scientific reports

OPEN

Local stability of cooperation in a continuous model of indirect reciprocity

Sanghun Lee[1], Yohsuke Murase[2] & Seung Ki Baek[1]✉

Reputation is a powerful mechanism to enforce cooperation among unrelated individuals through indirect reciprocity, but it suffers from disagreement originating from private assessment, noise, and incomplete information. In this work, we investigate stability of cooperation in the donation game by regarding each player's reputation and behaviour as continuous variables. Through perturbative calculation, we derive a condition that a social norm should satisfy to give penalties to its close variants, provided that everyone initially cooperates with a good reputation, and this result is supported by numerical simulation. A crucial factor of the condition is whether a well-reputed player's donation to an ill-reputed co-player is appreciated by other members of the society, and the condition can be reduced to a threshold for the benefit-cost ratio of cooperation which depends on the reputational sensitivity to a donor's behaviour as well as on the behavioural sensitivity to a recipient's reputation. Our continuum formulation suggests how indirect reciprocity can work beyond the dichotomy between good and bad even in the presence of inhomogeneity, noise, and incomplete information.

Reputation was an absolutely essential asset in trade of the illiterate in the premodern era[1], and it still plays a crucial role in markets and communities, making reputation management a central part of marketing and public relations. Also in a variety of social contexts starting from early childhood, we evaluate others based on third-party interactions[2] and adjust our own behaviour to earn good reputations from others[3]. In this regard, although some studies suggest the existence of social evaluation in species other than humans[4], *Homo sapiens* seems to have unique capability to use information of other social members through rumour and gossip.

Evolutionary biologists argue that the ability of social evaluation helps us extend the range of cooperation beyond kinship by encouraging cooperators and punishing defectors in a social dilemma[5–11]. A classical example of a social dilemma is the donation game, in which a player's cooperation benefits his or her co-player by an amount of b at the cost of c, where $0 < c < b$. The following payoff matrix defines the game:

$$\left(\begin{array}{c|cc} & C & D \\ \hline C & b-c & -c \\ D & b & 0 \end{array}\right), \qquad (1)$$

where we abbreviate cooperation and defection as C and D, respectively. As is clearly seen from this payoff matrix, choosing D is the rational choice for each player whereas mutual cooperation is better for both, hence a dilemma. The players can escape from mutual defection by the action of reciprocity if the game is repeated[12–19], but the price is that they have to remember the past and repeat interaction with sufficiently high probability, which is sometimes unfeasible. The basic idea of indirect reciprocity is that even a single encounter between two persons can be enough if that experience is reliably transferred in the form of reputation to those who will interact with these players in future. In other words, the problem is how to store, transmit, and retrieve information on each others's past behaviour in a distributed manner[9,20]. Experiments show that the notion of indirect reciprocity provides a useful explanation for cooperative human behaviour[21,22].

For this mechanism to work, we need two rules as a social norm: One is an assessment rule to assign reputation to a player based on his or her action to another player. The other is a behavioural rule to prescribe an action between C and D, when players' reputations are given. An early idea was a norm called Image Scoring, which judges the donor's C and D as good and bad, respectively[6]. According to this norm, cooperation can thrive when

[1]Department of Physics, Pukyong National University, Busan 48513, Korea. [2]RIKEN Center for Computational Science, Kobe, Hyogo 650-0047, Japan. ✉email: seungki@pknu.ac.kr

Rule	α_{1C1}	α_{1D1}	α_{1C0}	α_{1D0}	α_{0C1}	α_{0D1}	α_{0C0}	α_{0D0}	β_{11}	β_{10}	β_{01}	β_{00}
L1	1	0	1	1	1	0	1	0	*C*	*D*	*C*	*C*
L2 (Consistent Standing)	1	0	0	1	1	0	1	0	*C*	*D*	*C*	*C*
L3 (Simple Standing)	1	0	1	1	1	0	1	1	*C*	*D*	*C*	*D*
L4	1	0	1	1	1	0	0	1	*C*	*D*	*C*	*D*
L5	1	0	0	1	1	0	1	1	*C*	*D*	*C*	*D*
L6 (Stern Judging)	1	0	0	1	1	0	0	1	*C*	*D*	*C*	*D*
L7 (Staying)	1	0	1	1	1	0	0	0	*C*	*D*	*C*	*D*
L8 (Judging)	1	0	0	1	1	0	0	0	*C*	*D*	*C*	*D*

Table 1. Leading eight. Cooperation and defection are denoted as *C* and *D*, respectively, and a player's reputation is either good (1) or bad (0). By α_{uXv}, we mean the reputation assigned to a player who did $X \in \{C, D\}$ with reputation u to another player with reputation v. The behavioural rule β_{uv} prescribes what a player should do between *C* and *D* when he or she has reputation u and the co-player has reputation v. We note that L1 has been known as Contrite Tit-for-Tat in the context of direct reciprocity[27–30].

$$q > c/b, \tag{2}$$

where q means the probability of knowing someone's reputation[23]. On the one hand, this condition seems natural because it parallels Hamilton's rule for kin selection, and the only difference is that q has replaced genetic relatedness. On the other hand, if one asks what is an essential prerequisite for a norm to promote cooperation, it is not answered by Eq. (2), and we need a broader perspective on the structure of social norms.

According to Kandori's formalism[24], Image Scoring is an example of 'first-order' assessment rules because its judgment depends only on the donor's action. A 'second-order' assessment rule takes the recipient's reputation into account, and a 'third-order' assessment rule additionally refers to the donor's reputation. The number of possible third-order rules thus amounts to $2^{2^3} = 256$. On the other hand, the number of actions rules is $2^{2^2} = 16$ because a behavioural rule prescribes an action depending on the reputations of the donor and recipient. Among the $2^{2^3+2^2} = 4096$ combinations, we have the *leading eight*[25,26], the eight pairs of an assessment rule α and a behavioural rule β that make cooperative equilibrium evolutionarily stable against every mutant with $\beta' \neq \beta$ (Table 1).

The situation becomes complicated when reputations are not globally shared in the population: Misjudgement does occur in the presence of error, and some players may even have their own private rules of assessment[31–34]. Then, strict social norms such as 'Judging' and 'Stern Judging' completely fail to tell if other players are good or bad, although they successfully induce cooperation when reputation is always public information[35,36]. Communication rounds can be introduced to resolve disagreements[10], or one may need empathy or prudence in judgment to alleviate the problem[37,38], but these remedies imply the intrinsic instability of the reputation mechanism in its pure sense. We also point out that most of the existing models are based on an assumption that the dynamic variables are binary, although reputation is not really a simple dichotomy between good and bad, and some actions cannot be classified as either cooperation or defection[39,40].

In this work, we thus wish to investigate indirect reciprocity by taking reputations and actions as continuous variables. By doing so, we can naturally deal with the continuous dynamics between the existing norm and its close variants by means of analytic tools. We also expect that this formulation can be used to address the problems of error and incompleteness: The idea is that perception error will effectively replace a binary reputation by a probabilistic mixture between good and bad, just as a binary action can be replaced by a probabilistic mixture of cooperation and defection in the presence of implementation error. Although the number of possible social norms expands to infinity, we will restrict ourselves to local-stability analysis by assuming that mutants appear from a small neighbourhood of the existing social norm.

Analysis

Let us imagine a large population and denote the number of players as N. The basic setting is that a random pair of players are picked up to play the donation game [Eq. (1)]. In our model, the player chosen as a donor decides the degree of cooperation to the co-player between zero and one, which mean full defection and full cooperation, respectively, based on their reputations. Let m_{ij} denote player j's reputation from the viewpoint of player i. The player i also has a behavioural rule $\beta_i(m_{ii}, m_{ij})$, which determines how much he or she will do as a donor to j. Note that all of m_{ij}, α_i, and β_i for any i and j take real values inside the unit interval. Player k is observing the interaction between i and j, and it has its own assessment rule $\alpha_k(m_{ki}, \beta_i, m_{kj})$. With observation probability $q > 0$, the reputation that k assigns to i will be updated on average as follows:

$$m_{ki}^{t+1} = (1-q)m_{ki}^{t} + \frac{q}{N-1}\sum_{j \neq i} \alpha_k \left[m_{ki}^{t}, \beta_i \left(m_{ii}^{t}, m_{ij}^{t} \right), m_{kj}^{t} \right] \tag{3}$$

where the superscripts have been used as time indices. Equation (3) is to be analysed in this section. Before proceeding, let us note two points: First, as a deterministic equation, Eq. (3) does not include error explicitly. If the probability of error is low, Eq. (3) will nevertheless describe the dynamics for most of the time, and the main effect of error will be to perturb the output of α or β by a small amount at a point in time, say, $t = 0$. Second, from a mathematical point of view, it is preferable to treat both diagonal and off-diagonal elements on an equal footing as in Eq. (3), which implies that one has to observe even the self-reputation m_{ii} probabilistically. If that sounds unrealistic, we may alternatively assume that donors and recipients update their self-reputations with probability one. However, it is a reasonable guess that the difference between these two settings becomes marginal when N is large enough, and this guess is indeed verified by numerical calculation (not shown).

Throughout this work, α and β are assumed to be C^2-differentiable. In addition, we will focus on the cases where the system has a fixed point characterized by

$$\alpha(1,1,1) = 1 \tag{4a}$$

$$\beta(1,1) = 1 \tag{4b}$$

because otherwise the norm would not sustain cooperation among well-reputed players from the start. As concrete examples of α and β, let us extend the leading eight to deal with continuous variables by applying the trilinear (bilinear) interpolation to α (β) in Table 1. If we consider L3 (Simple Standing), for instance, it is described by

$$\alpha_{\mathrm{SS}}(x,y,z) = yz - z + 1 \tag{5a}$$

$$\beta_{\mathrm{SS}}(x,y) = y. \tag{5b}$$

If we define $A_\xi \equiv \partial\alpha/\partial\xi|_{(1,1,1)}$ and $B_\lambda \equiv \partial\beta/\partial\lambda|_{(1,1)}$ with $\xi \in \{x,y,z\}$ and $\lambda \in \{x,y\}$, all the leading eight have $A_y = B_y = 1$, together with $A_x = B_x = 0$, and these are related with the basic properties of the leading eight to be nice, retaliatory, apologetic, and forgiving[26].

Below, we will examine two aspects of stability: The first is recovery of full cooperation from disagreement in a homogeneous population where everyone uses the same α and β[36]. Starting from $m_{ij} = 1$ for every i and j, the dynamics of Eq. (3) will be investigated within the framework of linear-stability analysis. The second aspect is the stability against mutant norms, for which we have to check the long-term payoff difference between the resident and mutant norms in a stationary state. We again start this analysis from a nearly homogeneous population in which only one individual considers using a slightly different norm. Although private assignment of reputation is allowed, the point is that it will remain unrealised if no one has a reason to deviate from the prevailing norm, considering that such deviation will only decrease his or her payoff. In this sense, the homogeneity serves as a self-consistent assumption in the second part of the stability analysis.

Recovery from disagreement.

To understand the time evolution of disagreement in a homogeneous population with common α and β, let us rewrite Eq. (3):

$$m_{ki}^{t+1} = (1-q)m_{ki}^t + \frac{q}{N-1}\sum_{j\neq i}\alpha\left[m_{ki}^t, \beta\left(m_{ii}^t, m_{ij}^t\right), m_{kj}^t\right], \tag{6}$$

where $\alpha_k = \alpha$ and and $\beta_i = \beta$ in this homogeneous population. Initially, everyone starts with a good reputation, which can be perturbed by error. To see whether the magnitude of the perturbation grows with time, we set $m_{ki}^t \equiv 1 - \varepsilon_{ki}^t$ and expand the above equation to the first order of ε as follows:

$$1-\varepsilon_{ki}^{t+1} =(1-q)\left(1-\varepsilon_{ki}^t\right) + \frac{q}{N-1}\sum_{j\neq i}\alpha\left[1-\varepsilon_{ki}^t, \beta\left(1-\varepsilon_{ii}^t, 1-\varepsilon_{ij}^t\right), 1-\varepsilon_{kj}^t\right] \tag{7}$$

$$\approx(1-q)\left(1-\varepsilon_{ki}^t\right) + \frac{q}{N-1}\sum_{j\neq i}\alpha\left[1-\varepsilon_{ki}^t, 1-\left(B_x\varepsilon_{ii}^t + B_y\varepsilon_{ij}^t\right), 1-\varepsilon_{kj}^t\right] \tag{8}$$

$$\approx(1-q)\left(1-\varepsilon_{ki}^t\right) + \frac{q}{N-1}\sum_{j\neq i}\left\{1-\left[A_x\varepsilon_{ki}^t + A_y\left(B_x\varepsilon_{ii}^t + B_y\varepsilon_{ij}^t\right) + A_z\varepsilon_{kj}^t\right]\right\}, \tag{9}$$

or, equivalently,

$$\varepsilon_{ki}^{t+1} \approx(1-q)\varepsilon_{ki}^t + \frac{q}{N-1}\sum_{j\neq i}[A_x\varepsilon_{ki}^t + A_y(B_x\varepsilon_{ii}^t + B_y\varepsilon_{ij}^t) + A_z\varepsilon_{kj}^t] \tag{10}$$

$$=(1-q+qA_x)\varepsilon_{ki}^t + qA_yB_x\varepsilon_{ii}^t + \frac{q}{N-1}\sum_{j\neq i}[A_yB_y\varepsilon_{ij}^t + A_z\varepsilon_{kj}^t], \tag{11}$$

which leads to

$$\frac{d}{dt}\varepsilon_{ki} \approx -q(1-A_x)\varepsilon_{ki} + qA_yB_x\varepsilon_{ii} + \frac{q}{N-1}\sum_{j\neq i}[A_yB_y\varepsilon_{ij} + A_z\varepsilon_{kj}], \tag{12}$$

if time is regarded as a continuous variable. This is a linear-algebraic system with an $N^2 \times N^2$ matrix. In principle, we can find the stability at the origin as well as the speed of convergence toward it by calculating the eigenvalues. By attempting this calculation from $N = 2$ to 5 with a symbolic-algebra system[41], we see the following pattern in the eigenvalue structure:

$$\Lambda_1^{(N^2-2N+1)} = q\left(-1 + A_x - \frac{1}{N-1}A_z\right) \tag{13}$$

$$\Lambda_2^{(N-1)} = q(-1 + A_x + A_z) \tag{14}$$

$$\Lambda_3^{(N-1)} = q\left(-1 + A_x - \frac{1}{N-1}A_z + A_yB_x - \frac{1}{N-1}A_yB_y\right) \tag{15}$$

$$\Lambda_4^{(1)} = q(-1 + A_x + A_z + A_yB_x + A_yB_y), \tag{16}$$

where each superscript on the left-hand side means multiplicity of the corresponding eigenvalue. Based on this observation, we conjecture that this pattern is valid for general N. A sufficient condition for recovery to take place in this first-order calculation is that the largest eigenvalue is negative. The largest eigenvalue is the last one, $\Lambda_4^{(1)}$, because all the derivatives are non-negative. In other words, the first-order perturbation analysis gives a sufficient condition for local recovery as

$$Q \equiv -1 + A_x + A_z + A_y(B_x + B_y) < 0. \tag{17}$$

Suppression of mutants. To analyse the effect of a mutant norm, we will look at the long-time behaviour in Eq. (3). That is, for given sets of rules $\{\alpha_i\}$ and $\{\beta_i\}$, we assume that the image matrix $\{m_{ij}\}$ will converge to a stationary state as $t \to \infty$, satisfying

$$m_{ki} = \frac{1}{N-1}\sum_{j\neq i}\alpha_k\left[m_{ki}, \beta_i(m_{ii}, m_{ij}), m_{kj}\right]. \tag{18}$$

Note that q only affects the speed of convergence to stationarity: It is an irrelevant parameter as far as we work with a stationary state, which is in contrast with Eq. (2), where q appears as an essential condition for indirect reciprocity. In the donation game with benefit b and cost c [Eq. (1)], player j's expected payoff can be computed as

$$\pi_j = \frac{1}{N-1}\left[b\sum_{i\neq j}\beta_i(m_{ii}, m_{ij}) - c\sum_{i\neq j}\beta_j(m_{jj}, m_{ji})\right]. \tag{19}$$

For the sake of simplicity, let us assume that every person with index 1 to $N-1$ has the same rules and equal reputation, so that player $i = 1$ is representative for all of them in the resident population. Now, the situation is effectively reduced to a two-body problem between players 0 and 1. By assumption, the system initially starts from a fully cooperative state where everyone has good reputation, i.e., $m_{11} = \beta(1,1) = \alpha(1,1,1) = 1$. The rules used by the resident population will be denoted by $\alpha \equiv \alpha_1$ and $\beta \equiv \beta_1$ without the subscripts. Now, the focal player 0 attempts a slightly different norm, defined by $\alpha_0(x,y,z) = \alpha(x,y,z) - \delta(x,y,z)$ and $\beta_0(x,y) = \beta(x,y) - \eta(x,y)$ with $|\delta| \ll 1$ and $|\eta| \ll 1$. Let us assume that the introduction of δ and η causes small changes in the image matrix: Only the elements related to the focal player will be affected because the residents can still give $m_{11} = 1$ to each other when the mutant occupies a negligible fraction of the population, i.e., $N \gg 1$. Therefore, if mutation leads to $m_{00} = 1 - \varepsilon_{00}$, $m_{01} = 1 - \varepsilon_{01}$, and $m_{10} = 1 - \varepsilon_{10}$ with $\varepsilon_{ij} \ll 1$, by expanding Eq. (18) to the linear order of perturbation (see Methods), we obtain

$$\varepsilon_{00} = \frac{(1 - A_x + A_yB_y)\delta_1 + (1 - A_x - A_z)A_y\eta_1}{(1 - A_x - A_z)(1 - A_x - A_yB_x)} \tag{20}$$

$$\varepsilon_{01} = \frac{\delta_1}{1 - A_x - A_z} \tag{21}$$

$$\varepsilon_{10} = \frac{(B_x + B_y)\delta_1 + (1 - A_x - A_z)\eta_1}{(1 - A_x - A_z)(1 - A_x - A_yB_x)}A_y, \tag{22}$$

where $\delta_1 \equiv \delta(1,1,1) \geq 0$ and $\eta_1 \equiv \eta(1,1) \geq 0$, provided that

$$A_x + A_z < 1 \tag{23}$$

$$A_x + A_yB_x < 1. \tag{24}$$

We can now calculate the focal player 0's payoff as follows:

$$\pi_0 = \frac{1}{N-1}\left[b\sum_{i\neq 0}\beta_i(m_{ii}, m_{i0}) - c\sum_{i\neq 0}\beta_0(m_{00}, m_{0i}) \right] \tag{25}$$

$$= b\beta(m_{11}, m_{10}) - c\beta_0(m_{00}, m_{01}) \tag{26}$$

$$\approx b\big(1 - B_y\varepsilon_{10}\big) - c\big(1 - B_x\varepsilon_{00} - B_y\varepsilon_{01} - \eta_1\big). \tag{27}$$

If we plug Eqs. (20), (21), and (22) here, the payoff change $\Delta\pi_0 \equiv \pi_0 - (b-c)$ is given as

$$\Delta\pi_0 = -\frac{bA_yB_y - c(1-A_x)}{1 - A_x - A_yB_x}\left[\left(\frac{B_x + B_y}{1 - A_x - A_z}\right)\delta_1 + \eta_1\right], \tag{28}$$

and we require this quantity to be negative for any small positive δ_1 and η_1. Here, it is worth stressing that the signs of δ_1 and η_1 are determined because we start from a fully cooperative state with $m_{ij} = 1$: For other states where δ and η can take either sign, the first-order terms should vanish so that the second-order terms can determine the sign of $\Delta\pi_0$. In this respect, the payoff analysis is greatly simplified by choosing the specific initial state. Because of Eqs. (23) and (24), the negativity of Eq. (28) is reduced to the following inequality:

$$\frac{b}{c} > \frac{1 - A_x}{A_yB_y}, \tag{29}$$

which, together with Eqs. (23) and (24), characterizes a condition for a social norm to stabilize cooperation against local mutants, as an alternative to Eq. (2). This result is intuitively plausible because cooperation will be unstable if one does not lose reputation by decreasing the degree of cooperation (i.e., $A_y \approx 0$) or if no punishment is imposed on an ill-reputed player (i.e., $B_y \approx 0$).

Two remarks are in order: First, whether mutation occurs to a single individual or to a fraction of the population does not alter the final result in this first-order calculation. Suppose that the population is divided into two groups with fractions p and $1-p$, respectively. One group has α and β, and the other group has $\alpha + \delta$ and $\beta + \eta$. Then, the payoff difference between two players, each from a different group, is still the same as Eq. (28) (see Methods). Therefore, if an advantageous mutation occurs with $p \ll 1$, the mutants are always better off than the resident until they take over the whole population, i.e., $p \to 1$. In this sense, our condition determines not only the initial invasion but also the fixation of a mutant norm, as long as it is a close variant of the resident one. Second, one could ask what happens if a mutant differs only in the slopes while keeping $\delta_1 = \eta_1 = 0$. Equation (28) does not answer this question because it is based on an assumption that the $\partial\delta/\partial\xi|_{(1,1,1)}\varepsilon_{ij}$ and $\partial\eta/\partial\lambda|_{(1,1)}\varepsilon_{ij}$, where $\xi \in \{x, y, z\}$ and $\lambda \in \{x, y\}$, are all negligibly small in the first-order calculation. However, even if the derivatives are taken into consideration, we find that δ_1 or η_1 must still be positive to make a finite payoff change. In other words, the basic form of Eq. (28) is still useful, although the coefficients include correction terms. The performance of such a 'slope mutant' will be checked numerically at the end of the next section.

Results

In this section, we will numerically check the continuous-reputation system in the presence of inhomogeneity, noise, and incomplete information. More specifically, the simulation code should allow each player i to adopt a different set of α_i and β_i to simulate an inhomogeneous population. The outputs of α_i and β_i can be affected by random-number generation to simulate a noisy environment where misperception and misimplementation occur, and every interaction between a pair of players will update only some part of the reputation system, parametrized by the observation probability q, because information is incomplete.

Our numerical simulation code is based on a publicly available one[36] but has been modified to handle continuous variables. To simulate the dynamics of a society of N players, we work with an $N \times N$ image matrix $\{m_{ij}\}$ whose elements are all set to be ones at the beginning. Every player starts with zero payoff, i.e., $\pi_i = 0$ initially. In each round, we randomly pick up two players, say, i and j, so that i is the donor and j is the recipient of the donation game [Eq. (1)], which has $b = 2$ and $c = 1$ unless otherwise noted. Each other member of the population, say, k, independently observes the interaction with probability q and updates m_{ki} according to his or her own assessment rule α_k. Although the above analyses are generally applicable to any norms defined by α and β as long as Eq. (4) is true, we would like to focus on Simple Standing as a representative example of successful norms. Misperception may occur with probability e, whereby m_{ki} becomes a random number drawn from the unit interval. Implementation error is also simulated in a similar way by setting the output of β to a random number between zero (defection) and one (cooperation) with probability γ. This process is repeated for M rounds, during which every player's payoff is accumulated. Equation (18) suggests that q will only affect the convergence rate toward a stationary state. For this reason, we will fix this parameter at $q = 0.4$ throughout the simulation unless otherwise mentioned. Note also that we have deliberately made this parameter low enough to violate the inequality in Eq. (2).

To see the effect of Q on recovery [Eq. (17)], we have tested three norms one by one in a homogeneous population with $e = \gamma = 0$ (Fig. 1). All these norms have $\alpha(1,1,1) = 1$ and $\beta(1,1) = 1$ in common but their local slopes are different to make Q positive, zero, or negative. The first norm under consideration has $(A_x, A_y, A_z) = (0.2, 0.9, 0.1)$ and $(B_x, B_y) = (0.2, 0.8)$, which together make $Q > 0$. If some members of the population initially have slightly imperfect reputations, they fail to recover under such a norm. If $Q < 0$, on the other hand, the recovery process indeed takes place with a finite time scale. Although Simple Standing violates

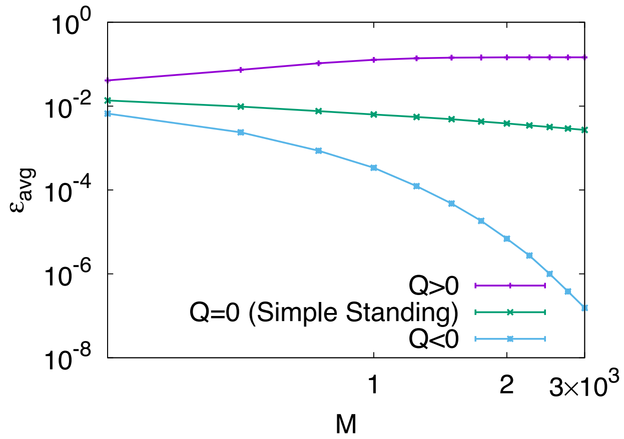

Figure 1. Recovery from disagreement when M rounds have elapsed in a population of size $N = 50$ with common α and β. Initially, we randomly pick up 20% of the image-matrix elements and change them to 0.9, whereas the rest of them remain as 1's, and the simulation has been repeated over 10^3 independent samples without error, i.e., $e = \gamma = 0$. In this log-log plot, the vertical axis shows the average difference from the state of perfect reputation, represented by the average of $\varepsilon_{ki} \equiv 1 - m_{ki}$. We have tested three norms, which all have $\alpha(1,1,1) = 1$ and $\beta(1,1) = 1$ but differ in their local slopes there: The first norm has $(A_x, A_y, A_z) = (0.2, 0.9, 0.1)$ and $(B_x, B_y) = (0.2, 0.8)$, which together yield $Q \equiv -1 + A_x + A_z + A_y(B_x + B_y) > 0$ [Eq. (17)]. The next one is Simple Standing with $(A_x, A_y, A_z) = (1, 0, 1)$ and $(B_x, B_y) = (0, 1)$, which has $Q = 0$. The last one for $Q < 0$ is a variant of Simple Standing with $(A_x, A_y, A_z) = (0, 0.9, 0)$ and $(B_x, B_y) = (0, 0.9)$.

Eq. (17) by having $Q = 0$, our simulation shows that it gets reputation recovered with the aid of higher-order terms, and it is a slow process with a diverging time scale. Among the leading eight, L1, L3 (Simple Standing), L4, and L7 (Staying) fall into this category of $Q = 0$, whereas the other four, i.e., L2 (Consistent Standing), L5, L6 (Stern Judging), and L8 (Judging), have positive Q. The difference between these two groups is whether $A_z = \alpha_{1C1} - \alpha_{1C0} = 1 - \alpha_{1C0}$ is zero or one: If a well-reputed player has to risk his or her own reputation in helping an ill-reputed co-player, i.e., $\alpha_{1C0} = 0$, it means $A_z = 1$ and $Q > 0$, so we can conclude that the initial state of $m_{ki} \approx 1$ will not be recovered. According to an earlier study on the leading eight[36], the latter four with $Q > 0$ have long recovery time from a single disagreement in reputation. Although it is not derived from a continuum formulation, the result is qualitatively consistent with ours.

As for the effect of mutation in assessment rules, let us consider the following scenario: One half of the population have adopted Simple Standing [Eq. (5)], whereas the other half are "mutants" using a different assessment rule $\alpha_{SS} - \delta$ with

$$\delta(x, y, z) = \delta_1(2yz - 2z + 1), \tag{30}$$

where δ_1 is a small number, say, $\delta_1 = 0.02$ in numerical calculation. Such a half-and-half configuration is being used because the payoff difference [Eq. (28)] is unaffected by the fraction of mutants, p (see Methods). Figure 2a shows that the level of cooperation is still high if $e \ll 1$, and the cooperation rate of Simple Standing in the continuous form converges to 100% in a monomorphic population (not shown). Furthermore, we see that mutants are worse off than the players of Simple Standing, i.e., $\pi_0 < \pi_1$, as expected.

From a theoretical viewpoint, an important question is how quickly the mutants' payoff difference $\Delta\pi_0 \equiv \pi_0 - \pi_1$ becomes negative: Although we have argued that the inequality will be true for Simple Standing, the calculation is based on several assumptions. In particular, one could say that Eq. (3) corresponds to $M \propto N^2$ because it seems to assume that everyone meets every other player with a weighting factor of $1/(N-1)$. If $M \propto N^2$, however, it would pose a serious obstacle to applying such a norm to the society where the number of interactions will grow linearly with N. Fortunately, the inset of Fig. 2a shows that $M \propto N$ indeed suffices to make $\Delta\pi_0$ negative. One could also point out that the payoff difference should be $\Delta\pi_0 = -\delta_1$ according to Eq. (28), whereas the result in Fig. 2a has smaller magnitude. A part of the reason is that Eq. (28) does not take perception error into account, so the numerical value recovers the predicted order of magnitude as $e \to 0$. In addition, Eq. (28) is based on a first-order approximation, and a higher-order calculation reproduces the observed value with greater precision (see Methods).

An important prediction of our analysis is the threshold of b/c to make a local mutant worse off than the resident population [Eq. (29)]. In Fig. 2b, we directly check Eq. (29) by measuring payoffs in equilibrium in a population of size $N = 50$. A variant of Simple Standing is chosen as the resident norm, which occupies $p = 0.5$ of the population with $\alpha(1,1,1) = \beta(1,1) = 1$ and $A_x = A_z = B_x = 0$. The only difference from Simple Standing is that $A_y = B_y = 0.9$, and the reason of this variation is that the first-order perturbation for the leading eight develops spurious singularity when p is finite (see Methods). When perception is free from error, i.e., $e = 0$, the results do not depend on the observation probability q, as expected from stationarity [Eq. (18)], and the

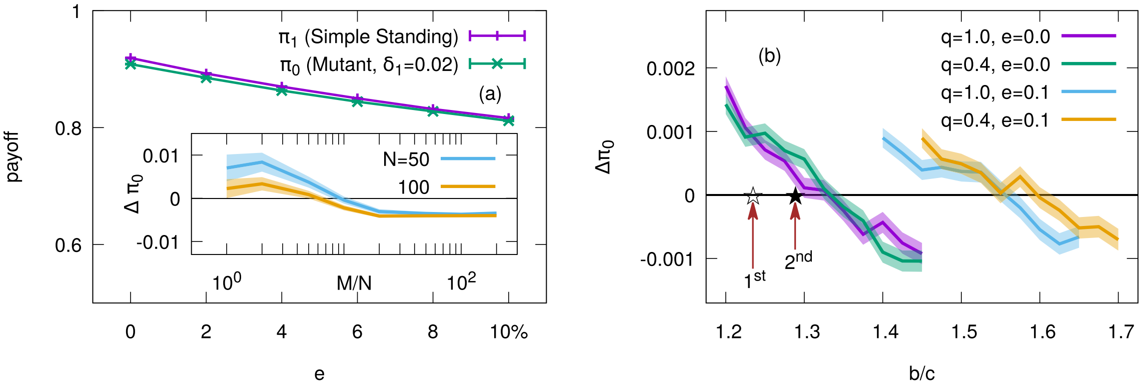

Figure 2. Stationary states of a population with $N = 50$ players, reached from an initial condition with $m_{ij} = 1$ for every i and j. In each case, the mutant norm differs from the resident one by $\delta_1 = 0.02$ and occupies one half of the population ($p = 0.5$). The game is defined by Eq. (1) with $b = 2$ and $c = 1$. (**a**) Average payoffs over 5×10^4 samples when the resident norm is Simple Standing. Everyone can observe each interaction with probability $q = 0.4$, and perception error and implementation error occur with probabilities $e = 0.1$ and $\gamma = 0.1$, respectively. Inset: Convergence of payoff difference $\Delta\pi_0 \equiv \pi_0 - \pi_1$ as M increases. If $M \propto gN$ with a sufficiently large constant $g \gtrsim O(10)$, the mutants will obtain less payoffs than Simple Standing, making $\Delta\pi_0 < 0$. This result has no significant dependence on N. (**b**) Payoff advantage of mutants with respect to the resident as a function of b/c, averaged over 5×10^4 samples per each, when $M = 10^4$. The resident norm, a variant of Simple Standing, has $\alpha(1,1,1) = \beta(1,1) = 1$ and $A_x = A_z = B_x = 0$ but $A_y = B_y = 0.9$ as in Fig. 1. Implementation error occurs with probability $\gamma = 0.1$, and the results are qualitatively the same for any small γ. The stars on the horizontal line indicate the predicted threshold values obtained from the first-order and second-order calculations, respectively. In both of these panels, the shaded areas represent error bars.

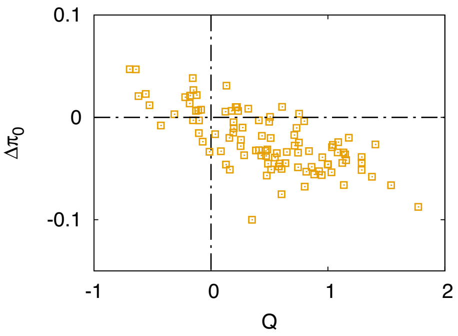

Figure 3. Payoff difference between the resident population using Simple Standing and its 'slope mutant', which has the same β and $\alpha(1,1,1) = 1$ but different slopes A_x and A_y and A_z. Each point denotes a randomly generated mutant through the trilinear interpolation among $\alpha(1,1,1) = 1$ and seven random values $\alpha(0,0,0), \alpha(0,0,1), \ldots, \alpha(1,1,0)$ within the unit interval. The mutant norm occupies 10% of the whole population whose size is $N = 100$. The horizontal axis shows the mutant's Q-value [Eq. (17)], and the vertical axis means its payoff difference $\Delta\pi_0$ with respect to the resident norm after a sufficiently long time, e.g., $M/N \sim O(10^3)$. As before, the game is defined with $b = 2$ and $c = 1$, and the observation probability is $q = 0.4$. Perception error and implementation error occur with probabilities $e = 0.1$ and $\gamma = 0.1$, respectively.

threshold value is consistent with the first- and second-order calculations [the arrows in Fig. 2b]. When $e > 0$, on the other hand, the threshold is pushed upward, implying that cooperation becomes harder to stabilize because of the perception error. In addition, we now see that incomplete information with $q < 1$ can shift the threshold further with the aid of positive e. We have also changed the value of γ, but it does not not change the average behaviour in the above results. Overall, the point of Fig. 2b is that our analysis does capture the correct picture.

Finally, we can numerically check the effect of a 'slope mutant', which has $\alpha(1,1,1) = 1$ as a fixed point and the same behavioural rule as Simple Standing but differs in the slopes A_x, A_y and A_z. To be more specific, let us assume that a mutant norm occupies 10% of the population whereas the rest of them are using Simple Standing. The values of $\alpha(x,y,z)$ at the vertices of the three-dimensional unit hypercube are randomly drawn from the unit interval, except for $\alpha(1,1,1) = 1$. Then, the trilinear interpolation is used to construct the continuous assessment rule. According to our simulation (Fig. 3), the performance of the mutant norm is strongly correlated with its Q-value [Eq. (17)]. Recall that the expression of Q has been derived in the context of recovery from small disagreement in a homogeneous population. Figure 3 nevertheless suggests that it can also serve as a useful indicator to tell if a minority of 'slope mutants' will be competitive with the resident norm, even when the difference between their assessment rules is not necessarily small.

Summary and discussion

In summary, we have studied indirect reciprocity with private, noisy, and incomplete information by extending the binary variables for reputation and behaviour to continuous ones. The extension to continuum is an idealization because it would impose an excessive cognitive burden to keep track of others' reputations without discretization; nonetheless, this abstraction allows us to overcome the fact that the sharp dichotomy between good and bad is often found insufficient in reporting an assessment[42–44]. In particular, this formulation makes it possible to check the role of sensitivity to new information in judging others and adjusting our own behaviour. That is, according to Eq. (29), the benefit-cost ratio of cooperation should increase for stabilizing the cooperative initial state, if reputation is insensitive to observed behaviour (low A_y) or if the level of cooperation is insensitive to the recipient's reputation (low B_y). At the same time, in contrast to the well-known condition for indirect reciprocity akin to Hamilton's rule [Eq. (2)], we have observed that incompleteness of information, controlled by $q < 1$, mainly affects the convergence toward a stationary state without altering the overall conclusion. This approach sheds light on difference among the leading eight in their recovery speeds from a single disagreement. Our analysis has identified the key factor α_{1C0} in Table 1, i.e., how to assign reputation to a well-reputed donor who chooses C against an ill-reputed recipient: If this choice is regarded as good according to $\alpha_{1C0} = 1$, making the assessment function $\alpha(x, y, z)$ insensitive to z, the recovery can take place smoothly. As a result, we conclude that α should respond to the donor's defection ($A_y > 0$) but not necessarily to the players' reputations (e.g., $A_x = A_z = 0$). A recent study also argues that helping an ill-reputed player should be regarded as good to maintain stable cooperation[45]. Such understanding of indirect reciprocity in terms of sensitivity is important because, as is usual, information processing through reputation has a trade-off between robustness and sensitivity: One could underestimate new information and fail to adapt, or, one could overestimate it and fail to distinguish noise from the signal. In practice, the best way of assessment seems to be updating little by little upon arrival of new information[46], and such a possibility is already incorporated in this continuum formulation.

It should be emphasized that our analysis has focused on local perturbation to the existing norm. Therefore, our inequalities cannot be interpreted as a condition for evolutionary stability against every possible mutant. Moreover, although $\Delta\pi_0$ is found independent of p in our analysis, one should keep in mind that it results from a first-order theory so that higher-order corrections generally show dependence on p. If a mutant is sufficiently different from the resident, then the first-order theory fails and the payoff difference may well depend on p. For instance, if we think of a population consisting of L1 and L8 (Table 1), we see that L1 is better off only when it comprises the majority of the population (not shown). Having said that, our local analysis can nevertheless provide a necessary condition which will hold for stronger notions of stability as well. We also believe that this locality assumption is usually plausible in reality, considering that a social norm is a complex construct that combines expectation and action in a mutually reinforcing manner and thus resists change but small ones[47]. An empirical analysis shows that even orthographic and lexical norms change so slowly that it takes centuries unless intervened by a formal institution[48]. Another restriction in our analytic approach is that the mutation is assumed to have positive δ_1 so that the mutant is not fully content with the initial cooperative state. If two norms have $\delta_1 = 0$ in common and differ only by slopes at the initial state, the first-order perturbation does not give a definite answer as to their dynamics. Having positive δ_1 can be interpreted from a myopic player's point of view as follows: A selfish player in a cooperating population may feel tempted to devalue others' cooperation and reduce his or her own cost of cooperation toward them. If our condition is met, however, such behaviour will eventually be punished by the social norm.

"Maturity of mind is the capacity to endure uncertainty," says a maxim. Although one lesson of the life is that we have to accept the grey area between good and bad, reputation is still something that can be easily driven to extremes, and what is worse is that it often goes in a different direction for each observer. Despite the theoretical achievement of indirect reciprocity, its real difficulties are thus manifested in the problem of private assessment, noise, and incomplete information. Our finding suggests that we can get a better grip on indirect reciprocity by preparing reputational and behavioural scales with finer gradations, which may be thought of as a form of systematic deliberation to protect each other's reputation from rash judgement.

Methods

Linear-order corrections.

Equation (18) in the large-N limit is written as follows:

$$m_{00} = \frac{1}{N-1}\sum_{j\neq 0}\alpha_0[m_{00}, \beta_0(m_{00}, m_{0j}), m_{0j}] = \alpha_0[m_{00}, \beta_0(m_{00}, m_{01}), m_{01}] \tag{31}$$

$$m_{01} = \frac{1}{N-1}\sum_{j\neq 1}\alpha_0[m_{01}, \beta_1(m_{11}, m_{1j}), m_{0j}] \approx \alpha_0[m_{01}, \beta_1(m_{11}, m_{11}), m_{01}] \tag{32}$$

$$m_{10} = \frac{1}{N-1}\sum_{j\neq 0}\alpha_1[m_{10}, \beta_0(m_{00}, m_{0j}), m_{1j}] = \alpha_1[m_{10}, \beta_0(m_{00}, m_{01}), m_{11}] \tag{33}$$

$$m_{11} = \frac{1}{N-1}\sum_{j\neq 1}\alpha_1[m_{11}, \beta_1(m_{11}, m_{1j}), m_{1j}] \approx \alpha_1[m_{11}, \beta_1(m_{11}, m_{11}), m_{11}]. \tag{34}$$

With $m_{00} = 1-\varepsilon_{00}$, $m_{01} = 1-\varepsilon_{01}$, and $m_{10} = 1-\varepsilon_{10}$, Eq. (32) becomes

$$1-\varepsilon_{01} = \alpha_0(1-\varepsilon_{01}, 1, 1-\varepsilon_{01}) = \alpha(1-\varepsilon_{01}, 1, 1-\varepsilon_{01}) - \delta(1-\varepsilon_{01}, 1, 1-\varepsilon_{01}) \tag{35}$$

$$\approx \alpha(1,1,1) - A_x\varepsilon_{01} - A_z\varepsilon_{01} - \delta(1,1,1) = 1 - A_x\varepsilon_{01} - A_z\varepsilon_{01} - \delta_1, \tag{36}$$

where $\alpha_\xi \equiv \partial\alpha/\partial\xi|_{(1,1,1)}$ and $\delta_1 \equiv \delta(1,1,1)$. Thus, we have

$$\varepsilon_{01} \approx (1-A_x-A_z)^{-1}\delta_1. \tag{37}$$

Likewise,

$$\beta_0(1-\varepsilon_{00}, 1-\varepsilon_{01}) = \beta(1-\varepsilon_{00}, 1-\varepsilon_{01}) - \eta(1-\varepsilon_{00}, 1-\varepsilon_{01}) \tag{38}$$

$$\approx 1 - B_x\varepsilon_{00} - B_y\varepsilon_{01} - \eta_1, \tag{39}$$

where $\beta_\lambda \equiv \partial\beta/\partial\lambda|_{(1,1)}$ and $\eta_1 \equiv \eta(1,1)$. Using this expression, we obtain from Eq. (33) the following:

$$1-\varepsilon_{10} = \alpha\left(1-\varepsilon_{10}, 1 - B_x\varepsilon_{00} - B_y\varepsilon_{01} - \eta_1, 1\right) \tag{40}$$

$$\approx 1 - A_x\varepsilon_{10} - A_y\left(B_x\varepsilon_{00} + B_y\varepsilon_{01} + \eta_1\right), \tag{41}$$

which means

$$\varepsilon_{10} = \frac{A_y}{1-A_x}(B_x\varepsilon_{00} + B_y\varepsilon_{01} + \eta_1) \tag{42}$$

$$= \frac{A_y}{1-A_x}\left[B_x\varepsilon_{00} + B_y(1-A_x-A_z)^{-1}\delta_1 + \eta_1\right]. \tag{43}$$

To get a closed-form expression for this, we need ε_{00} in addition to ε_{01} [Eq. (37)]. Thus, from Eq. (31), we derive

$$1-\varepsilon_{00} \approx \alpha[1-\varepsilon_{00}, \beta_0(1-\varepsilon_{00}, 1-\varepsilon_{01}), 1-\varepsilon_{01}] - \delta_1 \tag{44}$$

$$\approx \alpha\left(1-\varepsilon_{00}, 1 - B_x\varepsilon_{00} - B_y\varepsilon_{01} - \eta_1, 1-\varepsilon_{01}\right) - \delta_1 \tag{45}$$

$$\approx 1 - A_x\varepsilon_{00} - A_y\left(B_x\varepsilon_{00} + B_y\varepsilon_{01} + \eta_1\right) - A_z\varepsilon_{01} - \delta_1, \tag{46}$$

which gives

$$\varepsilon_{00} = \frac{1}{1-A_x-A_yB_x}\left[(A_yB_y + A_z)\varepsilon_{01} + A_y\eta_1 + \delta_1\right] \tag{47}$$

$$= \frac{1}{1-A_x-A_yB_x}\left[\frac{A_yB_y + A_z}{1-A_x-A_z}\delta_1 + A_y\eta_1 + \delta_1\right], \tag{48}$$

where we have used Eq. (37). By substituting Eq. (48) into Eq. (43), we can write ε_{10} explicitly.

Finite fraction of mutants.

If a mutant norm occupies a finite fraction p, Eqs. (31) to (34) are generalized to

$$m_{00} = p\alpha_0[m_{00}, \beta_0(m_{00}, m_{00}), m_{00}] + \bar{p}\alpha_0[m_{00}, \beta_0(m_{00}, m_{01}), m_{01}] \tag{49}$$

$$m_{01} = p\alpha_0[m_{01}, \beta_1(m_{11}, m_{10}), m_{00}] + \bar{p}\alpha_0[m_{01}, \beta_1(m_{11}, m_{11}), m_{01}] \tag{50}$$

$$m_{10} = p\alpha_1[m_{10}, \beta_0(m_{00}, m_{00}), m_{10}] + \bar{p}\alpha_1[m_{10}, \beta_0(m_{00}, m_{01}), m_{11}] \tag{51}$$

$$m_{11} = p\alpha_1[m_{11}, \beta_1(m_{11}, m_{10}), m_{10}] + \bar{p}\alpha_1[m_{11}, \beta_1(m_{11}, m_{11}), m_{11}], \tag{52}$$

where $\bar{p} \equiv 1-p$. Through linearisation, the above equations are rewritten as

$$\begin{aligned} 1-\varepsilon_{00} \approx\; & p[1 - A_x\varepsilon_{00} - A_y(B_x\varepsilon_{00} + B_y\varepsilon_{00} + \eta_1) - A_z\varepsilon_{00} - \delta_1] \\ & + \bar{p}[1 - A_x\varepsilon_{00} - A_y(B_x\varepsilon_{00} + B_y\varepsilon_{01} + \eta_1) - A_z\varepsilon_{01} - \delta_1] \end{aligned} \tag{53}$$

$$\begin{aligned}1-\varepsilon_{01} \approx & p[1-A_x\varepsilon_{01}-A_y(B_x\varepsilon_{11}+B_y\varepsilon_{10})-A_z\varepsilon_{00}-\delta_1] \\ &+\bar{p}[1-A_x\varepsilon_{01}-A_y(B_x\varepsilon_{11}+B_y\varepsilon_{11})-A_z\varepsilon_{01}-\delta_1]\end{aligned} \tag{54}$$

$$\begin{aligned}1-\varepsilon_{10} \approx & p[1-A_x\varepsilon_{10}-A_y(B_x\varepsilon_{00}+B_y\varepsilon_{00}+\eta_1)-A_z\varepsilon_{10}] \\ &+\bar{p}[1-A_x\varepsilon_{10}-A_y(B_x\varepsilon_{00}+B_y\varepsilon_{01}+\eta_1)-A_z\varepsilon_{11}]\end{aligned} \tag{55}$$

$$\begin{aligned}1-\varepsilon_{11} \approx & p[1-A_x\varepsilon_{11}-A_y(B_x\varepsilon_{11}+B_y\varepsilon_{10})-A_z\varepsilon_{10}] \\ &+\bar{p}[1-A_x\varepsilon_{11}-A_y(B_x\varepsilon_{11}+B_y\varepsilon_{11})-A_z\varepsilon_{11}].\end{aligned} \tag{56}$$

After some algebra, we find

$$\begin{aligned}\varepsilon_{00} =& \frac{\delta_1\left\{{A_x}^2+A_x(A_yB_x+A_z-2)-\bar{p}{A_y}^2B_xB_y-\bar{p}{A_y}^2{B_y}^2+A_z[A_y(pB_x-\bar{p}B_y)-1]-A_yB_x+1\right\}}{(1-A_x-A_z)(1-A_x-A_yB_x)(1-A_x-A_yB_x-A_yB_y-A_z)} \\ &+\frac{A_y\eta_1(1-A_x-A_z)(1-A_x-A_yB_x-\bar{p}A_yB_y-\bar{p}A_z)}{(1-A_x-A_z)(1-A_x-A_yB_x)(1-A_x-A_yB_x-A_yB_y-A_z)}\end{aligned} \tag{57}$$

$$\begin{aligned}\varepsilon_{01} =& \frac{A_y\eta_1 p(1-A_x-A_z)(A_yB_y+A_z)}{(1-A_x-A_z)(1-A_x-A_yB_x)(1-A_x-A_yB_x-A_yB_y-A_z)} \\ &+\frac{\delta_1\left[{A_x}^2+A_x(2A_yB_x+A_yB_y+A_z-2)+{A_y}^2{B_x}^2+{A_y}^2B_xB_yp+{A_y}^2B_xB_y+{A_y}^2{B_y}^2p\right]}{(1-A_x-A_z)(1-A_x-A_yB_x)(1-A_x-A_yB_x-A_yB_y-A_z)} \\ &+\frac{\delta_1\left\{A_z[A_y(pB_x+B_x+pB_y)-1]-2A_yB_x-A_yB_y+1\right\}}{(1-A_x-A_z)(1-A_x-A_yB_x)(1-A_x-A_yB_x-A_yB_y-A_z)}\end{aligned} \tag{58}$$

$$\varepsilon_{10} = \frac{A_y(1-A_x-A_yB_x-\bar{p}A_yB_y-\bar{p}A_z)[\eta_1(1-A_x-A_z)+(B_x+B_y)\delta_1]}{(1-A_x-A_z)(1-A_x-A_yB_x)(1-A_x-A_yB_x-A_yB_y-A_z)} \tag{59}$$

$$\varepsilon_{11} = \frac{A_yp(A_yB_y+A_z)(\eta_1(1-A_x-A_z)+(B_x+B_y)\delta_1)}{(1-A_x-A_z)(1-A_x-A_yB_x)(1-A_x-A_yB_x-A_yB_y-A_z)}, \tag{60}$$

from which one can reproduce the previous results [Eqs. (20) to (22)] by taking the limit of $p \to 0$. The denominators seem to require another inequality in addition to Eqs. (23) and (24), that is,

$$A_x+A_z+A_y(B_x+B_y) < 1, \tag{61}$$

which is equivalent to Eq. (17). Recall that the continuous versions of the leading eight always have $A_y = B_y = 1$ and $A_x = B_x = 0$ in common, which means that they all violate this inequality. However, in practice, no singularity arises for Simple Standing if higher-order corrections are included, and even the second-order calculation agrees moderately well with numerical results.

The payoff earned by a mutant is calculated as

$$\begin{aligned}\pi_0 =& b[p\beta_0(m_{00},m_{00})+(1-p)\beta_1(m_{11},m_{10})] \\ &-c[p\beta_0(m_{00},m_{00})+(1-p)\beta_0(m_{00},m_{01})]\end{aligned} \tag{62}$$

$$\begin{aligned}\approx & b[p(1-B_x\varepsilon_{00}-B_y\varepsilon_{00}-\eta_1)+(1-p)(1-B_x\varepsilon_{11}-B_y\varepsilon_{10})] \\ &-c[p(1-B_x\varepsilon_{00}-B_y\varepsilon_{00}-\eta_1)+(1-p)[1-B_x\varepsilon_{00}-B_y\varepsilon_{01}-\eta_1]],\end{aligned} \tag{63}$$

whereas a resident player earns

$$\begin{aligned}\pi_1 =& b[p\beta_0(m_{00},m_{01})+(1-p)\beta_1(m_{11},m_{11})] \\ &-c[p\beta_1(m_{11},m_{10})+(1-p)\beta_1(m_{11},m_{11})]\end{aligned} \tag{64}$$

$$\begin{aligned}\approx & b[p(1-B_x\varepsilon_{00}-B_y\varepsilon_{01}-\eta_1)+(1-p)(1-B_x\varepsilon_{11}-B_y\varepsilon_{11})] \\ &-c[p(1-B_x\varepsilon_{11}-B_y\varepsilon_{10})+(1-p)(1-B_x\varepsilon_{11}-B_y\varepsilon_{11})].\end{aligned} \tag{65}$$

If we plug Eqs. (57) to (60) here, the payoff difference $\Delta\pi_0 = \pi_0 - \pi_1$ becomes identical to Eq. (28) with no dependence on p.

Second-order corrections. We assume that δ, η, as well as their partial derivatives, and ε_{ij}'s are small parameters of the same order of magnitude. The second-order perturbation for β_1 can thus be written as follows:

$$\beta_1(m_{11},m_{1j}) = \beta(1-\varepsilon_{11},1-\varepsilon_{1j}) \tag{66}$$

$$\approx 1 - B_x\varepsilon_{11} - B_y\varepsilon_{1j} + \frac{1}{2}B_{xx}\varepsilon_{11}^2 + B_{xy}\varepsilon_{11}\varepsilon_{1j} + \frac{1}{2}B_{yy}\varepsilon_{1j}^2 \tag{67}$$

$$\equiv 1 - \kappa_1. \tag{68}$$

Here, we write $\kappa_1 \equiv \kappa_1^{(1)} + \kappa_1^{(2)}$, where $\kappa_1^{(1)} \equiv B_x\varepsilon_{11} + B_y\varepsilon_{1j}$ and $\kappa_1^{(2)} \equiv -\left(\frac{1}{2}B_{xx}\varepsilon_{11}^2 + B_{xy}\varepsilon_{11}\varepsilon_{1j} + \frac{1}{2}B_{yy}\varepsilon_{1j}^2\right)$ are first- and second-order corrections, respectively, and $B_{\mu\nu} \equiv \partial^2\beta/\partial\mu\partial\nu\big|_{(1,1)}$. Likewise,

$$\beta_0(m_{00}, m_{0j}) = \beta(m_{00}, m_{0j}) - \eta(m_{00}, m_{0j}) \tag{69}$$

$$= \beta(1-\varepsilon_{00}, 1-\varepsilon_{0j}) - \eta(1-\varepsilon_{00}, 1-\varepsilon_{0j}) \tag{70}$$

$$\begin{aligned} \approx &\left(1 - B_x\varepsilon_{00} - B_y\varepsilon_{0j} + \frac{1}{2}B_{xx}\varepsilon_{00}^2 + B_{xy}\varepsilon_{00}\varepsilon_{0j} + \frac{1}{2}B_{yy}\varepsilon_{0j}^2\right) \\ &- \left(\eta_1 - \eta_x\varepsilon_{00} - \eta_y\varepsilon_{0j}\right) \end{aligned} \tag{71}$$

$$\equiv 1 - \kappa_0, \tag{72}$$

where $\kappa_0 \equiv \kappa_0^{(1)} + \kappa_0^{(2)}$ with $\kappa_0^{(1)} \equiv B_x\varepsilon_{00} + B_y\varepsilon_{0j} + \eta_1$ and $\kappa_0^{(2)} \equiv -\left(\frac{1}{2}B_{xx}\varepsilon_{00}^2 + B_{xy}\varepsilon_{00}\varepsilon_{0j} + \frac{1}{2}B_{yy}\varepsilon_{0j}^2\right) - (\eta_x\varepsilon_{00} + \eta_y\varepsilon_{0j})$.

The second-order perturbation for α_1 is also straightforward:

$$\alpha_1[m_{1i}, \beta_i(m_{ii}, m_{ij}), m_{1j}] \approx \alpha(1-\varepsilon_{1i}, 1-\kappa_i, 1-\varepsilon_{1j}) \tag{73}$$

$$\begin{aligned} \approx &1 - A_x\varepsilon_{1i} - A_y\kappa_i - A_z\varepsilon_{1j} + \frac{1}{2}A_{xx}\varepsilon_{1i}^2 + \frac{1}{2}A_{yy}\left(\kappa_i^{(1)}\right)^2 + \frac{1}{2}A_{zz}\varepsilon_{1j}^2 \\ &+ A_{xy}\varepsilon_{1i}\kappa_i^{(1)} + A_{yz}\kappa_i^{(1)}\varepsilon_{1j} + A_{zx}\varepsilon_{1i}\varepsilon_{1j}, \end{aligned} \tag{74}$$

where $A_{\mu\nu} \equiv \partial^2\alpha/\partial\mu\partial\nu\big|_{(1,1,1)}$, and similarly,

$$\alpha_0[m_{0i}, \beta_i(m_{ii}, m_{ij}), m_{0j}] \approx \alpha(1-\varepsilon_{0i}, 1-\kappa_i, 1-\varepsilon_{0j}) - \delta(1-\varepsilon_{0i}, 1-\kappa_i, 1-\varepsilon_{0j}) \tag{75}$$

$$\begin{aligned} \approx &\Big[1 - A_x\varepsilon_{0i} - A_y\kappa_i - A_z\varepsilon_{0j} + \frac{1}{2}A_{xx}\varepsilon_{0i}^2 + \frac{1}{2}A_{yy}\left(\kappa_i^{(1)}\right)^2 + \frac{1}{2}A_{zz}\varepsilon_{0j}^2 \\ &+ A_{xy}\varepsilon_{0i}\kappa_i^{(1)} + A_{yz}\kappa_i^{(1)}\varepsilon_{0j} + A_{zx}\varepsilon_{0i}\varepsilon_{0j}\Big] \\ &- \left(\delta_1 - \delta_x\varepsilon_{0i} - \delta_y\kappa_i^{(1)} - \delta_z\varepsilon_{0j}\right). \end{aligned} \tag{76}$$

Data availability

The source code for this study is available at https://github.com/yohm/sim_game_continuous_reputation.

Received: 9 April 2021; Accepted: 28 June 2021
Published online: 09 July 2021

References

1. Burke, J. *The Day the Universe Changed* (London Writers Ltd., London, 1985).
2. Hamlin, J. K., Wynn, K. & Bloom, P. Social evaluation by preverbal infants. *Nature* **450**, 557–559 (2007).
3. Engelmann, J. M., Herrmann, E. & Tomasello, M. Five-year olds, but not chimpanzees, attempt to manage their reputations. *PLoS ONE* **7**, e48433 (2012).
4. Abdai, J. & Miklósi, Á. The origin of social evaluation, social eavesdropping, reputation formation, image scoring or what you will. *Front. Psychol.* **7**, 1772 (2016).
5. Alexander, R. *The Biology of Moral Systems* (A. de Gruyter, New York, 1987).
6. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573 (1998).
7. Leimar, O. & Hammerstein, P. Evolution of cooperation through indirect reciprocity. *Proc. R. Roc. Lond. B* **268**, 745–753 (2001).
8. Brandt, H. & Sigmund, K. Indirect reciprocity, image scoring, and moral hazard. *Proc. Natl. Acad. Sci. USA* **102**, 2666–2670 (2005).
9. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity. *Nature* **437**, 1291–1298 (2005).
10. Ohtsuki, H., Iwasa, Y. & Nowak, M. A. Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature* **457**, 79 (2009).
11. Nax, H. H., Perc, M., Szolnoki, A. & Helbing, D. Stability of cooperation under image scoring in group interactions. *Sci. Rep.* **5**, 12145 (2015).
12. Axelrod, R. *Evolution of Cooperation* (Basic Books, New York, 1984).
13. Baek, S. K. *et al.* Intelligent tit-for-tat in the iterated prisoner's dilemma game. *Phys. Rev. E* **78**, 011125 (2008).
14. Baek, S. K., Jeong, H.-C., Hilbe, C. & Nowak, M. A. Comparing reactive and memory-one strategies of direct reciprocity. *Sci. Rep.* **6**, 1–13 (2016).
15. Yi, S. D., Baek, S. K. & Choi, J.-K. Combination with anti-tit-for-tat remedies problems of tit-for-tat. *J. Theor. Biol.* **412**, 1–7 (2017).
16. Murase, Y. & Baek, S. K. Seven rules to avoid the tragedy of the commons. *J. Theor. Biol.* **449**, 94–102 (2018).
17. Murase, Y. & Baek, S. K. Automata representation of successful strategies for social dilemmas. *Sci. Rep.* **10**, 13370 (2020).

18. Murase, Y. & Baek, S. K. Five rules for friendly rivalry in direct reciprocity. *Sci. Rep.* **10**, 16904 (2020).
19. Murase, Y. & Baek, S. K. Friendly-rivalry solution to the iterated n-person public-goods game. *PLoS Comput. Biol.* **17**, e1008217 (2021).
20. Clark, D., Fudenberg, D. & Wolitzky, A. Indirect reciprocity with simple records. *Proc. Natl. Acad. Sci. USA* **117**, 11344–11349 (2020).
21. Wedekind, C. & Milinski, M. Cooperation through image scoring in humans. *Science* **288**, 850–852 (2000).
22. Milinski, M., Semmann, D. & Krambeck, H.-J. Reputation helps solve the 'tragedy of the commons'. *Nature* **415**, 424–426 (2002).
23. Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560–1563 (2006).
24. Kandori, M. Social norms and community enforcement. *Rev. Econ. Stud.* **59**, 63–80 (1992).
25. Ohtsuki, H. & Iwasa, Y. How should we define goodness? Reputation dynamics in indirect reciprocity. *J. Theor. Biol.* **231**, 107–120 (2004).
26. Ohtsuki, H. & Iwasa, Y. The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.* **239**, 435–444 (2006).
27. Sugden, R. *The Economics of Rights, Cooperation and Welfare* (Blackwell, Oxford, 1986).
28. Boyd, R. Mistakes allow evolutionary stability in the repeated prisoner's dilemma game. *J. Theor. Biol.* **136**, 47–56 (1989).
29. Panchanathan, K. & Boyd, R. A tale of two defectors: The importance of standing for evolution of indirect reciprocity. *J. Theor. Biol.* **224**, 115–126 (2003).
30. Brandt, H., Ohtsuki, H., Iwasa, Y. & Sigmund, K. A survey of indirect reciprocity. In Takeuchi, Y., Iwasa, Y. & Sato, K. (eds.) *Mathematics for ecology and environmental sciences*, 30 (Springer, Berlin, 2007).
31. Uchida, S. Effect of private information on indirect reciprocity. *Phys. Rev. E* **82**, 036111 (2010).
32. Uchida, S. & Sasaki, T. Effect of assessment error and private information on stern-judging in indirect reciprocity. *Chaos Solitons Fractals* **56**, 175–180 (2013).
33. Okada, I., Sasaki, T. & Nakai, Y. Tolerant indirect reciprocity can boost social welfare through solidarity with unconditional cooperators in private monitoring. *Sci. Rep.* **7**, 1–11 (2017).
34. Okada, I., Sasaki, T. & Nakai, Y. A solution for private assessment in indirect reciprocity using solitary observation. *J. Theor. Biol.* **455**, 7–15 (2018).
35. Santos, F. P., Santos, F. C. & Pacheco, J. M. Social norm complexity and past reputations in the evolution of cooperation. *Nature* **555**, 242–245 (2018).
36. Hilbe, C., Schmid, L., Tkadlec, J., Chatterjee, K. & Nowak, M. A. Indirect reciprocity with private, noisy, and incomplete information. *Proc. Natl. Acad. Sci. USA* **115**, 12241–12246 (2018).
37. Radzvilavicius, A. L., Stewart, A. J. & Plotkin, J. B. Evolution of empathetic moral evaluation.. *Elife* **8**, e44269 (2019).
38. Quan, J. *et al.* Withhold-judgment and punishment promote cooperation in indirect reciprocity under incomplete information. *EPL* **128**, 28001 (2020).
39. Tanabe, S., Suzuki, H. & Masuda, N. Indirect reciprocity with trinary reputations. *J. Theor. Biol.* **317**, 338–347 (2013).
40. Olejarz, J., Ghang, W. & Nowak, M. Indirect reciprocity with optional interactions and private information. *Games* **6**, 438–457 (2015).
41. *Mathematica, Version 10.0* (Wolfram Research, Inc., Champaign, IL, 2014).
42. Alwin, D. F. Feeling thermometers versus 7-point scales: Which are better?. *Sociol. Methods Res.* **25**, 318–340 (1997).
43. Preston, C. C. & Colman, A. M. Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychol.* **104**, 1–15 (2000).
44. Svensson, E. Comparison of the quality of assessments using continuous and discrete ordinal rating scales. *Biom. J* **42**, 417–434 (2000).
45. Okada, I. Two ways to overcome the three social dilemmas of indirect reciprocity. *Sci. Rep.* **10**, 1–9 (2020).
46. Tetlock, P. E. & Gardner, D. *Superforecasting: The art and science of prediction* (Random House, New York, 2015).
47. Mackie, G., Moneti, F., Denny, E. & Shakya, H. *What are Social Norms? How are They Measured?* (UNICEF/UCSD Center on Global Justice Project Cooperation Agreement Working Paper, San Diego, CA, 2014).
48. Amato, R., Lacasa, L., Díaz-Guilera, A. & Baronchelli, A. The dynamics of norm change in the cultural evolution of language. *Proc. Natl. Acad. Sci. USA* **115**, 8260–8265 (2018).

Acknowledgements

Y.M. acknowledges support from Japan Society for the Promotion of Science (JSPS) (JSPS KAKENHI; Grant no. 18H03621 and Grant no. 21K03362). S.K.B. acknowledges support by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1I1A2071670). We appreciate the APCTP for its hospitality during completion of this work.

Author contributions

S.L. carried out computation and analysed the results. Y.M. verified the method and reviewed the manuscript. S.K.B. conceived the work and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.K.B.

Reprints and permissions information is available at www.nature.com/reprints.

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