INTERFACE

royalsocietypublishing.org/rsif

CC BY

Research

Cite this article: Yuan Z, Shen C, Liu C, Perc M, Liu J, Wang Z. 2026 Public and private reputation systems for limiting centralized corruption. *J. R. Soc. Interface* **23**: 20260576.
https://doi.org/10.1098/rsif.2026.0576

Received: 25 May 2026
Accepted: 22 August 2026

Subject Category:
Life Sciences – Mathematics interface

Subject Areas:
evolution

Keywords:
evolutionary game theory, cooperation, centralized punishment, corruption, reputation systems, social norms

Authors for correspondence:
Chen Liu
e-mail: liuchen@nwpu.edu.cn
Matjaž Perc
e-mail: matjaz.perc@gmail.com
Jinzhuo Liu
e-mail: jinzhuo.liu@hotmail.com

Supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.8709999.

THE ROYAL SOCIETY PUBLISHING

Public and private reputation systems for limiting centralized corruption

Zheng Yuan[1], Chen Shen[3], Chen Liu[2], Matjaž Perc[4,5,6,7], Jinzhuo Liu[8,9] and Zhen Wang[1]

[1]School of Cybersecurity and [2]School of Life Science and Technology, Northwestern Polytechnical University, Xi'an, Shaanxi, People's Republic of China
[3]Faculty of Engineering Sciences, Kyushu University, Fukuoka, Japan
[4]Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor, Slovenia
[5]Community Healthcare Center Dr. Adolf Drolc Maribor, Maribor, Slovenia
[6]Department of Physics, Kyung Hee University, Seoul, Republic of Korea
[7]University College, Korea University, Seoul, Republic of Korea
[8]School of Software and AI, Yunnan University, Kunming, Yunnan, People's Republic of China
[9]Yunnan Key Laboratory of Software Engineering, Kunming, Yunnan, People's Republic of China

ZY, 0009-0004-9977-9401; CS, 0000-0002-2725-9763; MP, 0000-0002-3087-541X; JL, 0000-0002-3891-4663; ZW, 0000-0001-6729-2640

Corruption is widely recognized as a major challenge in human institutions. To prevent its spread, measures such as increasing transparency have attracted growing attention. Here, we develop an evolutionary game-theoretic framework with a reputation mechanism to address centralized corruption and sustain cooperation. In our framework, players participate in a donation game monitored by umpires, who enforce centralized punishment on free riders to support cooperation. However, the possibility of bribery by defectors creates incentives for umpires to abuse their authority, thereby undermining cooperation. To counter this, we construct public and private reputation systems that allow players to condition their participation decisions on the reputations of assigned umpires. Theoretical analysis based on our dynamical model, together with simulations, shows that a public reputation system, in which honest umpires receive good reputations and corrupt ones receive bad reputations, effectively reduces corruption. Interestingly, under certain social norms, private reputation systems can outperform public ones, while also offering an advantage in avoiding higher-order corruption problems inherent in centralized systems. Overall, our results highlight the potential of bottom-up transparency mechanisms in centralized institutions and demonstrate that the effectiveness of reputation systems critically depends on the careful design of social norms.

1. Introduction

Fostering cooperation among self-interested individuals has long been a central and challenging problem across fields such as biology, economics, social science and computer science [1–5]. Over the years, a wide range of mechanisms have been proposed, significantly advancing our understanding of the evolution of cooperation [6–12]. Among these, costly punishment, in which individuals incur a cost to penalize free riders, has been shown to be one of the most effective mechanisms for sustaining cooperation [13–16]. This can occur as peer punishment, where individuals directly punish others, or as centralized punishment, where a designated authority, often supported by taxes, enforces penalties.

© 2026 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0576/6176689/rsif.2026.0576.pdf
by guest
on 11 November 2026

Although both methods are effective, theoretical and field studies indicate a preference for centralized punishment to maintain cooperation [17–20].

Centralized punishment is widely used in human societies [21–23]. It appears in various forms, such as umpires penalizing rule-breakers in sports, social media platforms banning users for policy violations, regulators sanctioning firms for misconduct and judges sentencing criminals. Despite the prevalence of centralized punishment, its effectiveness can be undermined by corruption [24–27]. This occurs because those responsible for enforcing penalties may not be immune to selfish motives. The temptation to accept bribes from wrongdoers, neglect monitoring duties or assume the institution's integrity without scrutiny can lead to corruption and, ultimately, the breakdown of cooperation [28]. A widely accepted approach to mitigating corruption involves increasing the transparency of public institutions through measures such as direct disclosure of institutional information [29–31] and external supervision [32,33]. However, some measures, such as direct access to strategy information, are unrealistic and infeasible in practice. Other measures, such as external supervision, often require more powerful institutions to enforce them, which may be prone to higher-order corruption: higher-level supervisory bodies might accept bribes from lower-level institutions, rendering these measures ineffective.

To mitigate centralized corruption, we propose a novel reputation mechanism to enhance the transparency of centralized punishment enforcers, inspired by the extensive use of reputation mechanisms in social interactions and their important role in promoting reciprocity and cooperation [34–37]. We investigate how this reputation mechanism shapes population dynamics within the framework of evolutionary game theory. Specifically, we consider a population game consisting of a player population and an umpire population. For each interaction, two players are randomly selected to play a donation game, while an umpire is randomly assigned to supervise the interaction and enforce centralized punishment by penalizing defectors. Centralized corruption arises as defectors have an incentive to bribe umpires with an amount lower than the cost of cooperation to avoid punishment, while umpires may accept bribes for additional payoff. Therefore, umpires can be either honest or corrupt. Honest umpires reject bribes and punish defectors, whereas corrupt umpires accept bribes, neglect to punish defectors and may even impose anti-social punishment on cooperators. To limit the spread of corruption and the resulting collapse of cooperation in the donation game, our reputation mechanism allows players, prior to participating in the donation game, to decide whether to acquire information about the reputation of the assigned umpire. Players who access this information can further decide whether to proceed with the game or to exit based on it. Importantly, the reputation available to players in a given interaction is the result of reputation assessment based on information from previous interactions. This mechanism thus provides players with a more realistic information channel for evaluating the assigned umpire.

The introduction of the reputation mechanism raises two key questions: who evaluates umpire reputations and provides this information to players, and what norms govern such evaluations? To tackle these questions, we construct both public and private reputation systems and design different social norms that prescribe the assignment of good and bad reputations. For the public reputation system, a reputation institution evaluates umpire reputations based on their strategies, and all players unanimously agree with these assessments. By contrast, for the private reputation system, each player independently evaluates umpires based on personal observations and experiences, and stores this information for private use, with evaluations depending on both the player's strategy and the umpire's strategy.

We conduct a combined theoretical- and simulation-based analysis to investigate the role of reputation mechanisms in limiting centralized corruption. For the public reputation system, we derive a dynamical model in the form of a system of differential equations to capture the coevolution of strategies and reputations. Numerical solutions of the model, together with agent-based simulations, reveal that under a norm assigning good reputations to honest umpires and bad reputations to corrupt ones, sustained cyclic behaviour is observed, thereby preventing the population from falling into a fully corrupt state with collapsed cooperation. Moreover, compared with the baseline mechanism (i.e. a costly strategy acquisition mechanism), the public reputation system achieves higher levels of cooperation and lower levels of corruption. We further show that, under certain social norms, private reputation systems can outperform public ones. These findings highlight the importance of bottom-up approaches to enhancing transparency in centralized institutions and underscore the critical role of social norm design in ensuring the effectiveness of reputation mechanisms.

2. An evolutionary game-theoretic model with reputation mechanisms

In this section, we introduce a corruption game by extending the traditional donation game to include centralized punishment and the potential corruption of an umpire population. We establish both public and private reputation systems to enhance transparency, allowing players to make informed decisions about their participation based on available umpire reputation information.

2.1. Donation game with centralized punishment and corruption

We consider a standard donation game in which each player chooses to cooperate or defect [38–42]. A cooperator incurs a cost c to provide a benefit b to the co-player, whereas a defector incurs no cost. We use the following payoff matrix to present the resultant payoffs a focal player can get when adopting the strategies in the rows to interact with another player adopting the strategies in

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0576/6176689/rsif.2026.0576.pdf
by guest
on 11 November 2026

the columns:

	cooperate	defect
cooperate	$b-c$	$-c$
defect	b	0

Regardless of the co-player's choice, defection always yields a higher payoff, leading to mutual defection as the Nash equilibrium, which highlights the need for external mechanisms to sustain cooperation.

To sustain cooperation, we introduce centralized punishment: an umpire supervises the interaction and imposes a fine of $A > c$ on any player who defects. This fine does not contribute to the umpire's income. Each player participating in the game under the umpire's supervision pays a fee f, giving the umpire a total income of $2f$. Here, we assume that the fee f contributes to the umpire's income, whereas the fine A does not. This assumption is a typical setting in previous studies on centralized corruption [29,30,43]. Moreover, it is also consistent with many real-world settings involving third-party supervision, where individuals responsible for supervision or adjudication receive compensation for their services, while fines are typically paid to the institutions they represent rather than to the individuals themselves. The resulting payoff matrix for the donation game with centralized punishment is given by

	cooperate	defect
cooperate	$b-c-f$	$-c-f$
defect	$b-A-f$	$-A-f$

Now, cooperation becomes the dominant strategy for the focal player, provided that the umpire enforces punishment faithfully.

However, the effectiveness of centralized punishment depends critically on the umpire's honesty. In practice, the umpire may be corrupt. Corruption arises because a defector has an incentive to offer a bribe $B < c$ to the umpire to avoid severe punishment, while a corrupt umpire may accept this bribe to obtain additional payoff and allow the defector to evade penalties. In addition, we consider that corrupt umpires may exhibit another form of anti-social behaviour: they may unjustly impose an additional punishment P on cooperators, even though this yields no additional benefit to the umpire. Consequently, under centralized corruption, the payoff matrix is given as follows:

	cooperate	defect
cooperate	$b-c-f-P$	$-c-f-P$
defect	$b-B-f$	$-B-f$

As a result, centralized punishment becomes ineffective under corruption, and defection once again emerges as the dominant strategy.

2.2. Game set-up of player and umpire populations

We consider a population game consisting of a player population and an umpire population. Whenever two players are randomly selected from the player population to engage in a donation game, an umpire is simultaneously drawn at random from the umpire population to supervise the interaction. The selected umpire can be either *honest* (H) or *corrupt* (C). Based on the above analysis, a player's optimal strategy depends on the strategy of umpire randomly assigned to the game. From an evolutionary perspective, if umpires update their strategies through social learning by imitating more successful peers, corruption may gradually spread, as corrupt umpires can obtain additional income from bribes. Consequently, in the absence of additional mechanisms, corruption undermines centralized punishment; when players update their strategies by imitating more successful ones, cooperation collapses, and the population becomes trapped in a persistent social dilemma.

To address this critical issue, namely, the erosion of centralized punishment caused by corruption, this paper introduces a reputation-based mechanism to enhance the transparency of the umpire population.

2.3. A cost-free reputation acquisition mechanism

To mitigate corruption, we introduce a reputation mechanism that allows players to decide, before participating in the donation game, whether to acquire information about the umpire's reputation. Players who obtain this information can condition their participation decisions accordingly, that is, they can decide whether to proceed with or exit the game. We assume that acquiring reputation information is cost-free.

Specifically, under this reputation mechanism, players must make two decisions before engaging in the donation game: in addition to choosing whether to cooperate or defect, they must also decide whether to acquire information about the umpire's reputation. We denote players who do not acquire such information as *optimistic*, and those who do as *prudent*. Optimistic players trust their luck and always participate in the game. By contrast, prudent players first obtain the umpire's reputation and then decide whether to participate based on this information. Accordingly, based on players' strategies in the donation game (cooperation or defection) and their attitudes towards acquiring reputation information (optimistic or prudent), we redefine the player strategy

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0576/6176689/rsif.2026.0576.pdf
by guest
on 11 November 2026

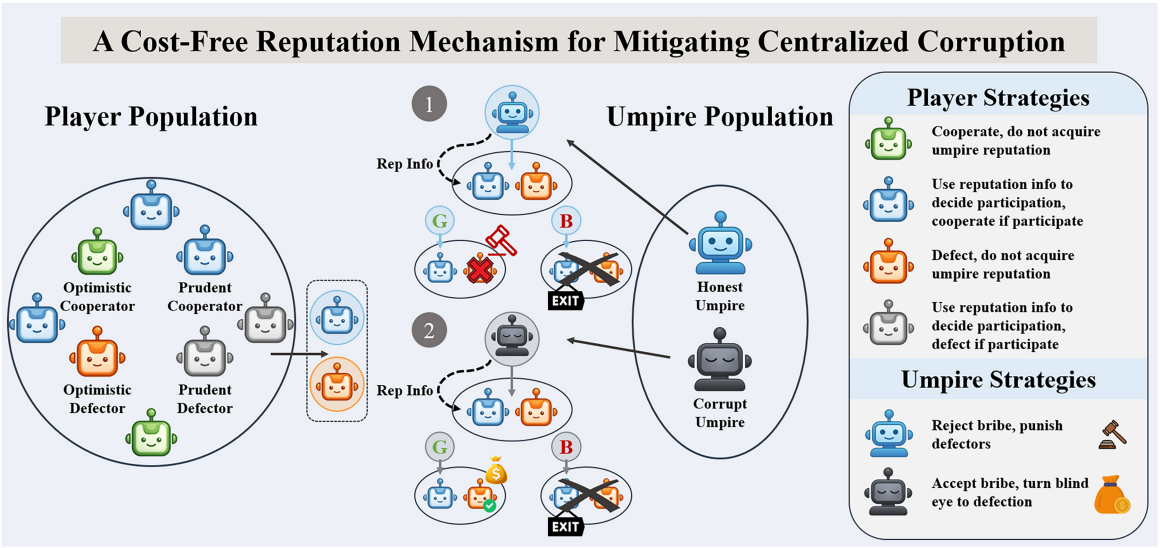


Figure 1. Schematic of the population game with a cost-free reputation mechanism. The population is assumed to be well-mixed and consists of a player population and an umpire population. In each interaction, two randomly selected players are paired to play a donation game under the supervision of a randomly selected umpire. The figure illustrates an example of interaction between a prudent cooperator and an optimistic defector. The prudent cooperator acquires the assigned umpire's reputation and decides whether to participate based on this information, whereas the optimistic defector always participates without acquiring reputation information. Depending on the strategy of the assigned umpire, two cases arise. (1) If the assigned umpire is honest, then if the acquired reputation information indicates that the umpire is good, both players proceed with the donation game and the defector is punished; otherwise, the prudent cooperator withdraws and the game does not take place. (2) If the assigned umpire is corrupt, then if the acquired reputation information indicates that the umpire is good, the game proceeds; however, the corrupt umpire accepts bribes from the defector and refrains from punishment, while the cooperator suffers anti-social punishment. The prudent cooperator withdraws when the acquired information indicates that the umpire has a bad reputation.

set in this population game to include four available strategies: *optimistically cooperate* (OC), *prudently cooperate* (PC), *optimistically defect* (OD) and *prudently defect* (PD).

Players adopting these four strategies are therefore referred to as optimistic cooperators (OCs), prudent cooperators (PCs), optimistic defectors (ODs), and prudent defectors (PDs), respectively. Although both PCs and PDs choose to acquire reputation information, they adopt different reputation-based participation rules. Under our binary reputation assessment, in which an umpire is labelled either *good* (G) or *bad* (B), PCs withdraw from the game when the acquired reputation information indicates that the assigned umpire is bad, whereas PDs withdraw when the umpire is good. This reputation-based exit behaviour affects the payoffs of both players and umpires: for each randomly matched pair of players, if either player withdraws, the game does not take place and neither the players nor the umpire receives any payoff. If both players choose to participate, they each pay a fee f to the umpire and receive the corresponding payoff from the donation game. For the umpire population, given the presence of reputation information, we define umpire types by combining their strategy and reputation. This yields four possible types: *good and honest* (GH), *good and corrupt* (GC), *bad and honest* (BH) and *bad and corrupt* (BC). Figure 1 illustrates the game environment, in which a player's payoff depends not only on its own strategy and that of its co-player, but also on the strategy and reputation of the assigned umpire. Similarly, an umpire's payoff depends on its own strategy, its reputation and the strategies of the two players.

2.4. Construction of public and private reputation systems

With the introduction of the reputation mechanism, two important questions arise. First, who evaluates the umpire's reputation and provides this information to players? Second, how is the umpire's reputation assessed, and based on what criteria? We address the first question by considering two distinct settings: public and private reputation systems. To address the second question, we design a set of social norms under each type of system to define the goodness or badness of umpires.

2.4.1. Public reputation system

We model a public reputation system by introducing a reputation institution that evaluates umpires' reputations. Specifically, in each interaction, if either player chooses to exit based on the acquired reputation of the umpire, the reputation institution does not update the umpire's reputation because the donation game is not actually played. If neither player exits, the reputation institution updates the umpire's reputation according to a prescribed social norm. As is common in models of public reputation [44–47],

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0576/6176689/rsif.2026.0576.pdf
by guest
on 11 November 2026

1 **Social Norms for Public Reputation System**

	Norm 00	Norm 01	Norm 10	Norm 11
Honest umpires are assessed as	G	G	B	B
Corrupt umpires are assessed as	G	B	G	B

2 **Social Norms for Private Reputation System**

	Norm 0000	Norm 0001	Norm 0010	Norm 0011	Norm 0100	Norm 0101	Norm 0110	Norm 0111	Norm 1000	Norm 1001	Norm 1010	Norm 1011	Norm 1100	Norm 1101	Norm 1110	Norm 1111
Cooperators assess **honest** umpires as	G	G	G	G	G	G	G	G	B	B	B	B	B	B	B	B
Cooperators assess **corrupt** umpires as	G	G	G	G	B	B	B	B	G	G	G	G	B	B	B	B
Defectors assess **honest** umpires as	G	G	B	B	G	G	B	B	G	G	B	B	G	G	B	B
Defectors assess **corrupt** umpires as	G	B	G	B	G	B	G	B	G	B	G	B	G	B	G	B

Figure 2. Social norms for two distinct reputation systems. In the public reputation system, the reputation assigned to an umpire by the public institution depends solely on the umpire's strategy. Accordingly, each norm can be represented by a 2-bit string, where the first bit denotes the reputation assigned to an honest umpire and the second bit to a corrupt umpire. Here, 0 indicates a good reputation and 1 indicates a bad reputation. In the private reputation system, when an interaction occurs and neither of the two involved players exits, both participants and all other players independently evaluate the assigned umpire's reputation. This evaluation depends on the evaluator's own strategy type (i.e. whether the player is cooperative or defective) and the umpire's strategy. Thus, each norm is encoded by a 4-bit string: the first two bits indicate the reputations assigned by cooperators to honest and corrupt umpires, respectively, and the last two bits indicate the reputations assigned by defectors to honest and corrupt umpires, respectively. Again, 0 denotes a good reputation, and 1 denotes a bad reputation.

we assume that reputation information is publicly available and that all players share the same assessment of the published evaluations.

As for the norms, we consider simple norms for public reputation assessment that assign reputations based solely on the umpire's strategy. Consequently, we consider four distinct social norms, as shown in figure 2. For instance, norm 01 prescribes assigning a good reputation to honest umpires and a bad reputation to corrupt ones. As this norm assigns reputations according to the moral quality of behaviour, it is analogous to the 'image scoring' in the traditional donation game [48], where cooperators are assigned good reputations and defectors bad reputations. Moreover, it is conceptually consistent with the widely adopted reputation assessment norm, under which cooperation increases reputation scores whereas defection decreases them [49–52].

2.4.2. Private reputation system

In many real-world situations, in addition to relying on evaluations from public institutions, individuals often form subjective moral judgments about institutions or other individuals based on personal observations and experiences [53–56]. Therefore, we also construct a private reputation system in which players independently evaluate and privately maintain the reputation of each umpire.

In this private reputation system, when an umpire is assigned to oversee an interaction between two players, each player decides whether to exit the game based on their own assessment of the umpire's reputation. If either player exits, the umpire's reputation remains unchanged in the eyes of all players. If the interaction proceeds, both participating players and all other players update their assessments of the umpire's reputation according to a common social norm.

For private reputation assessment, we consider norms that depend on both the evaluator's strategy type and the umpire's strategy. Here, the strategy type refers to whether a player is cooperative (including OCs and PCs) or defective (including ODs and PDs). Consequently, there are 16 possible norms, as shown in figure 2. For example, the norm 0110 specifies that cooperators assign a good reputation to honest umpires and a bad reputation to corrupt ones, while defectors assign the opposite reputations.

Notably, the private reputation system differs fundamentally from the public one. Under private assessment, players may hold divergent views of the same umpire. Specifically, in the public system, an umpire's reputation is identical across all players, whereas in the private system, it may vary across individuals. For instance, an honest umpire may be regarded as good by some players but bad by others. Nevertheless, under either reputation system, once reputations are established and no player withdraws, an umpire's current strategic behaviour affects their future reputation. This, in turn, influences players' participation decisions in subsequent interactions, ultimately shaping the umpire's future payoffs and strategic choices.

3. Theoretical analysis of corruption and cooperation dynamics

To analyse the evolutionary dynamics of the proposed population game, we assume that both players and umpires update their strategies through imitation and mutation. Occasionally, with a small probability, a player (or umpire) randomly adopts a new

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0576/6176689/rsif.2026.0576.pdf
by guest
on 11 November 2026

strategy, with all other available strategies having the same probability of being picked. Otherwise, the player (or umpire) imitates another player's (or umpire's) strategy with a certain likelihood. Note that under the reputation mechanism, reputation and strategy are updated on the same evolutionary timescale: strategy updating determines the new strategy for subsequent interactions, while reputation updating assigns a new reputation for subsequent interactions based on information from the current interaction.

In this section, we first introduce the mechanism proposed by Lee *et al.* [30]. Building on this framework, we construct a baseline mechanism, referred to as the costly strategy acquisition mechanism, for comparison with our cost-free reputation mechanism. We then derive dynamical models to characterize the imitation–mutation dynamics in a well-mixed player–umpire population under two distinct mechanisms: the baseline mechanism and our reputation mechanism under a public reputation system.

3.1. Evolutionary dynamics under the costly strategy acquisition mechanism

Different from our reputation-based mechanism, Lee *et al.* [30] propose a strategy information acquisition mechanism that allows prudent players to incur an information cost h to directly access the strategy of the assigned umpire. Under this mechanism, PCs exit the game when they learn that the umpire is corrupt, while PDs exit when the umpire is honest. In this paper, we incorporate anti-social punishment on cooperators imposed by corrupt umpires into this framework, thereby constructing a baseline mechanism that enables comparison with our proposed cost-free reputation mechanism.

For the population game under the baseline mechanism, we denote the set of player strategies as $\mathcal{S}_p = \{\mathrm{OC}, \mathrm{PC}, \mathrm{OD}, \mathrm{PD}\}$ and the set of umpire strategies as $\mathcal{S}_u = \{\mathrm{H}, \mathrm{C}\}$. The payoff of a player depends on its own strategy, that of its co-player and the strategy of the assigned umpire. Similarly, the payoff of an umpire depends on the strategies of the paired players as well as its own strategy. Therefore, for each umpire strategy, we construct the payoff matrices for players, as well as the corresponding payoff matrices for umpires. Specifically, let $\mathbf{Q}^{\mathrm{H}}$ and $\mathbf{Q}^{\mathrm{C}}$ denote the payoff matrices for players interacting under honest and corrupt umpires, respectively:

$$\mathbf{Q}^{\mathrm{H}} = \begin{pmatrix} b-c-f & b-c-f & -c-f & 0 \\ b-c-f-h & b-c-f-h & -c-f-h & -h \\ b-f-A & b-f-A & -f-A & 0 \\ -h & -h & -h & -h \end{pmatrix} \text{and } \mathbf{Q}^{\mathrm{C}} = \begin{pmatrix} b-c-f-P & 0 & -c-f-P & -c-f-P \\ -h & -h & -h & -h \\ b-f-B & 0 & -f-B & -f-B \\ b-f-B-h & -h & -f-B-h & -f-B-h \end{pmatrix}.$$

The entry $\mathbf{Q}^{\mathrm{H}}_{ij}$ (or $\mathbf{Q}^{\mathrm{C}}_{ij}$) represents the payoff of the row player when interacting under an honest (or corrupt) umpire, where the row player adopts the i-th strategy and the column player adopts the j-th strategy, with the ordering given by $\mathcal{S}_p = \{\mathrm{OC}, \mathrm{PC}, \mathrm{OD}, \mathrm{PD}\}$.

Similarly, let $\mathbf{P}^{\mathrm{H}}$ and $\mathbf{P}^{\mathrm{C}}$ denote the payoff matrices for honest and corrupt umpires, respectively:

$$\mathbf{P}^{\mathrm{H}} = \begin{pmatrix} 2f & 2f & 2f & 0 \\ 2f & 2f & 2f & 0 \\ 2f & 2f & 2f & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \text{and } \mathbf{P}^{\mathrm{C}} = \begin{pmatrix} 2f & 0 & 2f+B & 2f+B \\ 0 & 0 & 0 & 0 \\ 2f+B & 0 & 2f+2B & 2f+2B \\ 2f+B & 0 & 2f+2B & 2f+2B \end{pmatrix}.$$

The entry $\mathbf{P}^{\mathrm{H}}_{ij}$ (or $\mathbf{P}^{\mathrm{C}}_{ij}$) denotes the payoff of an honest (or corrupt) umpire when the row player adopts the i-th strategy and the column player adopts the j-th strategy, with the ordering given by $\mathcal{S}_p$.

We denote the frequencies of OCs, PCs, ODs and PDs in the player population by $f_{\mathrm{OC}}, f_{\mathrm{PC}}, f_{\mathrm{OD}}$ and f_{PD}, respectively. Similarly, we use f_{H} and f_{C} to denote the frequencies of honest and corrupt umpires in the umpire population. Following Lee *et al.* [30], in the limit of an infinitely large well-mixed player–umpire population, the time evolution of strategy frequencies is governed by the following replicator–mutator equations:

$$\begin{aligned} &\dot{f}_i = f_i(R^p_i - \overline{R}_p) - \mu f_i + \mu(1-f_i)/3, \quad i \in \mathcal{S}_p \\ &\text{and } \dot{f}_{\mathrm{H}} = f_{\mathrm{H}}(R^u_{\mathrm{H}} - \overline{R}_u) - \nu f_{\mathrm{H}} + \nu(1-f_{\mathrm{H}}), \end{aligned} \tag{3.1}$$

where R^p_i denotes the expected payoff of a player using strategy i, and $\overline{R}_p = \sum_{i\in\mathcal{S}_p} f_i R^p_i$ is the average payoff of the player population. Specifically,

$$R^p_i = \sum_{j\in\mathcal{S}_p} f_j \left(f_{\mathrm{H}} \mathbf{Q}^{\mathrm{H}}_{ij} + f_{\mathrm{C}} \mathbf{Q}^{\mathrm{C}}_{ij} \right).$$

Similarly, R^u_{H} and R^u_{C} denote the expected payoffs of honest and corrupt umpires, respectively, given by

$$R^u_{\mathrm{H}} = \sum_{i\in\mathcal{S}_p} \sum_{j\in\mathcal{S}_p} f_i f_j \mathbf{P}^{\mathrm{H}}_{ij} \text{ and} R^u_{\mathrm{C}} = \sum_{i\in\mathcal{S}_p} \sum_{j\in\mathcal{S}_p} f_i f_j \mathbf{P}^{\mathrm{C}}_{ij},$$

and $\overline{R}_u = \sum_{k\in\mathcal{S}_u} f_k R^u_k$ denotes the average payoff of the umpire population. In addition, μ and ν denote the mutation rates of players and umpires, respectively.

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0576/6176689/rsif.2026.0576.pdf
by guest
on 11 November 2026

3.2. Coevolutionary dynamics of reputation and strategy under the public reputation system

Under the proposed public reputation system, in addition to updating their strategies, umpires also have their reputations updated by a public institution according to a social norm, provided that neither of the interacting players exits the game. As a result, while the dynamics of the player population can still be described by the evolution of player strategies, the dynamics of the umpire population are governed by a coevolutionary process involving both reputation and strategy. This coevolutionary process can be characterized by the evolution of umpire types, which we model through a new dynamical model.

Before formulating the dynamical model, we first analyse how the payoff matrices for players and umpires change under the public reputation system. Under public reputation assessment, all players share a common view of each umpire's reputation. Therefore, the perceived type of each umpire is identical across all players, allowing us to construct payoff matrices conditioned on umpire types.

Let $\mathcal{T} = \{\text{GH, GC, BH, BC}\}$ denote the set of umpire types. We denote by $\mathbf{Q}^{\text{GH}}$, $\mathbf{Q}^{\text{GC}}$, $\mathbf{Q}^{\text{BH}}$ and $\mathbf{Q}^{\text{BC}}$ the payoff matrices for players under the corresponding umpire types. By construction, $\mathbf{Q}^{\text{GH}}$ coincides with $\mathbf{Q}^{\text{H}}$ in the baseline mechanism when $h = 0$, and $\mathbf{Q}^{\text{BC}}$ coincides with $\mathbf{Q}^{\text{C}}$ when $h = 0$. The payoff matrices under umpire types GC and BH are given by

$$\mathbf{Q}^{\text{GC}} = \begin{pmatrix} b-c-f-P & b-c-f-P & -c-f-P & 0 \\ b-c-f-P & b-c-f-P & -c-f-P & 0 \\ b-f-B & b-f-B & -f-B & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \text{and } \mathbf{Q}^{\text{BH}} = \begin{pmatrix} b-c-f & 0 & -c-f & -c-f \\ 0 & 0 & 0 & 0 \\ b-f-A & 0 & -f-A & -f-A \\ b-f-A & 0 & -f-A & -f-A \end{pmatrix}.$$

Similarly, we construct the payoff matrices for umpires under each type. Let $\mathbf{P}^{\text{GH}}$, $\mathbf{P}^{\text{GC}}$, $\mathbf{P}^{\text{BH}}$ and $\mathbf{P}^{\text{BC}}$ denote the payoff matrices for umpires of the corresponding types. Here, $\mathbf{P}^{\text{GH}}$ and $\mathbf{P}^{\text{BC}}$ coincide with $\mathbf{P}^{\text{H}}$ and $\mathbf{P}^{\text{C}}$ in the baseline mechanism, respectively. The matrices for types GC and BH are given by

$$\mathbf{P}^{\text{GC}} = \begin{pmatrix} 2f & 2f & 2f+B & 0 \\ 2f & 2f & 2f+B & 0 \\ 2f+B & 2f+B & 2f+2B & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \text{and } \mathbf{P}^{\text{BH}} = \begin{pmatrix} 2f & 0 & 2f & 2f \\ 0 & 0 & 0 & 0 \\ 2f & 0 & 2f & 2f \\ 2f & 0 & 2f & 2f \end{pmatrix}.$$

These results highlight a key feature of the reputation mechanism: while reputation information allows players to make exit decisions aligned with their interests in interactions under GH and BC umpires, the emergence of GC and BH umpires reveals potential mismatches between obtained reputation and actual strategies.

3.2.1. Evolutionary dynamics model of player strategies

We now derive the dynamical model governing the evolution of player strategies. Changes in the proportions of strategies arise from the combined effects of imitation and mutation. Following the framework of replicator–mutator dynamics [57], we first characterize the rates of decrease and increase owing to imitation, and then incorporate mutation.

Under the imitation mechanism, a player updates its strategy by randomly selecting another player as a role model and then adopting that player's strategy with a certain probability. Consequently, for an arbitrary strategy $i \in S_p$, its frequency decreases when players currently adopting strategy i imitate players using different strategies. This leads to

$$\frac{\mathrm{d}f_i^-}{\mathrm{d}t} = f_i \sum_{j\in S_p} \sum_{k\in\mathcal{T}} f_j f_k \left[\sum_{i'\in S_p\setminus\{i\}} \sum_{j'\in S_p} \sum_{k'\in\mathcal{T}} f_{i'} f_{j'} f_{k'} W_{i\to i'}(\mathbf{Q}_{ij}^k, \mathbf{Q}_{i'j'}^{k'}) \right], \tag{3.2}$$

where $\mathcal{T} = \{\text{GH, GC, BH, BC}\}$ denotes the set of umpire types. $\mathbf{Q}_{ij}^k$ represents the payoff of a player using strategy i against a co-player using strategy j under an umpire of type k. $W_{i\to i'}(\mathbf{Q}_{ij}^k, \mathbf{Q}_{i'j'}^{k'})$ denotes the probability that a player using strategy i, who obtains payoff $\mathbf{Q}_{ij}^k$, imitates a role model using strategy i', who obtains payoff $\mathbf{Q}_{i'j'}^{k'}$. This probability is determined by the strategy updating rule. We consider the widely used Fermi updating rule [58,59], given by

$$W_{i\to i'}(\mathbf{Q}_{ij}^k, \mathbf{Q}_{i'j'}^{k'}) = \left[1 + \exp\left(s(\mathbf{Q}_{ij}^k - \mathbf{Q}_{i'j'}^{k'})\right)\right]^{-1},$$

where s denotes the selection strength.

Similarly, the increase in the frequency of strategy i arises from players using other strategies imitating strategy i:

$$\frac{\mathrm{d}f_i^+}{\mathrm{d}t} = \sum_{i'\in S_p\setminus\{i\}} \sum_{j'\in S_p} \sum_{k'\in\mathcal{T}} f_{i'} f_{j'} f_{k'} \left[f_i \sum_{j\in S_p} \sum_{k\in\mathcal{T}} f_j f_k W_{i'\to i}(\mathbf{Q}_{i'j'}^{k'}, \mathbf{Q}_{ij}^k) \right]. \tag{3.3}$$

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0576/6176689/rsif.2026.0576.pdf
by guest
on 11 November 2026

Finally, incorporating mutation, the overall dynamics of strategy frequencies are given by

$$\frac{\mathrm{d}f_i}{\mathrm{d}t} = \frac{\mathrm{d}f_i^+}{\mathrm{d}t} - \frac{\mathrm{d}f_i^-}{\mathrm{d}t} - \mu f_i + \mu(1-f_i)/3, \quad i \in S_p. \tag{3.4}$$

3.2.2. Coevolutionary dynamics of umpire strategy and reputation

We derive a dynamical model that captures the evolution of umpire types by analysing how strategy updating and reputation updating jointly induce transitions between different types. Although the transition patterns differ across umpire types, the underlying analytical approach is similar. Therefore, we focus on the evolution of the proportion of GH umpires to illustrate the derivation of the dynamical model for the umpire population.

We first consider the case in which umpire strategies are updated solely through the imitation mechanism, and examine how strategy updating and reputation updating jointly reduce the proportion of GH umpires. Specifically, a decrease in GH umpires may arise in two ways. First, the reputation of a GH umpire may be updated from G to B. In this case, regardless of whether its strategy changes, the umpire transitions to type BH or BC. Second, the reputation may remain unchanged, but the umpire switches its strategy from H to C by imitating a GC or BC role model. Combining these effects, the rate at which the proportion of GH umpires decreases is given by

$$\frac{\mathrm{d}f_{\mathrm{GH}}^-}{\mathrm{d}t} = f_{\mathrm{GH}} \sum_{i,j\in S_p} f_i f_j \Big[p(\mathrm{G},i,j)p(\mathrm{B}|\mathrm{H}) + \big(p(\mathrm{G},i,j)p(\mathrm{G}|\mathrm{H}) + (1-p(\mathrm{G},i,j))\big) \sum_{i',j'\in S_p} f_{i'} f_{j'} \big(f_{\mathrm{GC}} W_{\mathrm{GH}\to\mathrm{GC}}(\mathbf{P}_{ij}^{\mathrm{GH}}, \mathbf{P}_{i'j'}^{\mathrm{GC}}) + f_{\mathrm{BC}} W_{\mathrm{GH}\to\mathrm{BC}}(\mathbf{P}_{ij}^{\mathrm{GH}}, \mathbf{P}_{i'j'}^{\mathrm{BC}}) \big) \Big]. \tag{3.5}$$

Here, the social norm is characterized by the probabilities $p(\mathrm{G}|\mathrm{H})$, $p(\mathrm{B}|\mathrm{H})$, $p(\mathrm{G}|\mathrm{C})$ and $p(\mathrm{B}|\mathrm{C})$, where $p(Y|X)$ denotes the probability that an umpire using strategy $X \in \{\mathrm{H},\mathrm{C}\}$ is assigned reputation $Y \in \{\mathrm{G},\mathrm{B}\}$. Since reputation is updated only when no player exits, we define $p(\mathrm{G},i,j)$ as the probability that neither player exits given that the interacting players adopt strategies (i,j), and the umpire is currently assessed as good. Analogously, $p(\mathrm{B},i,j)$ denotes the probability that neither player exits when the umpire is assessed as bad, and the interacting players adopt strategies (i,j). In addition, $W_{\mathrm{GH}\to\mathrm{GC}}(\mathbf{P}_{ij}^{\mathrm{GH}}, \mathbf{P}_{i'j'}^{\mathrm{GC}})$ denotes the probability that a GH umpire imitates a GC role model based on payoff comparison. Specifically, a GH umpire observing a pair of players with strategies (i,j) obtains payoff $\mathbf{P}_{ij}^{\mathrm{GH}}$, and compares it with the payoff $\mathbf{P}_{i'j'}^{\mathrm{GC}}$ obtained by a GC role model observing (i',j'). The imitation probability is given by the Fermi updating rule. The definition of $W_{\mathrm{GH}\to\mathrm{BC}}(\mathbf{P}_{ij}^{\mathrm{GH}}, \mathbf{P}_{i'j'}^{\mathrm{BC}})$ is analogous.

Next, we examine how the proportion of GH umpires increases under the combined effects of strategy imitation and reputation updating. The increase in GH umpires arises from transitions from the other three types. First, GC umpires may become GH if their reputation remains good and they switch their strategy to H by imitating GH or BH umpires. Second, BH umpires may become GH if their reputation is updated to good and they retain strategy H. Third, BC umpires may become GH if their reputation is updated to good and they switch their strategy to H by imitating GH or BH umpires. Accordingly, the increase rate of GH umpires is given by

$$\begin{aligned}
\frac{\mathrm{d}f_{\mathrm{GH}}^+}{\mathrm{d}t} =& f_{\mathrm{GC}} \sum_{i\in S_p}\sum_{j\in S_p} f_i f_j \Big[\Big(p(\mathrm{G},i,j)p(\mathrm{G}|\mathrm{C}) + \big(1-p(\mathrm{G},i,j)\big) \Big) \sum_{i'\in S_p}\sum_{j'\in S_p} f_{i'} f_{j'} \Big(f_{\mathrm{GH}} W_{\mathrm{GC}\to\mathrm{GH}}(\mathbf{P}_{ij}^{\mathrm{GC}}, \mathbf{P}_{i'j'}^{\mathrm{GH}}) + f_{\mathrm{BH}} W_{\mathrm{GC}\to\mathrm{BH}}(\mathbf{P}_{ij}^{\mathrm{GC}}, \mathbf{P}_{i'j'}^{\mathrm{BH}}) \Big) \Big] \\
&+ f_{\mathrm{BH}} \sum_{i\in S_p}\sum_{j\in S_p} f_i f_j \Big[p(\mathrm{B},i,j)p(\mathrm{G}|\mathrm{H}) \sum_{i'\in S_p}\sum_{j'\in S_p} f_{i'} f_{j'} \Big(f_{\mathrm{GH}} + f_{\mathrm{BH}} + f_{\mathrm{GC}}\big(1 - W_{\mathrm{BH}\to\mathrm{GC}}(\mathbf{P}_{ij}^{\mathrm{BH}}, \mathbf{P}_{i'j'}^{\mathrm{GC}})\big) + f_{\mathrm{BC}}\big(1 - W_{\mathrm{BH}\to\mathrm{BC}}(\mathbf{P}_{ij}^{\mathrm{BH}}, \mathbf{P}_{i'j'}^{\mathrm{BC}})\big) \Big) \Big] \\
&+ f_{\mathrm{BC}} \sum_{i\in S_p}\sum_{j\in S_p} f_i f_j \Big[p(\mathrm{B},i,j)p(\mathrm{G}|\mathrm{C}) \sum_{i'\in S_p}\sum_{j'\in S_p} f_{i'} f_{j'} \big(f_{\mathrm{GH}} W_{\mathrm{BC}\to\mathrm{GH}}(\mathbf{P}_{ij}^{\mathrm{BC}}, \mathbf{P}_{i'j'}^{\mathrm{GH}}) + f_{\mathrm{BH}} W_{\mathrm{BC}\to\mathrm{BH}}(\mathbf{P}_{ij}^{\mathrm{BC}}, \mathbf{P}_{i'j'}^{\mathrm{BH}}) \big) \Big].
\end{aligned} \tag{3.6}$$

Consequently, the rate of change in the proportion of GH umpires under the imitation mechanism and reputation dynamics is given by the difference between the increase rate (equation (3.6)) and the decrease rate (equation (3.5)).

To derive the imitation–mutation dynamics for the umpire population, we now consider the case in which umpires update their strategies through mutation, and examine how strategy updating and reputation updating jointly affect the proportion of GH umpires. Specifically, under the mutation mechanism, a GH umpire may switch its strategy from H to C. Consequently, regardless of whether its reputation changes, mutation always leads to a decrease in the proportion of GH umpires. Conversely, the increase in the proportion of GH umpires arises from the conversion of GC and BC umpires. In particular, GC umpires contribute to the increase when their reputations remain unchanged and their strategies mutate from C to H. Similarly, BC umpires contribute to the increase when their strategies mutate from C to H and their reputations are updated from B to G.

With mutation rate ν, the imitation–mutation dynamics for GH umpires is given by

$$\frac{\mathrm{d}f_{\mathrm{GH}}}{\mathrm{d}t} = \frac{\mathrm{d}f_{\mathrm{GH}}^+}{\mathrm{d}t} - \frac{\mathrm{d}f_{\mathrm{GH}}^-}{\mathrm{d}t} - \nu f_{\mathrm{GH}} + \nu f_{\mathrm{GC}} \sum_{i,j} f_i f_j \Big[p(\mathrm{G},i,j)p(\mathrm{G}|\mathrm{C}) + \big(1-p(\mathrm{G},i,j)\big) \Big] + \nu f_{\mathrm{BC}} \sum_{i,j} f_i f_j \, p(\mathrm{B},i,j)p(\mathrm{G}|\mathrm{C}). \tag{3.7}$$

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0576/6176689/rsif.2026.0576.pdf
by guest
on 11 November 2026

The dynamics for the other umpire types (GC, BH and BC) can be derived analogously and are provided in the supplementary material.

4. Experimental results

In this section, we investigate the effectiveness of the proposed cost-free reputation mechanism in limiting centralized corruption. We further examine how different social norms, under both public and private reputation systems, influence the suppression of corruption and the promotion of cooperative behaviour.

4.1. Experiment settings

We conduct experiments by combining theoretical analysis with agent-based simulations. For both the baseline mechanism and the reputation mechanism under the public reputation system, we provide theoretical results by numerically solving the corresponding dynamical models, and present simulation results to validate the effectiveness of our analytical results. For the private reputation system, owing to the analytical intractability, we rely on simulations to gain insights into the coevolutionary dynamics of reputation and strategies. Unless otherwise specified, we set the game parameters as follows: $b = 1$, $c = 0.5$, $f = 0.2$, $A = 2$, $B = 0.2$ and $P = 0.1$.

We perform simulations using the Monte Carlo method. Specifically, we assume that the player and umpire populations are of equal size ($N_p = N_u = N$). Each simulation run consists of multiple Monte Carlo steps (MCSs), where each MCS comprises N elementary steps. On average, each player and each umpire has one opportunity to update their strategy per MCS. In each elementary step, a focal player X and a co-player are randomly selected to play the game. Simultaneously, a focal umpire Y is randomly chosen to oversee the interaction. The focal player X and the focal umpire Y then receive their respective payoffs, denoted by P_X and P_Y. After receiving payoffs, both the focal player and the focal umpire update their strategies via mutation or imitation. Specifically, with probability μ, the focal player mutates to a randomly selected alternative strategy, where all other strategies are chosen with equal probability. Otherwise, with probability $1 - \mu$, the focal player selects a role model X' from the player population. The role model interacts with a randomly chosen co-player under the supervision of a randomly assigned umpire, obtaining a payoff $P_{X'}$. Player X then adopts the strategy of X' with probability $1/\left[1 + \exp\left(s(P_X - P_{X'})\right)\right]$. Similarly, the focal umpire updates its strategy as follows: with probability ν, it switches to a randomly selected alternative strategy; otherwise, with probability $1 - \nu$, it selects a role model Y' from the umpire population. The role model Y' obtains its payoff by supervising the interaction between two randomly selected players. Umpire Y then adopts the strategy of Y' with probability $1/\left[1 + \exp\left(s(P_Y - P_{Y'})\right)\right]$. In addition, if a reputation system is in place and neither the focal player X nor its co-player exits the game, the focal umpire Y's reputation is updated according to the specified social norm, in parallel with the strategy-updating process. The above procedure constitutes a single elementary step.

4.2. Cyclic behaviour under reputation mechanisms

To demonstrate the impact of our reputation mechanism on mitigating corruption, we analyse the temporal evolution of population behaviour under both the baseline mechanism and the proposed reputation mechanism.

We first present the time evolution of player and umpire strategies under the baseline mechanism in figure 3. Both the theoretical predictions and agent-based simulation results exhibit sustained oscillatory dynamics, where decreases (or increases) in corruption correspond to increases (or decreases) in cooperation. Specifically, these oscillations arise as follows. An increase in honest umpires benefits OCs, allowing them to dominate. However, mutations gradually reintroduce corruption, which undermines OCs and enables PDs to gain the upper hand, leading to the growth of corruption. As defection and corruption peak, PCs become dominant, which in turn promotes the resurgence of honest umpires and cooperative behaviour.

Compared to the case without any transparency-enhancing mechanism—where all players remain optimistic and bribery becomes the dominant strategy for umpires, resulting in pervasive corruption and the collapse of cooperation—the oscillatory dynamics induced by the baseline mechanism prevent the population from being trapped in a fully corrupt state. Nevertheless, this mechanism relies on direct access to umpire strategies and is therefore overly idealized.

We next examine our proposed reputation mechanism, which provides a more practical approach to enhancing transparency. When reputation information replaces direct strategy information, its effectiveness depends on whether it enables players to make participation decisions aligned with their self-interest. This, in turn, depends on the design of the social norm. Recall that PCs participate only when the umpire is perceived as good, whereas PDs participate only when the umpire is perceived as bad. Therefore, we adopt norm 01 for the public reputation system, which assigns a good reputation to honest umpires and a bad reputation to corrupt ones. Among the four possible public evaluation rules, this norm most effectively ensures that reputation reflects the underlying strategy of the umpire and thereby supports informed decision-making by players.

The evolutionary dynamics under this norm are illustrated in figure 4. As shown, our theoretical model, which captures the coevolution of reputation and strategy, closely matches the simulation results. Moreover, sustained oscillatory behaviour is observed: as the proportion of GH umpires increases, OCs benefit and become more prevalent. However, once cooperation reaches a sufficiently high level, interactions between defectors and BC umpires promote their growth, leading to losses for OCs. As BC umpires increase, PCs withdraw from the game to avoid losses, reducing the profitability of corruption and driving BC umpires

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0576/6176689/rsif.2026.0576.pdf
by guest
on 11 November 2026

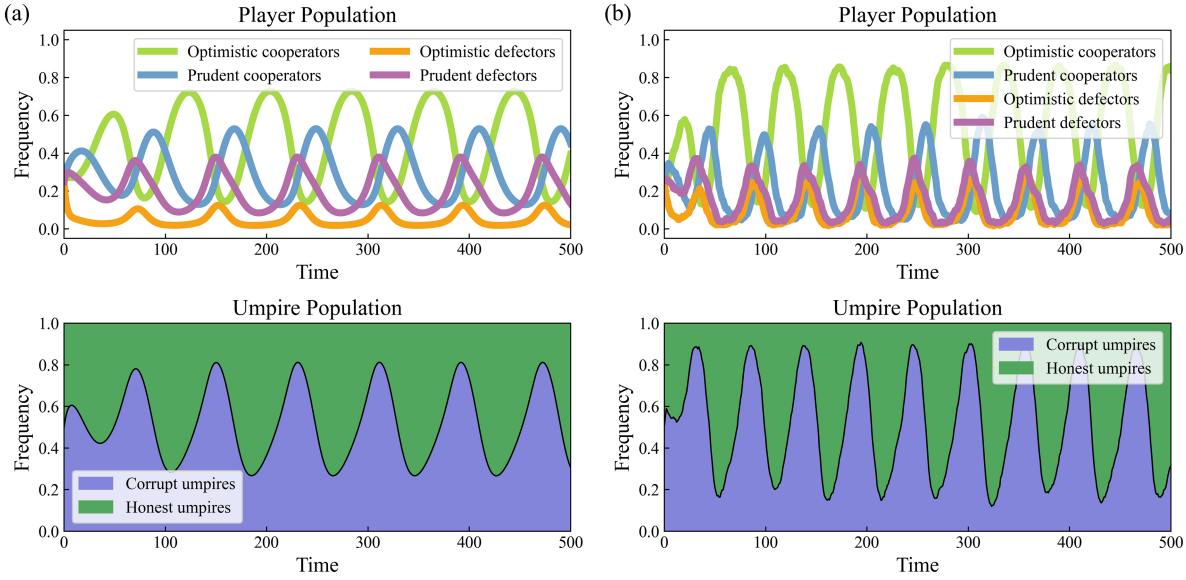


Figure 3. Population dynamics under the baseline mechanism. (a) Evolutionary dynamics predicted by the replicator–mutator equations. (b) Results obtained from agent-based simulations. Unless otherwise specified, we set the mutation rates to $\mu = \nu = 0.05$ and the strategy information cost to $h = 0.1$. The initial conditions for the numerical solution are $f_{OC} = 0.25$, $f_{PC} = 0.25$, $f_{OD} = 0.25$, $f_{PD} = 0.25$, $f_H = 0.5$ and $f_C = 0.5$. For agent-based simulations, we set $N = 3000$ and $s = 5$, with all individuals randomly assigned strategies at initialization.

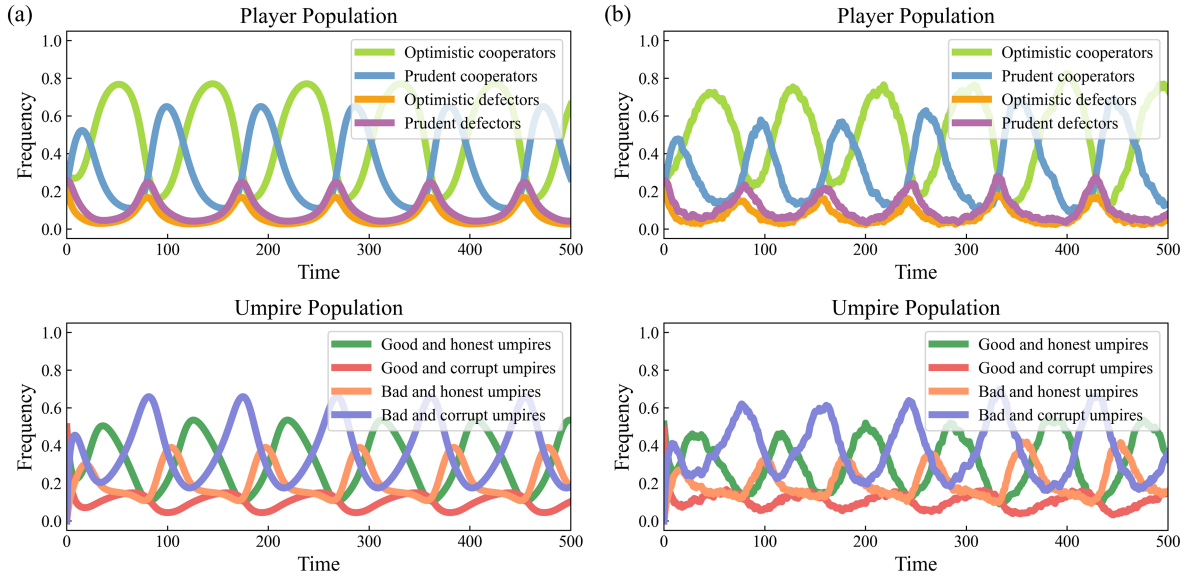


Figure 4. Long-term oscillatory population dynamics under the public reputation system with social norm 01, which assigns good reputations to honest umpires and bad reputations to corrupt umpires. (a) Population dynamics predicted by the theoretical model. (b) Population dynamics observed in agent-based simulations. For agent-based simulations, all umpires are initialized with a good reputation.

to adopt honest strategies. As PCs decrease, more OCs re-enter interactions with BH umpires, improving their reputations. The subsequent increase in GH umpires further promotes the resurgence of OCs. Therefore, similar to the baseline mechanism, our reputation system prevents pervasive corruption and the collapse of cooperation, while relying on a more realistic form of information.

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0576/6176689/rsif.2026.0576.pdf
by guest
on 11 November 2026

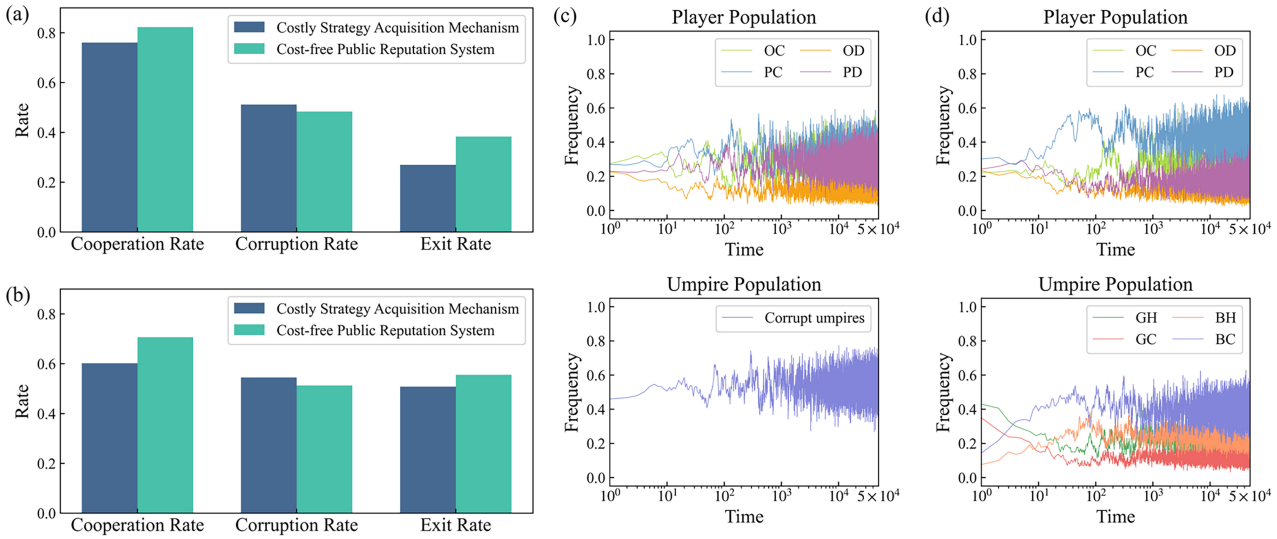


Figure 5. Public reputation system outperforms the baseline mechanism in reducing corruption and promoting cooperation. Average cooperation rate, corruption rate and exit rate under the baseline mechanism and the public reputation system are compared. (a) Average results for a large population ($N = 3000$) with strong selection intensity ($s = 5$). (b) Average results for a small population ($N = 500$) with weak selection intensity ($s = 0.5$). The results in panels (a) and (b) are averaged over 20 independent runs, with each run averaged over the last 5000 MCSs out of a total of 50 000 MCSs. The exit rate in a given MCS is defined as the proportion of interactions (out of all N interactions) in which at least one player chooses to exit. (c) Temporal evolution under the baseline mechanism for $N = 500$ and $s = 0.5$. (d) Temporal evolution under the public reputation system for $N = 500$ and $s = 0.5$.

4.3. Comparison of the public reputation system and the baseline mechanism

To provide a more detailed comparison between the baseline mechanism and our public reputation system, we conduct a quantitative analysis based on the average outcomes of multiple independent simulation runs. Specifically, focusing on the public reputation system under norm 01, which assigns good reputations to honest umpires and bad reputations to corrupt ones, we compare it with the baseline mechanism in terms of cooperation level, corruption rate and exit rate.

As shown in figure 5a, under a large population size ($N = 3000$) and strong selection intensity ($s = 5$), where cyclic dynamics are observed (figures 3 and 4), the averaged results indicate that the cost-free public reputation system with norm 01 yields a higher proportion of cooperative players and a lower level of corruption compared to the baseline mechanism. In addition, we observe a higher exit rate under the reputation system. This suggests that the way in which reputation information is used, namely to decide whether to participate in interactions, plays a crucial role in the effectiveness of the mechanism.

To examine the robustness of these findings, we further consider a different parameter setting with a smaller population ($N = 500$) and weaker selection intensity ($s = 0.5$), as shown in figure 5b. The results show that the public reputation system continues to outperform the baseline mechanism, leading to higher cooperation levels and lower corruption rates. Furthermore, figure 5c,d presents the evolutionary dynamics obtained from agent-based simulations under $N = 500$ and $s = 0.5$.

Additional comparative experiments under different parameter settings, including $N = 3000$, $s = 1$ and $N = 3000$, $s = 10$, are provided in the supplementary material. Overall, these findings demonstrate that the public reputation system under norm 01 enables prudent players to effectively use reputation information, thereby reducing corruption and promoting cooperation.

4.4. Private reputation assessment norms that better promote prosocial behaviour than public assessment

Under a simple social norm that assigns good reputations to honest umpires and bad reputations to corrupt ones, the public reputation system is shown to outperform the baseline mechanism. However, the positive effects of our proposed reputation mechanism extend beyond this setting. In particular, the introduction of the private reputation system gives rise to a range of interesting and non-trivial behavioural outcomes.

As shown in figure 6, we present the evolutionary outcomes under all 16 possible social norms for the private reputation system. Notably, four of these norms, namely 0000, 0101, 1010 and 1111, are effectively equivalent to the public reputation assessment norms 00, 01, 10 and 11, respectively. This is because, under these private norms, cooperators and defectors do not differ in how they evaluate umpires. As a result, all players assess umpire reputations based solely on the umpire's strategy using a common criterion, which eliminates disagreement across players. This feature allows figure 6 to provide a direct comparison between the private and public reputation systems.

Compared to these four public assessment norms, some private assessment norms, such as 0001, 1001 and 1011, perform worse in curbing corruption and promoting cooperation. This is because, under these norms, cooperative players fail to assign bad reputations to corrupt umpires, whereas defectors are able to do so. As a result, this asymmetry prevents cooperators from effectively

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0576/6176689/rsif.2026.0576.pdf
by guest
on 11 November 2026

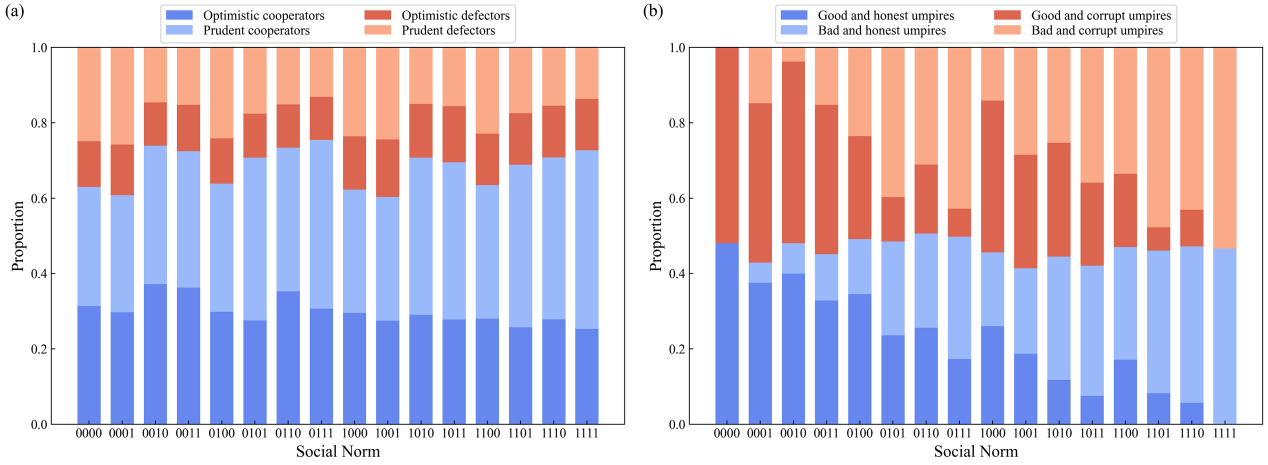


Figure 6. Evolutionary outcomes under different social norms for the private reputation system. (a) Average proportions of players adopting different strategies under various social norms. (b) Average proportions of umpires of different types under different social norms. In the private reputation system, each player forms its own evaluation of all umpires; therefore, the reported average proportions of umpire types are obtained by averaging across all players' assessments. These results are obtained from 20 independent runs, each averaged over the last 5000 MCSs of 50 000 MCSs, with $N = 500$ and $s = 0.5$.

avoiding corrupt umpires, leaving them vulnerable to exploitation. At the same time, defectors can benefit from interacting with corrupt umpires, thereby reinforcing corrupt behaviour.

By contrast, certain private norms outperform public ones in reducing corruption and fostering cooperation. A representative example is norm 0110. Under this norm, cooperators correctly assign good reputations to honest umpires and bad reputations to corrupt ones, while defectors show poor assessment, assigning bad reputations to honest umpires and good reputations to corrupt ones. The way defectors evaluate umpires reflects that they assess umpires not based on objective moral standards, but according to whether the umpire's behaviour aligns with their own interests. Consequently, defectors are more likely to be exposed to punishment from honest umpires, whereas cooperators benefit from favourable interactions. This leads to a reduction in corruption and promotes the evolution of cooperation.

These results further suggest that the design of social norms should be considered together with the adopted participation rule. Under the reputation-based participation rule in our model, norm 0101, which aligns reputation with the moral quality of umpire behaviour, enables both cooperators and defectors to make participation decisions that best serve their own interests. By contrast, norm 0110 leverages the asymmetry between cooperators and defectors in their participation decisions: it preserves reliable reputation assessment among cooperators while preventing defectors from consistently exploiting reputation information to avoid unfavourable interactions. Therefore, the advantage of some private assessment norms does not simply arise from incorporating additional information (i.e. the player's own strategy) into reputation assessment, but also from their ability to leverage the adopted participation rule.

5. Conclusion

We propose a novel evolutionary game-theoretic framework to investigate how reputation mechanisms can limit centralized corruption. By allowing players to access umpire reputation information and decide whether to participate in interactions accordingly, the proposed mechanism induces non-trivial coevolutionary dynamics between strategy and reputation. We establish both public and private reputation systems to investigate how reputation information shapes strategy evolution.

For public reputation systems, we develop a formal model inspired by the replicator–mutator dynamics to characterize the population dynamics. Combining results from the dynamical model with agent-based simulations, we demonstrate that under a public reputation system with a norm where honest umpires are assigned good reputations and corrupt ones are assigned bad reputations, the population achieves lower levels of corruption and higher cooperation rates compared with the costly strategy acquisition mechanism. Moreover, for private reputation systems, we show that even stronger performance can emerge under certain social norms, such as when cooperators assign reputations according to the moral quality of umpire behaviour while defectors evaluate umpires based on self-interest. These findings highlight the importance of carefully designing social norms within reputation systems.

It is worth noting that, while the existing literature on indirect reciprocity provides important insights and some of our norms share conceptual similarities with norms studied in traditional donation games, our norms still differ substantially from most norms designed for player reputation assessment, such as the well-known 'leading eight' norms [60]. This distinction stems from the fact that our model incorporates centralized punishment and focuses on the reputation assessment of third-party umpires. Therefore, our norms represent a class of reputation assessment norms tailored to third-party supervision.

Overall, our results demonstrate that reputation mechanisms provide a robust and practical pathway to enhancing transparency, thereby sustaining cooperation and limiting corruption. Building on our framework, promising directions for future

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0576/6176689/rsif.2026.0576.pdf
by guest
on 11 November 2026

research include exploring the evolution of social norms under both public and private reputation systems from a bottom-up perspective, extending the framework to consider quantitative or finer-grained reputation representations, and investigating the role of player reputation systems in mitigating centralized corruption.

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. All code is placed on the Zenodo platform [61].

Supplementary material is available online [62].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. Z.Y.: conceptualization, formal analysis, methodology, software, writing—original draft; C.S.: investigation, methodology, writing—original draft; C.L.: investigation, visualization, writing—review and editing; M.P.: investigation, methodology, supervision, validation; J.L.: conceptualization, methodology, project administration, supervision, validation; Z.W.: conceptualization, investigation, project administration, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This research was supported by the National Natural Science Foundation of China (Grant Nos. 62476221, U22B2036 and 62366058), the Yunnan Province XingDian Talent Support Program (Grant No. YNWR-QNBJ-2020-041), the Fundamental Research Funds for the Central Universities (Grant No. G2026KY06057), the Technological Innovation Team of Shaanxi Province (Grant No. 2025RS-CXTD-009), the International Cooperation Project of Shaanxi Province (Grant No. 2025GH-YBXM-017), and the XPLORER PRIZE 2021 of the Tencent Foundation. M.P. was supported by the Slovenian Research and Innovation Agency (Grant No. P1-0403).

References

1. Pennisi E. 2005 How did cooperative behavior evolve? *Science* **309**, 93–93. (doi:10.1126/science.309.5731.93)
2. Wang Z, Wang L, Szolnoki A, Perc M. 2015 Evolutionary games on multilayer networks: a colloquium. *Eur. Phys. J. B* **88**, 124. (doi:10.1140/epjb/e2015-60270-7)
3. Wang Z, Kokubo S, Jusup M, Tanimoto J. 2015 Universal scaling for the dilemma strength in evolutionary games. *Phys. Life Rev.* **14**, 1–30. (doi:10.1016/j.plrev.2015.04.033)
4. Wang Z *et al*. 2020 Communicating sentiment and outlook reverses inaction against collective risks. *Proc. Natl Acad. Sci. USA* **117**, 17 650–17 655. (doi:10.1073/pnas.1922345117)
5. Mu C, Guo H, Chen Y, Shen C, Hu D, Hu S, Wang Z. 2024 Multi-agent, human–agent and beyond: a survey on cooperation in social dilemmas. *Neurocomputing* **610**, 128514. (doi:10.1016/j.neucom.2024.128514)
6. Nowak MA. 2006 Five rules for the evolution of cooperation. *Science* **314**, 1560–1563. (doi:10.1126/science.1133755)
7. Wang Z, Jusup M, Wang RW, Shi L, Iwasa Y, Moreno Y, Kurths J. 2017 Onymity promotes cooperation in social dilemma experiments. *Sci. Adv.* **3**, e1601444. (doi:10.1126/sciadv.1601444)
8. Wang Z, Jusup M, Shi L, Lee JH, Iwasa Y, Boccaletti S. 2018 Exploiting a cognitive bias promotes cooperation in social dilemma experiments. *Nat. Commun.* **9**, 2954. (doi:10.1038/s41467-018-05259-5)
9. Schmid L, Chatterjee K, Hilbe C, Nowak MA. 2021 A unified framework of direct and indirect reciprocity. *Nat. Hum. Behav.* **5**, 1292–1302. (doi:10.1038/s41562-021-01114-8)
10. Chu C, Yuan Z, Hu S, Mu C, Wang Z. 2023 A pair-approximation method for modelling the dynamics of multi-agent stochastic games. In *Proc. 37th AAAI Conf. Artif. Intell., Washington, DC, 7–14 February*, vol. 37, pp. 5565–5572. Washington, DC: AAAI Press. (doi:10.1609/aaai.v37i5.25691)
11. Jia D, Romić I, Shi L, Su Q, Liu C, Liu J, Holme P, Li X, Wang Z. 2025 Social networking agency and prosociality are inextricably linked in economic games. *Nat. Hum. Behav.* **9**, 2620–2631. (doi:10.1038/s41562-025-02289-0)
12. Jia D, Dai X, Xing J, Tao P, Shi Y, Wang Z. 2025 Asymmetric interaction preference induces cooperation in human-agent hybrid game. *Sci. China Inf. Sci.* **68**, 212201. (doi:10.1007/s11432-023-4287-2)
13. Fehr E, Gächter S. 2000 Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**, 980–994. (doi:10.1257/aer.90.4.980)
14. Fehr E, Gächter S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140. (doi:10.1038/415137a)
15. Hauert C, Traulsen A, Brandt H, Nowak MA, Sigmund K. 2007 Via freedom to coercion: the emergence of costly punishment. *Science* **316**, 1905–1907. (doi:10.1126/science.1141588)
16. Li X, Jusup M, Wang Z, Li H, Shi L, Podobnik B, Stanley HE, Havlin S, Boccaletti S. 2018 Punishment diminishes the benefits of network reciprocity in social dilemma experiments. *Proc. Natl Acad. Sci. USA* **115**, 30–35. (doi:10.1073/pnas.1707505115)
17. Fehr E, Fischbacher U. 2004 Third-party punishment and social norms. *Evol. Hum. Behav.* **25**, 63–87. (doi:10.1016/s1090-5138(04)00005-4)
18. Sigmund K, De Silva H, Traulsen A, Hauert C. 2010 Social learning promotes institutions for governing the commons. *Nature* **466**, 861–863. (doi:10.1038/nature09203)
19. Baldassarri D, Grossman G. 2011 Centralized sanctioning and legitimate authority promote cooperation in humans. *Proc. Natl Acad. Sci. USA* **108**, 11 023–11 027. (doi:10.1073/pnas.1105456108)
20. Shen C, He Z, Shi L, Wang Z, Tanimoto J. 2024 Prosocial punishment bots breed social punishment in human players. *J. R. Soc. Interface* **21**, 20240019. (doi:10.1098/rsif.2024.0019)
21. Raihani NJ, Bshary R. 2019 Punishment: one tool, many uses. *Evol. Hum. Sci.* **1**, e12. (doi:10.1017/ehs.2019.12)
22. Morris-Martin A, De Vos M, Padget J. 2019 Norm emergence in multiagent systems: a viewpoint paper. *Auton. Agents Multi-Agent Syst.* **33**, 706–749. (doi:10.1007/s10458-019-09422-0)
23. Wang Z, Song Z, Shen C, Hu S. 2023 Emergence of punishment in social dilemma with environmental feedback. In *Proc. 37th AAAI Conf. Artif. Intell., Washington, DC, 7–14 February*, vol. 37, pp. 11 708–11 716. Washington, DC: AAAI Press. (doi:10.1609/aaai.v37i10.26383)
24. Robbins P. 2000 The rotten institution: corruption in natural resource management. *Polit. Geogr.* **19**, 423–443. (doi:10.1016/s0962-6298(99)00087-6)
25. Sundström A. 2015 Covenants with broken swords: corruption and law enforcement in governance of the commons. *Glob. Environ. Change* **31**, 253–262. (doi:10.1016/j.gloenvcha.2015.02.002)
26. Ambraseys N, Bilham R. 2011 Corruption kills. *Nature* **469**, 153–155. (doi:10.1038/469153a)
27. Weisel O, Shalvi S. 2015 The collaborative roots of corruption. *Proc. Natl Acad. Sci. USA* **112**, 10 651–10 656. (doi:10.1073/pnas.1423035112)

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0576/6176689/rsif.2026.0576.pdf
by guest
on 11 November 2026

28. Ostrom E. 2009 *Understanding institutional diversity*. Princeton, NJ: Princeton University Press.
29. Lee JH, Sigmund K, Dieckmann U, Iwasa Y. 2015 Games of corruption: how to suppress illegal logging. *J. Theor. Biol.* **367**, 1–13. (doi:10.1016/j.jtbi.2014.10.037)
30. Lee JH, Iwasa Y, Dieckmann U, Sigmund K. 2019 Social evolution leads to persistent corruption. *Proc. Natl Acad. Sci. USA* **116**, 13 276–13 281. (doi:10.1073/pnas.1900078116)
31. Muthukrishna M, Francois P, Pourahmadi S, Henrich J. 2017 Corrupting cooperation and how anti-corruption strategies may backfire. *Nat. Hum. Behav.* **1**, 0138. (doi:10.1038/s41562-017-0138)
32. Malagueño R, Albrecht C, Ainge C, Stephens N. 2010 Accounting and corruption: a cross-country analysis. *J. Money Laund. Control* **13**, 372–393. (doi:10.1108/13685201011083885)
33. Zhou X, Belloum A, Lees MH, van Engers T, de Laat C. 2023 The dynamics of corruption under an optional external supervision service. *Appl. Math. Comput.* **457**, 128172. (doi:10.1016/j.amc.2023.128172)
34. Manrique HM, Zeidler H, Roberts G, Barclay P, Walker M, Samu F, Fariña A, Bshary R, Raihani N. 2021 The psychological foundations of reputation-based cooperation. *Phil. Trans. R. Soc. B* **376**, 20200287. (doi:10.1098/rstb.2020.0287)
35. Murase Y, Hilbe C. 2024 Indirect reciprocity under opinion synchronization. *Proc. Natl Acad. Sci. USA* **121**, e2418364121. (doi:10.1073/pnas.2418364121)
36. Yue H, Xiong X, Feng M, Szolnoki A. 2025 Dynamic evolution of cooperation based on adaptive reputation threshold and game transition. *Chaos Solit. Fractals* **199**, 116693. (doi:10.1016/j.chaos.2025.116693)
37. He S, Li Q, Feng M, Szolnoki A. 2026 Reputation assimilation mechanism for sustaining cooperation. *Chaos Solit. Fractals* **202**, 117586. (doi:10.1016/j.chaos.2025.117586)
38. Sigmund K. 2010 *The calculus of selfishness*. Princeton, NJ: Princeton University Press.
39. Santos FP, Santos FC, Pacheco JM. 2016 Social norms of cooperation in small-scale societies. *PLoS Comput. Biol.* **12**, e1004709. (doi:10.1371/journal.pcbi.1004709)
40. Hilbe C, Šimsa Š, Chatterjee K, Nowak MA. 2018 Evolution of cooperation in stochastic games. *Nature* **559**, 246–249. (doi:10.1038/s41586-018-0277-x)
41. Su Q, McAvoy A, Wang L, Nowak MA. 2019 Evolutionary dynamics with game transitions. *Proc. Natl Acad. Sci. USA* **116**, 25 398–25 404. (doi:10.1073/pnas.1908936116)
42. Yuan Z, Jiang G, Hu S, Perc M, Chu C, Liu J. 2026 Dynamics of Q-learning in networked stochastic games. *IEEE Trans. Neural Netw. Learn. Syst.* **37**, 2854–2868. (doi:10.1109/tnnls.2025.3641365)
43. Li Q, Ling T, Feng M, Szolnoki A. 2026 Supervised tax compliance and evasion from a spatial evolutionary game perspective. *Humanit. Soc. Sci. Commun.* **13**, 447. (doi:10.1057/s41599-026-06802-2)
44. Radzvilavicius AL, Kessinger TA, Plotkin JB. 2021 Adherence to public institutions that foster cooperation. *Nat. Commun.* **12**, 3567. (doi:10.1038/s41467-021-23783-9)
45. Kessinger TA, Tarnita CE, Plotkin JB. 2023 Evolution of norms for judging social behavior. *Proc. Natl Acad. Sci. USA* **120**, e2219480120. (doi:10.1073/pnas.2219480120)
46. Xia C, Wang J, Perc M, Wang Z. 2023 Reputation and reciprocity. *Phys. Life Rev.* **46**, 8–45. (doi:10.1016/j.plrev.2023.05.002)
47. Kawakatsu M, Michel-Mata S, Kessinger TA, Tarnita CE, Plotkin JB. 2024 When do stereotypes undermine indirect reciprocity? *PLoS Comput. Biol.* **20**, e1011862. (doi:10.1371/journal.pcbi.1011862)
48. Brandt H, Sigmund K. 2005 Indirect reciprocity, image scoring, and moral hazard. *Proc. Natl Acad. Sci. USA* **102**, 2666–2670. (doi:10.1073/pnas.0407370102)
49. Feng M, Pi B, Deng LJ, Kurths J. 2023 An evolutionary game with the game transitions based on the Markov process. *IEEE Trans. Syst. Man Cybern. Syst.* **54**, 609–621. (doi:10.1109/TSMC.2023.3315963)
50. Feng K, Han S, Feng M, Szolnoki A. 2024 An evolutionary game with reputation-based imitation-mutation dynamics. *Appl. Math. Comput.* **472**, 128618. (doi:10.1016/j.amc.2024.128618)
51. Zhang G, Yao Y, Zeng Z, Feng M, Chica M. 2025 The evolution of cooperation in spatial public goods game with tolerant punishment based on reputation threshold. *Chaos* **35**, 013104. (doi:10.1063/5.0250120)
52. Zhang G, Xiong X, Pi B, Feng M, Perc M. 2025 Spatial public goods games with queueing and reputation. *Appl. Math. Comput.* **505**, 129533. (doi:10.1016/j.amc.2025.129533)
53. Okada I, Sasaki T, Nakai Y. 2018 A solution for private assessment in indirect reciprocity using solitary observation. *J. Theor. Biol.* **455**, 7–15. (doi:10.1016/j.jtbi.2018.06.018)
54. Hilbe C, Schmid L, Tkadlec J, Chatterjee K, Nowak MA. 2018 Indirect reciprocity with private, noisy, and incomplete information. *Proc. Natl Acad. Sci. USA* **115**, 12 241–12 246. (doi:10.1073/pnas.1810565115)
55. Schmid L, Ekbatani F, Hilbe C, Chatterjee K. 2023 Quantitative assessment can stabilize indirect reciprocity under imperfect information. *Nat. Commun.* **14**, 2086. (doi:10.1038/s41467-023-37817-x)
56. Shi J, Cao Z, Liu J, Chen C, Wang Z. 2025 A unified model of direct and indirect reciprocity in multichannel games. In *Proc. 39th AAAI Conf. Artif. Intell., Philadelphia, Pennsylvania, 25 February–4 March*, vol. 39, pp. 14 111–14 119. Washington, DC: AAAI Press. (doi:10.1609/aaai.v39i13.33545)
57. Page KM, Nowak MA. 2002 Unifying evolutionary dynamics. *J. Theor. Biol.* **219**, 93–98. (doi:10.1006/jtbi.2002.3112)
58. Blume LE. 1993 The statistical mechanics of strategic interaction. *Games Econ. Behav.* **5**, 387–424. (doi:10.1006/game.1993.1023)
59. Guo H, Shen C, Zou R, Tao P, Shi Y, Wang Z, Xing J. 2024 Complex pathways to cooperation emergent from asymmetry in heterogeneous populations. *Chaos* **34**, 023139. (doi:10.1063/5.0188177)
60. Ohtsuki H, Iwasa Y. 2006 The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.* **239**, 435–444. (doi:10.1016/j.jtbi.2005.08.008)
61. Yuan Z, Shen C, Liu C, Perc M, Liu J, Wang Z. 2026 Public and Private Reputation Systems for Limiting Centralized Corruption [Computer software]. Zenodo. (doi:10.5281/zenodo.21489795)
62. Yuan Z, Shen C, Liu C, Perc M, Liu J,Wang Z. 2026 Supplementary material from: Public and private reputation systems for limiting centralized corruption. Figshare. (doi:10.6084/m9.figshare.c.8709999)

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0576/6176689/rsif.2026.0576.pdf
by guest
on 11 November 2026