is known to have defected (e.g., shipped substandard goods or canceled at the last minute), future buyers may refuse to trust them—leading to lost business. However, if the short-term profit from cheating is high (for instance, listing an attractive rental property and then canceling to rebook at a higher rate), then even a robust review system may fail to sustain cooperation unless a sufficient fraction of participants are honest by nature. Thus, an influx of honest users can help anchor trust, but in certain parameter ranges where temptation is large, the mere presence of honest participants might still not prevent opportunistic behavior.

Rotating Savings and Credit Associations. ROSCAs feature a group of individuals who commit to making regular contributions, taking turns to withdraw a collective pot of funds. Although each participant interacts with the "group," actual money exchanges can be viewed as pairwise or small-subset interactions that happen once in a cycle. A participant who defaults on her contributions may gain a one-time monetary advantage but will be branded untrustworthy in subsequent cycles, losing eligibility for future rounds or being ostracized in local social networks. This stigma can deter opportunism—much like in the model, where a defector is flagged as risky for future matches. However, if the short-term gain from defaulting is substantial, reputational sanctions alone may not suffice to sustain full cooperation unless the membership includes enough intrinsically honest individuals or there are more formal enforcement mechanisms in place.

Inter-Generational Family Businesses. In many local business communities, such as family-run supplier–buyer networks, relationships and reputations are passed across generations. A newly appointed manager (the "young" player) interacts with an incumbent manager (the "old" player) for a limited time. Even if those two do not expect to do business with each other again, word of opportunistic behavior—like reneging on a final payment—circulates and can tarnish the firm's name. Future managers (the next "young" players) might refuse to contract with a firm that has a reputation for deceit. Hence, a single act of defection can carry long-run consequences across

overlapping generations. Still, if profit margins from cheating one partner are very high, a firm may choose to defect despite the risk to its reputation. This mirrors the trade-off in your model between immediate gains from defection and the longer-term cost of stigma.

Academic or Professional Collaborations. In research collaborations, consultants' engagements, or professional project teams, often each pair of collaborators works together only briefly before moving on to new co-authors or new clients. However, norms of integrity still deter free-riding or misappropriating joint work: once someone is discovered to have "defected" (e.g., taken sole credit for a group project, or billed incorrectly for consulting hours), others in the field are likely to learn of it. Future partners then become wary of collaborating with that person. Thus, even absent repeated interactions with the same partner, the threat of community-wide stigma can enforce cooperative behavior—just as in your framework. Yet if immediate reputational penalties are weak, or the immediate benefits from cheating are very large, such community enforcement may not suffice—particularly when only a moderate proportion of researchers or professionals are intrinsically honest.

7 Conclusion

We have analyzed an overlapping-generations environment in which two groups repeatedly play a Prisoner's Dilemma under sporadic matching. A fraction of the population consists of *honest* types, who cooperate unless they detect a past defection by their opponent, while strategic types weigh the short-run gain from defection against the stigma it imposes on future interactions. We show that in equilibrium, each strategic player's decision whether to cooperate follows a threshold rule based on his privately observed cost of being exploited. When the net benefit from defection is moderate, an increase in the fraction of honest types unambiguously raises the overall level of cooperation. However, when defection becomes very tempting, the presence of additional honest individuals can paradoxically reduce cooperation in some equilibria. Furthermore, even partial "forgiveness" through imperfect record-keeping does not boost cooperation; instead, it weakens the deterrent power of stigma.

These findings suggest that while honest types can underpin a culture of cooperation, their mere presence is not sufficient to ensure a cooperative outcome. High returns to defection or diminished punishment mechanisms may undermine this equilibrium, even if many players are intrinsically moral. We believe these conclusions have broad relevance, ranging from workplace settings and community finance to international relations, where limited or imperfectly disseminated information about past misdeeds can profoundly influence cooperation incentives.

References

Acemoglu, D. and Wolitzky, A. (2014). Cycles of conflict: An economic model. *American Economic Review*, 104(4):1350–67.

Acemoglu, D. and Wolitzky, A. (2024). Employment and Community: Socioeconomic Cooperation and Its Breakdown. NBER Working Papers 32773, National Bureau of Economic Research, Inc.

Alger, I. and Weibull, J. W. (2013). Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica*, 81(6):2269–2302.

Bahel, E., Ball, S., and Sarangi, S. (2022). Communication and cooperation in prisoner's dilemma games. *Games and Economic Behavior*, 133:126–137.

Baker, G., Gibbons, R., and Murphy, K. J. (2002). Relational contracts and the theory of the firm. *The Quarterly Journal of Economics*, 117(1):39–84.

Bar-Isaac, H. (2007). Something to prove: reputation in teams. *The RAND Journal of Economics*, 38(2):495–511.

Besley, T., Coate, S., and Loury, G. (1993). The economics of rotating savings and credit associations. *The American Economic Review*, 83(4):792–810.

Bolton, G. E., Katok, E., and Ockenfels, A. (2004). How effective are electronic reputation mechanisms? an experimental investigation. *Management Science*, 50(11):1587–1602.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

Fafchamps, M. (2000). Ethnicity and credit in african manufacturing. *Journal of Development Economics*, 61(1):205–235.

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.

Fudenberg, D., Kreps, D., and Maskin, E. (1998). Repeated games with long-run and short-run players. *David K. Levine, Levine's Working Paper Archive*, 57.

Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying aversion and the size of the lie. *American Economic Review*, 108(2):419–53.

Kets, W. and Sandroni, A. (2020). A theory of strategic uncertainty and cultural diversity. *The Review of Economic Studies*, 88(1):287–333.

Kreps, D. M., Milgrom, P., Roberts, J., and Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, 27(2):245–252.

López-Pérez, R. (2012). The power of words: A model of honesty and fairness. *Journal of Economic Psychology*, 33(3):642–658.

Lukyanov, G. and Li, D. (2025). Belief diversity and cooperation. *Journal of Economic Behavior & Organization*, 229:106815.

Miettinen, T. (2013). Promises and conventions – an approach to pre-play agreements. *Games and Economic Behavior*, 80:68–84.

Resnick, P., Zeckhauser, R., Swanson, J., and Lockwood, K. (2006). The value of reputation on ebay: A controlled experiment. *Experimental Economics*, 9(2):79–101.

Tadelis, S. (2002). The market for reputations as an incentive mechanism. *Journal of Political Economy*, 110(4):854–882.

van den Assem, M., Dolder, D., and Thaler, R. (2012). Split or steal? cooperative behavior when the stakes are large. *Management Science*, 58:2–20.