Supplementary Information for

Evolutionary stability of cooperation in indirect reciprocity with noisy and private reputation

Yuma Fujimoto, and Hisashi Ohtsuki

Yuma Fujimoto.
E-mail: fujimoto_yuma@soken.ac.jp

This PDF file includes:

Supporting Information Text

1. Calculation of joint distribution of goodnesses

This section proposes an analytical method to obtain the reputation structure under indirect reciprocity. We assume a situation where rare mutants with norm M of ratio δ invade other wild-types with norm W of ratio $1-\delta$. We denote the population ratio of norm $A \in \{\mathrm{W},\mathrm{M}\}$ as ρ_A; $\rho_A = 1-\delta$ when $A = \mathrm{W}$, while $\rho_A = \delta$ when $A = \mathrm{M}$. To characterize the reputation structure, we define p_{iA} as a proportion of individuals of norm A who assign good reputations to individual i. We call p_{iA} a goodness of individual i from norm $A \in \{\mathrm{W},\mathrm{M}\}$.

In the following, let us consider a stochastic transition of p_{iA} in each round. In a single round, a recipient and a donor are chosen and labeled as i_R and i_D, respectively. In this round, $p_{i_\mathrm{D}A_\mathrm{O}}$, i.e., the goodness of donor from norm A_O, changes into the next goodness $p'_{i_\mathrm{D}A_\mathrm{O}}$ for all $A_\mathrm{O} \in \{\mathrm{W},\mathrm{M}\}$. Below, we formulate the stochastic change separately for cases that the donor chooses to cooperate or defect.

C-map case: First, we consider a case that the donor cooperates with the recipient, occurring with a probability of

$$h(p_{i_\mathrm{R}A_\mathrm{D}}) := p_{i_\mathrm{R}A_\mathrm{D}}(1-e_1) + (1-p_{i_\mathrm{R}A_\mathrm{D}})e_1. \tag{1}$$

Here, e_1 is the probability of action error, with which the donor chooses an opposite action to the intended one. In this case, $N\rho_{A_\mathrm{O}}p'_{i_\mathrm{D}A_\mathrm{O}}$, i.e., the number of observers with norm A_O who give good reputations to the donor in the next round, follows a probability distribution of

$$N\rho_{A_\mathrm{O}}p'_{i_\mathrm{D}A_\mathrm{O}} \sim N_1 + N_2, \tag{2}$$
$$N_1 \sim \mathcal{B}(N\rho_{A_\mathrm{O}}p_{i_\mathrm{R}A_\mathrm{O}}, a^{\mathrm{GC}}_{A_\mathrm{O}}), \tag{3}$$
$$N_2 \sim \mathcal{B}(N\rho_{A_\mathrm{O}}(1-p_{i_\mathrm{R}A_\mathrm{O}}), a^{\mathrm{BC}}_{A_\mathrm{O}}). \tag{4}$$

Here, $\mathcal{B}(n,a)$ denotes a binomial distribution with success probability a and trial number n. In addition, $a^{XY}_{A_\mathrm{O}}$ denotes the probability that an observer with norm A_O who evaluates the recipient as $X \in \{\mathrm{G},\mathrm{B}\}$ newly gives a good reputation to the donor whose action is $Y \in \{\mathrm{C},\mathrm{D}\}$. $a^{\mathrm{GC}}_{A_\mathrm{O}}$, $a^{\mathrm{BC}}_{A_\mathrm{O}}$, $a^{\mathrm{GD}}_{A_\mathrm{O}}$, and $a^{\mathrm{BD}}_{A_\mathrm{O}}$ are obtained by converting corresponding G and B pivots into $1-e_2$ and e_2 in Table 1 of the main manuscript. Instead of Eq. (3), we use a shorthand notation;

$$N\rho_{A_\mathrm{O}}p'_{i_\mathrm{D}A_\mathrm{O}} \sim \mathcal{B}(N\rho_{A_\mathrm{O}}p_{i_\mathrm{R}A_\mathrm{O}}, a^{\mathrm{GC}}_{A_\mathrm{O}}) + \mathcal{B}(N\rho_{A_\mathrm{O}}(1-p_{i_\mathrm{R}A_\mathrm{O}}), a^{\mathrm{BC}}_{A_\mathrm{O}}). \tag{5}$$

Because $N\rho_{A_\mathrm{O}}$ is sufficiently large, the mean and variance of $p'_{i_\mathrm{D}A_\mathrm{O}}$ are given by

$$\mathrm{E}[p'_{i_\mathrm{D}A_\mathrm{O}}] = p_{i_\mathrm{R}A_\mathrm{O}}(\underbrace{a^{\mathrm{GC}}_{A_\mathrm{O}} - a^{\mathrm{BC}}_{A_\mathrm{O}}}_{=:\Delta f^{\mathrm{C}}_{A_\mathrm{O}}}) + a^{\mathrm{BC}}_{A_\mathrm{O}} \quad (=: f^{\mathrm{C}}_{A_\mathrm{O}}(p_{i_\mathrm{R}A_\mathrm{O}})), \tag{6}$$

$$\mathrm{Var}[p'_{i_\mathrm{D}A_\mathrm{O}}] = \frac{p_{i_\mathrm{R}A_\mathrm{O}}a^{\mathrm{GC}}_{A_\mathrm{O}}(1-a^{\mathrm{GC}}_{A_\mathrm{O}}) + (1-p_{i_\mathrm{R}A_\mathrm{O}})a^{\mathrm{BC}}_{A_\mathrm{O}}(1-a^{\mathrm{BC}}_{A_\mathrm{O}})}{N\rho_{A_\mathrm{O}}} = \frac{e_2(1-e_2)}{N\rho_{A_\mathrm{O}}} \quad (=: \rho^{-1}_{A_\mathrm{O}}s^2). \tag{7}$$

Here, e_2 is the probability of assessment error, with which an observer assigns an opposite reputation to the intended one. In Eq. (6), $f^{\mathrm{C}}_{A_\mathrm{O}}$ represents a map from the recipient's goodness in the present round to the donor's goodness in the next round. Because this map is applied only when the donor cooperates, we call it "C-map".

D-map case: On the other hand, we consider a case that the donor defects with the recipient, occurring with a probability of

$$1 - h(p_{i_\mathrm{R}A_\mathrm{D}}) = (1-p_{i_\mathrm{R}A_\mathrm{D}})(1-e_1) + p_{i_\mathrm{R}A_\mathrm{D}}e_1. \tag{8}$$

In this case, $N\rho_{A_\mathrm{O}}p'_{i_\mathrm{D}A_\mathrm{O}}$ follows a probability distribution of

$$N\rho_{A_\mathrm{O}}p'_{i_\mathrm{D}A_\mathrm{O}} \sim \mathcal{B}(N\rho_{A_\mathrm{O}}p_{i_\mathrm{R}A_\mathrm{O}}, a^{\mathrm{GD}}_{A_\mathrm{O}}) + \mathcal{B}(N\rho_{A_\mathrm{O}}(1-p_{i_\mathrm{R}A_\mathrm{O}}), a^{\mathrm{BD}}_{A_\mathrm{O}}). \tag{9}$$

From this equation, the mean and variance of $p'_{i_\mathrm{D}A_\mathrm{O}}$ are given by

$$\mathrm{E}[p'_{i_\mathrm{D}A_\mathrm{O}}] = p_{i_\mathrm{R}A_\mathrm{O}}(\underbrace{a^{\mathrm{GD}}_{A_\mathrm{O}} - a^{\mathrm{BD}}_{A_\mathrm{O}}}_{=:\Delta f^{\mathrm{D}}_{A_\mathrm{O}}}) + a^{\mathrm{BD}}_{A_\mathrm{O}} \quad (=: f^{\mathrm{D}}_{A_\mathrm{O}}(p_{i_\mathrm{R}A_\mathrm{O}})), \tag{10}$$

$$\mathrm{Var}[p'_{i_\mathrm{D}A_\mathrm{O}}] = \frac{p_{i_\mathrm{R}A_\mathrm{O}}a^{\mathrm{GD}}_{A_\mathrm{O}}(1-a^{\mathrm{GD}}_{A_\mathrm{O}}) + (1-p_{i_\mathrm{R}A_\mathrm{O}})a^{\mathrm{BD}}_{A_\mathrm{O}}(1-a^{\mathrm{BD}}_{A_\mathrm{O}})}{N\rho_{A_\mathrm{O}}} = \frac{e_2(1-e_2)}{N\rho_{A_\mathrm{O}}} \quad (=: \rho^{-1}_{A_\mathrm{O}}s^2). \tag{11}$$

Because the map $f^{\mathrm{D}}_{A_\mathrm{O}}$ is applied when the donor defects, we call it D-map in the same way as C-map.

The above C-map $f^{\mathrm{C}}_{S_k}$ and D-map $f^{\mathrm{D}}_{S_k}$ are illustrated in Fig. S1 for all $S_k \in \mathcal{S}$.

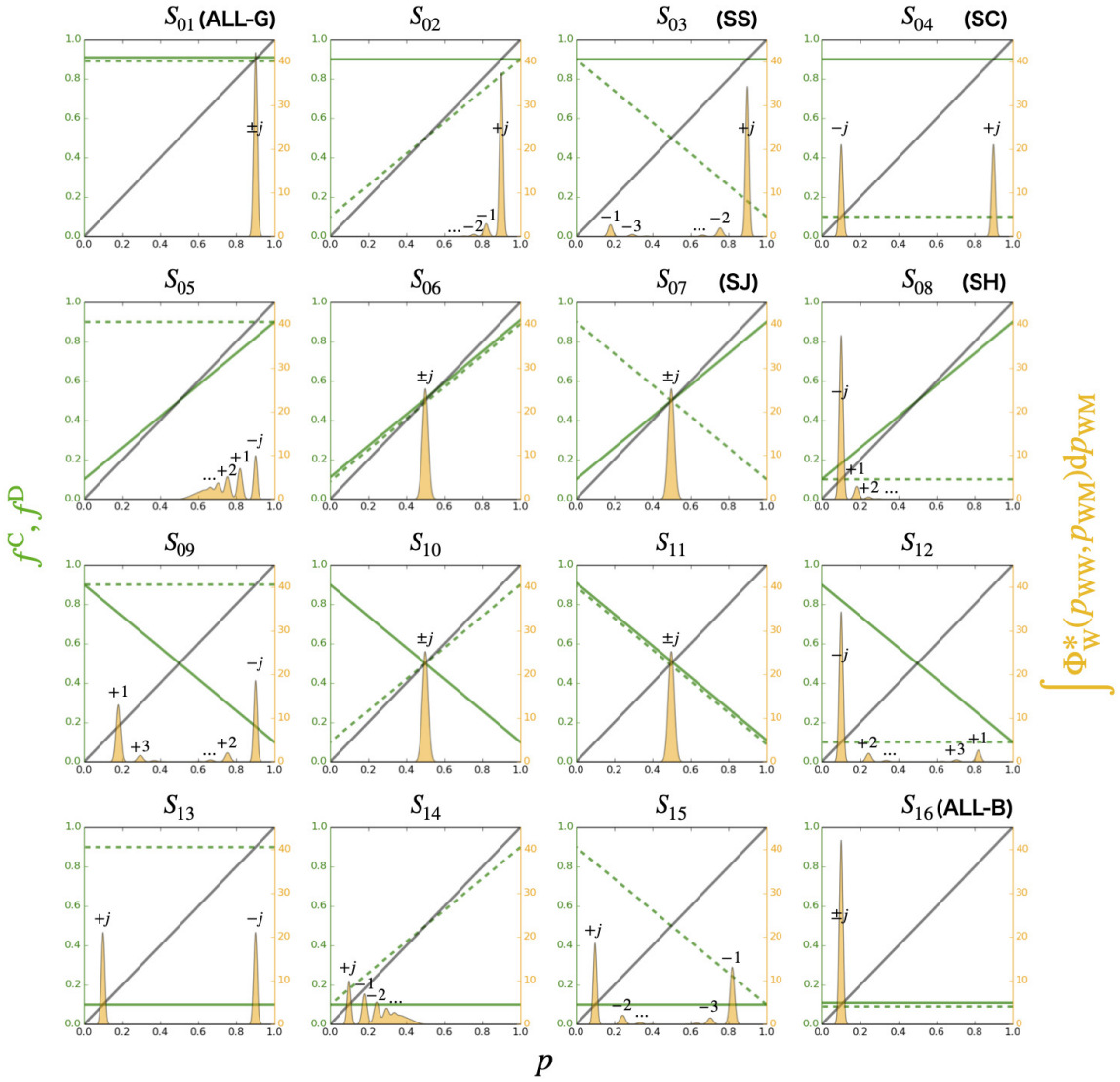

Fig. S1. Materials for the reputation structure for all the second-order norms $\mathrm{W} = S_k$. Green solid (resp. dashed) lines indicate C-map $f^{\mathrm{C}}_{\mathrm{W}}$ (resp. D-map $f^{\mathrm{D}}_{\mathrm{W}}$) of the norm. Gray lines indicate the identity map, which shows the fixed points of the C-map and D-map as the crossing points with these maps. The orange distribution shows the probability density function of goodnesses p_{WW} when $\mathrm{W} = S_k$. All the panels are output under $N = 2000$, $\delta = 0$, and $(e_1, e_2) = (0, 0.1)$. Numbers over each peak indicate j.

2. Time evolution of reputation structure

Because the population of wild-types and mutants are sufficiently large, we can continualize the distribution of individual goodness p_{iA} with separating into the cases that the norm of individual i is W or M. In the following, $p_{AA'}$ denotes a continualized goodness of an individual with norm A in the eyes of individuals with norm A'. Let us consider a time change of the distribution of $p_{AA'}$. As shown above, however, we should keep in mind that $p_{A\mathrm{W}}$ and $p_{A\mathrm{M}}$ are simultaneously changed by the C-map or D-map. Thus, we consider dynamics of $\Phi_A(p_{A\mathrm{W}}, p_{A\mathrm{M}})$, a joint probability distribution of $p_{A\mathrm{W}}$ and $p_{A\mathrm{M}}$. Note that a norm of the chosen recipient is M only with a probability of δ, which contributes the dynamics of Φ_A only in a scale of $O(\delta)$. By ignoring this scale of $O(\delta)$, the dynamics of Φ_A is given by

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\Phi_A(p_{A\mathrm{W}}, p_{A\mathrm{M}}) = -\Phi_A(p_{A\mathrm{W}}, p_{A\mathrm{M}}) + \int_0^1\int_0^1 \{&h(p'_{\mathrm{W}A})g(p_{A\mathrm{W}}; f^{\mathrm{C}}_{\mathrm{W}}(p'_{\mathrm{WW}}), \rho_{\mathrm{W}}^{-1}s^2)g(p_{A\mathrm{M}}; f^{\mathrm{C}}_{\mathrm{M}}(p'_{\mathrm{WM}}), \rho_{\mathrm{M}}^{-1}s^2)\\ &+ (1-h(p'_{\mathrm{W}A}))g(p_{A\mathrm{W}}; f^{\mathrm{D}}_{\mathrm{W}}(p'_{\mathrm{WW}}), \rho_{\mathrm{W}}^{-1}s^2)g(p_{A\mathrm{M}}; f^{\mathrm{D}}_{\mathrm{M}}(p'_{\mathrm{WM}}), \rho_{\mathrm{M}}^{-1}s^2)\}\\ &\times \Phi_{\mathrm{W}}(p'_{\mathrm{WW}}, p'_{\mathrm{WM}})\mathrm{d}p'_{\mathrm{WW}}\mathrm{d}p'_{\mathrm{WM}}.\end{aligned} \tag{12}$$

Here, $g(p;\mu,\sigma^2)$ denotes a Gaussian function with the mean μ and variance σ^2 as

$$g(p;\mu,\sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(p-\mu)^2}{2\sigma^2}\right). \tag{13}$$

Equation (12) explains an update of the donor's goodness per time. The first (resp. second) term on the right side represents decrements (increments) by updating goodnesses. In detail, $\Phi_{\mathrm{W}}(p'_{\mathrm{WW}}, p'_{\mathrm{WM}})$ in the second term shows the density that the recipient's goodness is p'_{WW} (resp. p'_{WM}) in the eyes of wild-types (resp. mutants). $h(p'_{\mathrm{W}A})$ shows the probability that the donor cooperates, and the donor's goodnesses after the update in the eyes of observers with norm W and M are described by $g(p_{A\mathrm{W}}; f^{\mathrm{C}}_{\mathrm{W}}(p'_{\mathrm{WW}}), \rho_{\mathrm{W}}^{-1}s^2)$ and $g(p_{A\mathrm{M}}; f^{\mathrm{C}}_{\mathrm{M}}(p'_{\mathrm{WM}}), \rho_{\mathrm{M}}^{-1}s^2)$, respectively. A similar explanation holds when the donor chooses to defect.

The equilibrium state of Eq. (12), i.e., Φ^*_A, satisfies

$$\begin{aligned}\Phi^*_A(p_{A\mathrm{W}}, p_{A\mathrm{M}}) = \int_0^1\int_0^1 \{&h(p'_{\mathrm{W}A})g(p_{A\mathrm{W}}; f^{\mathrm{C}}_{\mathrm{W}}(p'_{\mathrm{WW}}), \rho_{\mathrm{W}}^{-1}s^2)g(p_{A\mathrm{M}}; f^{\mathrm{C}}_{\mathrm{M}}(p'_{\mathrm{WM}}), \rho_{\mathrm{M}}^{-1}s^2)\\ &+ (1-h(p'_{\mathrm{W}A}))g(p_{A\mathrm{W}}; f^{\mathrm{D}}_{\mathrm{W}}(p'_{\mathrm{WW}}), \rho_{\mathrm{W}}^{-1}s^2)g(p_{A\mathrm{M}}; f^{\mathrm{D}}_{\mathrm{M}}(p'_{\mathrm{WM}}), \rho_{\mathrm{M}}^{-1}s^2)\}\\ &\times \Phi^*_{\mathrm{W}}(p'_{\mathrm{WW}}, p'_{\mathrm{WM}})\mathrm{d}p'_{\mathrm{WW}}\mathrm{d}p'_{\mathrm{WM}}.\end{aligned} \tag{14}$$

To solve this equation, we assume that the equilibrium state can be described by a summation of two-dimensional Gaussian functions without correlation as

$$\Phi^*_A(p_{A\mathrm{W}}, p_{A\mathrm{M}}) = \sum_j q_{Aj}g(p_{A\mathrm{W}}; \mu_{A\mathrm{W}j}, \rho_{\mathrm{W}}^{-1}\sigma^2_{A\mathrm{W}j})g(p_{A\mathrm{M}}; \mu_{A\mathrm{M}j}, \rho_{\mathrm{M}}^{-1}\sigma^2_{A\mathrm{M}j}). \tag{15}$$

The assumption of Gaussian is justified by the above transition process of the donor's goodness, where the goodness is virtually determined only by the mean and variance in a sufficiently large population. No correlation is assumed because the variance is given independently by observers with different norms.

We now derive equations in which the equilibrium state satisfies for each norm $A \in \{\mathrm{W}, \mathrm{M}\}$. First, substituting Eq. (15)

into Eq. (14) for $A = \mathrm{W}$, we obtain

$$
\begin{aligned}
&\sum_j q_{\mathrm{W}j} g(p_{\mathrm{WW}}; \mu_{\mathrm{WW}j}, \rho_{\mathrm{W}}^{-1}\sigma^2_{\mathrm{WW}j}) g(p_{\mathrm{WM}}; \mu_{\mathrm{WM}j}, \rho_{\mathrm{M}}^{-1}\sigma^2_{\mathrm{WM}j}) \\
&= \int_0^1 \int_0^1 \{h(p'_{\mathrm{WW}}) g(p_{\mathrm{WW}}; f^{\mathrm{C}}_{\mathrm{W}}(p'_{\mathrm{WW}}), \rho_{\mathrm{W}}^{-1}s^2) g(p_{\mathrm{WM}}; f^{\mathrm{C}}_{\mathrm{M}}(p'_{\mathrm{WM}}), \rho_{\mathrm{M}}^{-1}s^2) \\
&\qquad + (1-h(p'_{\mathrm{WW}})) g(p_{\mathrm{WW}}; f^{\mathrm{D}}_{\mathrm{W}}(p'_{\mathrm{WW}}), \rho_{\mathrm{W}}^{-1}s^2) g(p_{\mathrm{WM}}; f^{\mathrm{D}}_{\mathrm{M}}(p'_{\mathrm{WM}}), \rho_{\mathrm{M}}^{-1}s^2)\} \\
&\qquad \times \sum_j q_{\mathrm{W}j} g(p'_{\mathrm{WW}}; \mu_{\mathrm{WW}j}, \rho_{\mathrm{W}}^{-1}\sigma^2_{\mathrm{WW}j}) g(p'_{\mathrm{WM}}; \mu_{\mathrm{WM}j}, \rho_{\mathrm{M}}^{-1}\sigma^2_{\mathrm{WM}j}) \mathrm{d}p'_{\mathrm{WW}} \mathrm{d}p'_{\mathrm{WM}}, \\
&= \sum_j q_{\mathrm{W}j} \int_0^1 \int_0^1 \{h(p'_{\mathrm{WW}}) g(p_{\mathrm{WW}}; f^{\mathrm{C}}_{\mathrm{W}}(p'_{\mathrm{WW}}), \rho_{\mathrm{W}}^{-1}s^2) g(p_{\mathrm{WM}}; f^{\mathrm{C}}_{\mathrm{M}}(p'_{\mathrm{WM}}), \rho_{\mathrm{M}}^{-1}s^2) \\
&\qquad + (1-h(p'_{\mathrm{WW}})) g(p_{\mathrm{WW}}; f^{\mathrm{D}}_{\mathrm{W}}(p'_{\mathrm{WW}}), \rho_{\mathrm{W}}^{-1}s^2) g(p_{\mathrm{WM}}; f^{\mathrm{D}}_{\mathrm{M}}(p'_{\mathrm{WM}}), \rho_{\mathrm{M}}^{-1}s^2)\} \\
&\qquad \times g(p'_{\mathrm{WW}}; \mu_{\mathrm{WW}j}, \rho_{\mathrm{W}}^{-1}\sigma^2_{\mathrm{WW}j}) g(p'_{\mathrm{WM}}; \mu_{\mathrm{WM}j}, \rho_{\mathrm{M}}^{-1}\sigma^2_{\mathrm{WM}j}) \mathrm{d}p'_{\mathrm{WW}} \mathrm{d}p'_{\mathrm{WM}}, \\
&\simeq \sum_j q_{\mathrm{W}j} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{h(\mu_{\mathrm{WW}j}) g(p_{\mathrm{WW}}; f^{\mathrm{C}}_{\mathrm{W}}(p'_{\mathrm{WW}}), \rho_{\mathrm{W}}^{-1}s^2) g(p_{\mathrm{WM}}; f^{\mathrm{C}}_{\mathrm{M}}(p'_{\mathrm{WM}}), \rho_{\mathrm{M}}^{-1}s^2) \\
&\qquad + (1-h(\mu_{\mathrm{WW}j})) g(p_{\mathrm{WW}}; f^{\mathrm{D}}_{\mathrm{W}}(p'_{\mathrm{WW}}), \rho_{\mathrm{W}}^{-1}s^2) g(p_{\mathrm{WM}}; f^{\mathrm{D}}_{\mathrm{M}}(p'_{\mathrm{WM}}), \rho_{\mathrm{M}}^{-1}s^2)\} \\
&\qquad \times g(p'_{\mathrm{WW}}; \mu_{\mathrm{WW}j}, \rho_{\mathrm{W}}^{-1}\sigma^2_{\mathrm{WW}j}) g(p'_{\mathrm{WM}}; \mu_{\mathrm{WM}j}, \rho_{\mathrm{M}}^{-1}\sigma^2_{\mathrm{WM}j}) \mathrm{d}p'_{\mathrm{WW}} \mathrm{d}p'_{\mathrm{WM}}, \\
&= \sum_j q_{\mathrm{W}j} \{h(\mu_{\mathrm{WW}j}) g(p_{\mathrm{WW}}; f^{\mathrm{C}}_{\mathrm{W}}(\mu_{\mathrm{WW}j}), \rho_{\mathrm{W}}^{-1}(s^2 + (\Delta f^{\mathrm{C}}_{\mathrm{W}})^2\sigma^2_{\mathrm{WW}j})) g(p_{\mathrm{WM}}; f^{\mathrm{C}}_{\mathrm{M}}(\mu_{\mathrm{WM}j}), \rho_{\mathrm{M}}^{-1}(s^2 + (\Delta f^{\mathrm{C}}_{\mathrm{M}})^2\sigma^2_{\mathrm{WM}j}))\} \\
&\qquad + (1-h(\mu_{\mathrm{WW}j})) g(p_{\mathrm{WW}}; f^{\mathrm{D}}_{\mathrm{W}}(\mu_{\mathrm{WW}j}), \rho_{\mathrm{W}}^{-1}(s^2 + (\Delta f^{\mathrm{D}}_{\mathrm{W}})^2\sigma^2_{\mathrm{WW}j})) g(p_{\mathrm{WM}}; f^{\mathrm{D}}_{\mathrm{M}}(\mu_{\mathrm{WM}j}), \rho_{\mathrm{M}}^{-1}(s^2 + (\Delta f^{\mathrm{D}}_{\mathrm{M}})^2\sigma^2_{\mathrm{WM}j}))\}. \qquad [16]
\end{aligned}
$$

This equation gives a constraint for $(q_{\mathrm{W}j}, \mu_{\mathrm{WW}j}, \sigma^2_{\mathrm{WW}j}, \mu_{\mathrm{WM}j}, \sigma^2_{\mathrm{WM}j})$. Next, when $A = \mathrm{M}$, in a similar manner, we obtain

$$
\begin{aligned}
&\sum_j q_{\mathrm{M}j} g(p_{\mathrm{MW}}; \mu_{\mathrm{MW}j}, \rho_{\mathrm{W}}^{-1}\sigma^2_{\mathrm{MW}j}) g(p_{\mathrm{MM}}; \mu_{\mathrm{MM}j}, \rho_{\mathrm{M}}^{-1}\sigma^2_{\mathrm{MM}j}) \\
&= \int_0^1 \int_0^1 \{h(p'_{\mathrm{WM}}) g(p_{\mathrm{MW}}; f^{\mathrm{C}}_{\mathrm{W}}(p'_{\mathrm{MW}}), \rho_{\mathrm{W}}^{-1}s^2) g(p_{\mathrm{MM}}; f^{\mathrm{C}}_{\mathrm{M}}(p'_{\mathrm{MM}}), \rho_{\mathrm{M}}^{-1}s^2) \\
&\qquad + (1-h(p'_{\mathrm{WM}})) g(p_{\mathrm{MW}}; f^{\mathrm{D}}_{\mathrm{W}}(p'_{\mathrm{MW}}), \rho_{\mathrm{W}}^{-1}s^2) g(p_{\mathrm{MM}}; f^{\mathrm{D}}_{\mathrm{M}}(p'_{\mathrm{MM}}), \rho_{\mathrm{M}}^{-1}s^2)\} \\
&\qquad \times \sum_j q_{\mathrm{M}j} g(p'_{\mathrm{MW}}; \mu_{\mathrm{MW}j}, \rho_{\mathrm{W}}^{-1}\sigma^2_{\mathrm{MW}j}) g(p'_{\mathrm{MM}}; \mu_{\mathrm{MM}j}, \rho_{\mathrm{M}}^{-1}\sigma^2_{\mathrm{MM}j}) \mathrm{d}p'_{\mathrm{MW}} \mathrm{d}p'_{\mathrm{MM}} \\
&= \sum_j q_{\mathrm{M}j} \int_0^1 \int_0^1 \{h(p'_{\mathrm{WM}}) g(p_{\mathrm{MW}}; f^{\mathrm{C}}_{\mathrm{W}}(p'_{\mathrm{MW}}), \rho_{\mathrm{W}}^{-1}s^2) g(p_{\mathrm{MM}}; f^{\mathrm{C}}_{\mathrm{M}}(p'_{\mathrm{MM}}), \rho_{\mathrm{M}}^{-1}s^2) \\
&\qquad + (1-h(p'_{\mathrm{WM}})) g(p_{\mathrm{MW}}; f^{\mathrm{D}}_{\mathrm{W}}(p'_{\mathrm{MW}}), \rho_{\mathrm{W}}^{-1}s^2) g(p_{\mathrm{MM}}; f^{\mathrm{D}}_{\mathrm{M}}(p'_{\mathrm{MM}}), \rho_{\mathrm{M}}^{-1}s^2)\} \\
&\qquad \times \sum_j q_{\mathrm{M}j} g(p'_{\mathrm{MW}}; \mu_{\mathrm{MW}j}, \rho_{\mathrm{W}}^{-1}\sigma^2_{\mathrm{MW}j}) g(p'_{\mathrm{MM}}; \mu_{\mathrm{MM}j}, \rho_{\mathrm{M}}^{-1}\sigma^2_{\mathrm{MM}j}) \mathrm{d}p'_{\mathrm{MW}} \mathrm{d}p'_{\mathrm{MM}} \\
&\simeq \sum_j q_{\mathrm{M}j} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{h(\mu_{\mathrm{WM}j}) g(p_{\mathrm{MW}}; f^{\mathrm{C}}_{\mathrm{W}}(p'_{\mathrm{MW}}), \rho_{\mathrm{W}}^{-1}s^2) g(p_{\mathrm{MM}}; f^{\mathrm{C}}_{\mathrm{M}}(p'_{\mathrm{MM}}), \rho_{\mathrm{M}}^{-1}s^2) \\
&\qquad + (1-h(\mu_{\mathrm{WM}j})) g(p_{\mathrm{MW}}; f^{\mathrm{D}}_{\mathrm{W}}(p'_{\mathrm{MW}}), \rho_{\mathrm{W}}^{-1}s^2) g(p_{\mathrm{MM}}; f^{\mathrm{D}}_{\mathrm{M}}(p'_{\mathrm{MM}}), \rho_{\mathrm{M}}^{-1}s^2)\} \\
&\qquad \times \sum_j q_{\mathrm{M}j} g(p'_{\mathrm{MW}}; \mu_{\mathrm{MW}j}, \rho_{\mathrm{W}}^{-1}\sigma^2_{\mathrm{MW}j}) g(p'_{\mathrm{MM}}; \mu_{\mathrm{MM}j}, \rho_{\mathrm{M}}^{-1}\sigma^2_{\mathrm{MM}j}) \mathrm{d}p'_{\mathrm{MW}} \mathrm{d}p'_{\mathrm{MM}} \\
&= \sum_j q_{\mathrm{M}j} \{h(\mu_{\mathrm{WM}j}) g(p_{\mathrm{MW}}; f^{\mathrm{C}}_{\mathrm{W}}(\mu_{\mathrm{MW}j}), \rho_{\mathrm{W}}^{-1}(s^2 + (\Delta f^{\mathrm{C}}_{\mathrm{W}})^2\sigma^2_{\mathrm{MW}j})) g(p_{\mathrm{MM}}; f^{\mathrm{C}}_{\mathrm{M}}(\mu_{\mathrm{MM}j}), \rho_{\mathrm{M}}^{-1}(s^2 + (\Delta f^{\mathrm{C}}_{\mathrm{M}})^2\sigma^2_{\mathrm{MM}j}))\} \\
&\qquad + (1-h(\mu_{\mathrm{WM}j})) g(p_{\mathrm{MW}}; f^{\mathrm{D}}_{\mathrm{W}}(\mu_{\mathrm{MW}j}), \rho_{\mathrm{W}}^{-1}(s^2 + (\Delta f^{\mathrm{D}}_{\mathrm{W}})^2\sigma^2_{\mathrm{MW}j})) g(p_{\mathrm{MM}}; f^{\mathrm{D}}_{\mathrm{M}}(\mu_{\mathrm{MM}j}), \rho_{\mathrm{M}}^{-1}(s^2 + (\Delta f^{\mathrm{D}}_{\mathrm{M}})^2\sigma^2_{\mathrm{MM}j}))\}. \qquad [17]
\end{aligned}
$$

This equation gives a constraint for $(q_{\mathrm{M}j}, \mu_{\mathrm{MW}j}, \sigma^2_{\mathrm{MW}j}, \mu_{\mathrm{MM}j}, \sigma^2_{\mathrm{MM}j})$.

To solve Eq. (16), let us consider a set of solutions $\{(\mu_{\mathrm{WW}j}, \mu_{\mathrm{WM}j})\}_j$. From the equilibrium condition of Eq. (16), the equal set must be restored by applying C-map and D-map to all the elements of the set. In other words, the condition is given by

$$
\{(\mu_{\mathrm{WW}j}, \mu_{\mathrm{WM}j})\}_j = \{(f^{\mathrm{C}}_{\mathrm{W}}(\mu_{\mathrm{WW}j}), f^{\mathrm{C}}_{\mathrm{M}}(\mu_{\mathrm{WM}j}))\}_j \cup \{(f^{\mathrm{D}}_{\mathrm{W}}(\mu_{\mathrm{WW}j}), f^{\mathrm{D}}_{\mathrm{M}}(\mu_{\mathrm{WM}j}))\}_j. \qquad [18]
$$

Similarly, to obtain the equilibrium condition for Eq. (17), we should consider a set of solutions $\{(\mu_{\mathrm{MW}j}, \mu_{\mathrm{MM}j})\}_j$ satisfying

$$\{(\mu_{\mathrm{MW}j}, \mu_{\mathrm{MM}j})\}_j = \{(f^{\mathrm{C}}_{\mathrm{W}}(\mu_{\mathrm{MW}j}), f^{\mathrm{C}}_{\mathrm{M}}(\mu_{\mathrm{MM}j}))\}_j \cup \{(f^{\mathrm{D}}_{\mathrm{W}}(\mu_{\mathrm{MW}j}), f^{\mathrm{D}}_{\mathrm{M}}(\mu_{\mathrm{MM}j}))\}_j. \quad [19]$$

Here, although the appearance of the variables is different, the problems are essentially between Eq. (18) and Eq. (19). Thus, the problem to be solved is

$$\{(\mu_{j,\mathrm{W}}, \mu_{j,\mathrm{M}})\}_j = \{(f^{\mathrm{C}}_{\mathrm{W}}(\mu_{j,\mathrm{W}}), f^{\mathrm{C}}_{\mathrm{M}}(\mu_{j,\mathrm{M}}))\}_j \cup \{(f^{\mathrm{D}}_{\mathrm{W}}(\mu_{j,\mathrm{W}}), f^{\mathrm{D}}_{\mathrm{M}}(\mu_{j,\mathrm{M}}))\}_j. \quad [20]$$

Furthermore, because $\mathrm{W}, \mathrm{M} \in \mathcal{S}$, we can generalize the problem as

$$\{(\mu_{j,S_{01}}, \cdots, \mu_{j,S_{16}})\}_j = \{(f^{\mathrm{C}}_{S_{01}}(\mu_{j,S_{01}}), \cdots, f^{\mathrm{C}}_{S_{16}}(\mu_{j,S_{16}}))\}_j \cup \{(f^{\mathrm{D}}_{S_{01}}(\mu_{j,S_{01}}), \cdots, f^{\mathrm{D}}_{S_{16}}(\mu_{j,S_{16}}))\}_j. \quad [21]$$

Now, for all $S_k \in \mathcal{S}$ and, let us consider a set $\{\mu_{j,S_k}\}_{j\in\mathbb{Z}\setminus\{0\}}$ satisfying

$$\mu_{+1,S_k} = f^{\mathrm{C}}_{S_k}(\mu_{-1,S_k}) = f^{\mathrm{C}}_{S_k}(\mu_{-2,S_k}) = \cdots, \quad [22]$$
$$\mu_{+(j+1),S_k} = f^{\mathrm{C}}_{S_k}(\mu_{+j,S_k}), \quad [23]$$
$$\mu_{-1,S_k} = f^{\mathrm{D}}_{S_k}(\mu_{+1,S_k}) = f^{\mathrm{D}}_{S_k}(\mu_{+2,S_k}) = \cdots, \quad [24]$$
$$\mu_{-(j+1),S_k} = f^{\mathrm{D}}_{S_k}(\mu_{-j,S_k}), \quad [25]$$

(the proof for these equations will be given later). This set $\{\mu_{j,S_k}\}_{j\in\mathbb{Z}\setminus\{0\}}$ gives a solution to problem Eq. (21), and thus solves Eq. (16) and Eq. (17). In Eq. (22)-Eq. (25), we consistently label each μ_{j,S_k} such that sequentially applying C-map (resp. D-map) $j(>0)$ times leads to label $+j$ (resp. $-j$) (see the illustration in Fig. S1). In the following, we show that such $\{\mu_{j,S_k}\}_{j\in\mathbb{Z}\setminus\{0\}}$ actually exists for all k.

1. When neither C-map nor D-map is constant: First, we consider a case of $\Delta f^{\mathrm{C}}_{S_k} \neq 0$ and $\Delta f^{\mathrm{D}}_{S_k} \neq 0$. Four norms of $k = 06, 07, 09, 10$ correspond to this case. In this case, C-map and D-map have the same fixed point (at $1/2$). Only the position given by this fixed point is achieved at equilibrium. Indeed, if we substitute $\mu_{j,S_k} = 1/2$ for all $j \in \mathbb{Z}\setminus\{0\}$, Eq. (22)-Eq. (25) are simultaneously satisfied without any contradiction.

2. When both C-map and D-map are constant: Second, we consider a case of $\Delta f^{\mathrm{C}}_{S_k} = 0$ and $\Delta f^{\mathrm{D}}_{S_k} = 0$. Four norms of $k = 01, 04, 13, 16$ correspond to this case. Because $f^{\mathrm{C}}_{S_k}$ is a constant map, Eq. (22) and Eq. (23) are satisfied by substituting the mapped value of this map into μ_{+j,S_k} for all $j = 1, 2, \cdots$. In the same way, because $f^{\mathrm{D}}_{S_k}$ is a constant map, Eq. (24) and Eq. (25) are satisfied by substituting the mapped value into μ_{-j,S_k} for all $j = 1, 2, \cdots$. Thus, no contradiction occurs.

3. When only C-map is constant: Third, we consider a case of $\Delta f^{\mathrm{C}}_{S_k} = 0$ and $\Delta f^{\mathrm{D}}_{S_k} \neq 0$. Four norms $k = 02, 03, 14, 15$ correspond to this case. Because $f^{\mathrm{C}}_{S_k}$ is a constant map, Eq. (22) and Eq. (23) are satisfied by substituting the mapped value of this map into μ_{+j,S_k} for all $j = 1, 2, \cdots$. Then, we define μ_{-1,S_k} as the value to which D-map maps all the same value $\mu_{+1,S_k} = \mu_{+2,S_k} = \cdots$, and Eq. (24) is satisfied. Finally, we sequentially define $\mu_{-2,S_k}, \mu_{-3,S_k}, \cdots$ by applying D-map to μ_{-1,S_k} one by one. Thus, no contradiction occurs.

4. When only D-map is constant: Finally, we consider a case of $\Delta f^{\mathrm{C}}_{S_k} \neq 0$ and $\Delta f^{\mathrm{D}}_{S_k} = 0$. Four norms $k = 05, 08, 09, 12$ correspond to this case. Because $f^{\mathrm{D}}_{S_k}$ is a constant map, Eq. (24), Eq. (25) are satisfied by substituting the mapped value of this map into μ_{-j,S_k} for all $j = 1, 2, \cdots$. Then, we define μ_{+1,S_k} as the value to which D-map maps all the same value $\mu_{-1,S_k} = \mu_{-2,S_k} = \cdots$, and Eq. (24) is satisfied. Finally, we sequentially define $\mu_{+2,S_k}, \mu_{+3,S_k}, \cdots$ by applying C-map to μ_{+1,S_k} one by one. Thus, no contradiction occurs.

As summarized in Table S1, the set $\{\mu_{j,S_k}\}_{j\in\mathbb{Z}\setminus\{0\}}$ can be analytically described. Furthermore, we also define σ^2_{j,S_k} as the variance in Gaussian corresponding to the mean μ_{j,S_k}. Similarly to the mean values above, we solve the variances as

$$(\sigma^2_{\mathrm{WW}j}, \sigma^2_{\mathrm{WM}j}) = (\sigma^2_{\mathrm{MW}j}, \sigma^2_{\mathrm{MM}j}) = (\sigma^2_{j,\mathrm{W}}, \sigma^2_{j,\mathrm{M}}). \quad [26]$$

The recursion that the set $\{\sigma^2_{j,S_k}\}_{j\in\mathbb{Z}\setminus\{0\}}$ should satisfy is

$$\sigma^2_{+1,S_k} = s^2 + (\Delta f^{\mathrm{C}}_{S_k})^2\sigma^2_{-1,S_k} = s^2 + (\Delta f^{\mathrm{C}}_{S_k})^2\sigma^2_{-2,S_k} = \cdots, \quad [27]$$
$$\sigma^2_{+(j+1),S_k} = s^2 + (\Delta f^{\mathrm{C}}_{S_k})^2\sigma^2_{+j,S_k}, \quad [28]$$
$$\sigma^2_{-1,S_k} = s^2 + (\Delta f^{\mathrm{D}}_{S_k})^2\sigma^2_{+1,S_k} = s^2 + (\Delta f^{\mathrm{D}}_{S_k})^2\sigma^2_{+2,S_k} = \cdots, \quad [29]$$
$$\sigma^2_{-(j+1),S_k} = s^2 + (\Delta f^{\mathrm{D}}_{S_k})^2\sigma^2_{-j,S_k}, \quad [30]$$

and the solution exists for all S_k (see Table. S1 for the solution of these equations). Table. S1 shows $\{(\mu_{j,S_k}, \sigma^2_{j,S_k})\}_{j\in\mathbb{Z}\setminus\{0\}}$.

S_k	μ_{+j,S_k}	μ_{-j,S_k}	σ^2_{+j,S_k}	σ^2_{-j,S_k}
S_{01}	$1-e_2$	$1-e_2$	$\frac{e_2(1-e_2)}{N}$	$\frac{e_2(1-e_2)}{N}$
S_{02}	$1-e_2$	$\frac{1+(1-2e_2)^{j+1}}{2}$	$\frac{e_2(1-e_2)}{N}$	$\frac{1-(1-2e_2)^{2(j+1)}}{4N}$
S_{03}	$1-e_2$	$\frac{1-\{-(1-2e_2)\}^{j+1}}{2}$	$\frac{e_2(1-e_2)}{N}$	$\frac{1-(1-2e_2)^{2(j+1)}}{4N}$
S_{04}	$1-e_2$	e_2	$\frac{e_2(1-e_2)}{N}$	$\frac{e_2(1-e_2)}{N}$
S_{05}	$\frac{1+(1-2e_2)^{j+1}}{2}$	$1-e_2$	$\frac{1-(1-2e_2)^{2(j+1)}}{4N}$	$\frac{e_2(1-e_2)}{N}$
S_{07}	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4N}$	$\frac{1}{4N}$
S_{08}	$\frac{1-(1-2e_2)^{j+1}}{2}$	e_2	$\frac{1-(1-2e_2)^{2(j+1)}}{4N}$	$\frac{e_2(1-e_2)}{N}$
S_{09}	$\frac{1-\{-(1-2e_2)\}^{j+1}}{2}$	$1-e_2$	$\frac{1-(1-2e_2)^{2(j+1)}}{4N}$	$\frac{e_2(1-e_2)}{N}$
S_{12}	$\frac{1+\{-(1-2e_2)\}^{j+1}}{2}$	e_2	$\frac{1-(1-2e_2)^{2(j+1)}}{4N}$	$\frac{e_2(1-e_2)}{N}$
S_{13}	e_2	$1-e_2$	$\frac{e_2(1-e_2)}{N}$	$\frac{e_2(1-e_2)}{N}$
S_{14}	e_2	$\frac{1-(1-2e_2)^{j+1}}{2}$	$\frac{e_2(1-e_2)}{N}$	$\frac{1-(1-2e_2)^{2(j+1)}}{4N}$
S_{15}	e_2	$\frac{1+\{-(1-2e_2)\}^{j+1}}{2}$	$\frac{e_2(1-e_2)}{N}$	$\frac{1-(1-2e_2)^{2(j+1)}}{4N}$
S_{16}	e_2	e_2	$\frac{e_2(1-e_2)}{N}$	$\frac{e_2(1-e_2)}{N}$

Table S1. Analytical solution of Gaussian functions. We omit S_{06}, S_{10}, and S_{11} because the results are identical to those of S_{07}(SJ).

We also calculate a set of the masses of Gaussian functions, i.e., $\{q_{\mathrm{W}j}\}_{j\in\mathbb{Z}\setminus\{0\}}$ and $\{q_{\mathrm{M}j}\}_{j\in\mathbb{Z}\setminus\{0\}}$. By substituting the values in Table S1 into Eq. (16), we obtain the following relational expressions

$$q_{\mathrm{W}+1} = \sum_{j=1}^{\infty} h(\mu_{-j,\mathrm{W}}) q_{\mathrm{W}-j}, \tag{31}$$

$$q_{\mathrm{W}+j} = h(\mu_{+(j-1),\mathrm{W}}) q_{\mathrm{W}+(j-1)} \quad (j=2,\cdots,\infty), \tag{32}$$

$$q_{\mathrm{W}-1} = \sum_{j=1}^{\infty} (1-h(\mu_{+j,\mathrm{W}})) q_{\mathrm{W}+j}, \tag{33}$$

$$q_{\mathrm{W}-j} = (1-h(\mu_{-(j-1),\mathrm{W}})) q_{\mathrm{W}-(j-1)} \quad (j=2,\cdots,\infty). \tag{34}$$

Similarly, by substituting the values in Table S1 into Eq. (17), we obtain

$$q_{\mathrm{M}+1} = \sum_{j=1}^{\infty} h(\mu_{-j,\mathrm{M}}) q_{\mathrm{W}-j}, \tag{35}$$

$$q_{\mathrm{M}+j} = h(\mu_{+(j-1),\mathrm{M}}) q_{\mathrm{W}+(j-1)} \quad (j=2,\cdots,\infty), \tag{36}$$

$$q_{\mathrm{M}-1} = \sum_{j=1}^{\infty} (1-h(\mu_{+j,\mathrm{M}})) q_{\mathrm{W}+j}, \tag{37}$$

$$q_{\mathrm{M}-j} = (1-h(\mu_{-(j-1),\mathrm{M}})) q_{\mathrm{W}-(j-1)} \quad (j=2,\cdots,\infty). \tag{38}$$

Eq. (31)-Eq. (38) includes the infinite summations. Because these infinite summations cannot be analytically calculated, one should set a cutoff of the summations in the numerical calculation of Eq. (31)-Eq. (38).

Fig. 2-B in the main manuscript shows an example of $\Phi^*_{\mathrm{W}}(p_{\mathrm{WW}}, p_{\mathrm{WM}})$. In this example, the elements in $\{(\mu_{j,\mathrm{W}}, \mu_{j,\mathrm{M}})\}_{j\in\mathbb{Z}\setminus\{0\}}$ are all different for different j, and thus the labeling in this study is at least necessary for the description of the reputation structure. This figure also shows that the obtained analytical solutions well approximate simulations of the image matrix.

3. Calculation of expected payoff

In order to consider an evolutionary process, we derive expected payoffs of wild-types W and mutants M from joint probability distribution of goodnesses, i.e., Φ^*_{W} and Φ^*_{M}. In the limit that mutants are rare $\delta \to 0$, the expected payoffs of the wild-types

$u_{\rm W}$ and mutants $u_{\rm M}$ are given by

$$\begin{aligned} u_{\rm W} &= (b-c)\bar{p}_{\rm WW}, \\ u_{\rm M} &= b\bar{p}_{\rm MW} - c\bar{p}_{\rm WM}, \end{aligned} \qquad [39]$$

Here, $\bar{p}_{AA'}$ indicates the average goodnesses of $p_{AA'}$, i.e., described as

$$\begin{aligned} \bar{p}_{\rm WW} &= \int_0^1\int_0^1 p_{\rm WW}\Phi^*_{\rm W}(p_{\rm WW},p_{\rm WM})\mathrm{d}p_{\rm WW}\mathrm{d}p_{\rm WM}, \\ \bar{p}_{\rm WM} &= \int_0^1\int_0^1 p_{\rm WM}\Phi^*_{\rm W}(p_{\rm WW},p_{\rm WM})\mathrm{d}p_{\rm WW}\mathrm{d}p_{\rm WM}, \\ \bar{p}_{\rm MW} &= \int_0^1\int_0^1 p_{\rm MW}\Phi^*_{\rm M}(p_{\rm MW},p_{\rm MM})\mathrm{d}p_{\rm MW}\mathrm{d}p_{\rm MM}, \end{aligned} \qquad [40]$$

This average goodness can be analytically calculated by Gaussian approximation of $\Phi^*_A(p_{A\rm W},p_{A\rm M})$. According to the conditions for the ESS, mutants can invade the population of wild-types if $u_{\rm W} > u_{\rm M}$.

Regions, where a mutant norm can invade a wild-type norm, are given by Fig. S2. From this figure, we can obtain the invasibilities for each b/c, as shown in Fig. 3-A in the main manuscript.

mutant

wild-type	01	02	03	04	05	07	08	09	12	13	14	15	16
01		always	always	always	always	always	always	always	always	always	always	always	always
02	> 5.964		< 6.811	< 6.434	< 13.13	< 11.58	< 10.99	< 12.92	< 10.63	< 12.71	< 11.58	< 12.28	< 11.44
03	> 4.376	> 6.352		< 1.920	never	< 1.306	< 1.441	< 1.134	< 1.457	< 1.261	< 1.402	< 1.306	< 1.424
04	> 1.250	> 1.250	> 1.250		> 1.250		< 1.250	> 1.250	< 1.250		< 1.250	< 1.250	< 1.250
05	never	never	never	< 2.789		always	always	always	always	always	always	always	always
07	never	never	never		never		always	never	always		always	always	always
08	> 11.44	> 10.99	> 10.63	> 6.434	> 11.58	> 11.58		> 12.28	> 6.811	> 12.71	> 13.13	> 12.92	< 5.964
09	never	never	always	always	never	always	always		always	always	always	always	always
12	> 1.424	> 1.441	> 1.457	> 1.920	> 1.402	> 1.306	< 6.352	> 1.306		> 1.261	always	> 1.134	< 4.376
13	never	never	never		never		always	never	always		always	always	always
14	never	never	never	> 2.789	never	never	always	never	always	never		never	always
15	never	never	never	never	never	never	always	never	never	never	always		always
16	never	never	never	never	never	never	never	never	never	never	never	never	

Fig. S2. Regions for possible invasions in the evolutionary processes. The row and column indicate the wild-type and mutant norms, respectively. The matrix shows the region of $b/c(>1)$ where the mutant invades the wild-type. In some pairs of wild-type and mutant norms, the mutant always or never succeeds in invading the wild-types for all $b/c(>1)$. The calculation is based on $e_1 = 0$ and $e_2 = 0.1$.

4. Calculation of equilibrium state in public reputation

In this section, we derive the equilibrium distribution of reputations under public assessment, based on the previous study (1). The basic setting is the same whether the reputation is publicly shared or privately held. We assume a population of size N which consists of mutants with norm M and wild-type individuals with norm W $\neq$ M. A donor and a recipient are randomly chosen every round. The donor chooses cooperation to the good recipient and defection to the bad recipient. Here, the donor erroneously chooses the opposite action to the intended one with probability $0 \leq e_1 < 1/2$. Then, all the individuals update their reputations of the donor. The difference between the public and private reputation cases is seen in the observers' ways to update reputations. We assume that one mutant observer and one wild-type observer are chosen as representatives of each norm, and each gives a good or bad reputation to the donor according to its norm. Here, each representative observer commits an assignment error independently, in which case it erroneously assigns the opposite reputation to the intended one with probability $0 < e_2 < 1/2$. (such an assessment error was not assumed in (1)) Then, all the individuals with the same norm copy the reputation of the donor assigned by their representative. Thus, the reputation of the same individual, even an erroneously assigned one, is shared among all the individuals with the same norm. In other words, each individual at any given time has two reputations, one is shared by all the mutant individuals, and the other is shared by all the wild-type individuals in the population.

Here we specifically consider the situation where rare mutants with norm M $= S_{16}$(ALLB) invades a wild-type population with norm W $\neq$ M. We use the same definition of $p_{AA'}$, i.e., goodness of an individual with norm A in the eyes of norm A' users. Because reputations are public, $p_{AA'}$ can be either 1 (the individual is assigned as good from all) or 0 (the individual is assigned as bad from all). Below we will derive $\bar{p}_{AA'}$, the probability that a norm A user has a good reputation in the eyes of norm A' users.

Since the mutant norm is ALLB, the probability that mutants assign a good reputation to the donor is always $a_{\rm M}^{\rm GC} = a_{\rm M}^{\rm BC} = a_{\rm M}^{\rm GD} = a_{\rm M}^{\rm BD} = e_2$. Thus we obtain $\bar{p}_{\rm WM} = \bar{p}_{\rm MM} = e_2$.

Next we aim to solve the equilibrium average goodnesses in the eyes of wild-types, $\bar{p}_{\rm WW}$ and $\bar{p}_{\rm MW}$. First, let us calculate $\bar{p}_{\rm WW}$, which is relevant when the donor and the observer use norm W. Note that we can assume that the recipient uses norm W, because mutants are rare. $\bar{p}_{\rm WW}$ should satisfy

$$\bar{p}_{\rm WW} = \bar{p}_{\rm WW}\{(1-e_1)a_{\rm W}^{\rm GC} + e_1 a_{\rm W}^{\rm GD}\} + (1-\bar{p}_{\rm WW})\{e_1 a_{\rm W}^{\rm BC} + (1-e_1)a_{\rm W}^{\rm BD}\}, \quad [41]$$

The equality between the left- and right-hand sides shows that the proportion of good individuals balances before and after updating the chosen donor's reputation. In the right-hand side, $\bar{p}_{\rm WW}$ and $1-\bar{p}_{\rm WW}$ in the first and the second terms indicate the probabilities that a randomly chosen recipient of norm W is good or bad from the viewpoint of norm W, respectively. When the recipient is good, the donor chooses cooperation or defection with probabilities $(1-e_1)$ and e_1. Then, $a_{\rm W}^{\rm GC}$ and $a_{\rm W}^{\rm GD}$ indicate the probabilities that the cooperating or defecting donor receives a good reputation from observers of norm W. When the recipient is bad, the donor chooses cooperation or defection with probabilities e_1 and $(1-e_1)$. Then, $a_{\rm W}^{\rm BC}$ and $a_{\rm W}^{\rm BD}$ indicate the probabilities that the cooperating or defecting donor receives a good reputation from observers of norm W. The solution is

$$\bar{p}_{\rm WW} = \frac{(1-e_1)a_{\rm W}^{\rm BD} + e_1 a_{\rm W}^{\rm BC}}{1-\{(1-e_1)(a_{\rm W}^{\rm GC} - a_{\rm W}^{\rm BD}) + e_1(a_{\rm W}^{\rm GD} - a_{\rm W}^{\rm BC})\}}. \quad [42]$$

Second, let us calculate $\bar{p}_{\rm MW}$, which is relevant when the donor uses norm M and the observer uses norm W. Note that we can once again assume that the recipient uses norm W because mutants are rare. $\bar{p}_{\rm MW}$ should satisfy

$$\bar{p}_{\rm MW} = \bar{p}_{\rm WW}\{h(e_2)a_{\rm W}^{\rm GC} + (1-h(e_2))a_{\rm W}^{\rm GD}\} + (1-\bar{p}_{\rm WW})\{h(e_2)a_{\rm W}^{\rm BC} + (1-h(e_2))a_{\rm W}^{\rm BD}\}. \quad [43]$$

Here, $\bar{p}_{\rm WW}$ and $(1-\bar{p}_{\rm WW})$ in the first and second terms of the right-hand side indicate the probabilities that the recipient is good or bad from the viewpoint of norm W, respectively. In both terms, $h(e_2)(= e_2(1-e_1) + (1-e_2)e_1)$ and $1-h(e_2)$ are the probabilities that the donor with norm M executes cooperation or defection, which is independent of whether the recipient is good or bad from the viewpoint of norm W. In the first term, $a_{\rm W}^{\rm GC}$ and $a_{\rm W}^{\rm GD}$ indicate the probabilities that the cooperating and defecting donor receives a good reputation from the observers of norm W. In the second term, $a_{\rm W}^{\rm BC}$ and $a_{\rm W}^{\rm BD}$ indicate the probabilities that the cooperating and defecting donor receives a good reputation from the observers of norm W.

We summarize the solutions, $\bar{p}_{\rm WW}$ and $\bar{p}_{\rm MW}$, in Table S2.

S_k	$\bar{p}_{\mathrm{WW}}$		$\bar{p}_{\mathrm{MW}}$	$\bar{p}_{\mathrm{WW}}\|_{e_1=0}$	$\bar{p}_{\mathrm{MW}}\|_{e_1=0}$
S_{01}	$1-e_2$	$=$	$1-e_2$	$1-e_2$	$1-e_2$
S_{02}	$\frac{e_1+e_2-2e_1e_2}{e_1+2e_2-2e_1e_2}$	$<$	$\frac{e_1+e_2-2e_1e_2+e_2^2-2e_1e_2^2-2e_2^3+4e_1e_2^3}{e_1+2e_2-2e_1e_2}$	$\frac{1}{2}$	$\frac{1}{2}+\frac{1}{2}e_2-e_2^2$
S_{03}	$\frac{1-e_2}{1+e_1-2e_1e_2}$	$>$	$\frac{(1-e_2)(2e_1+3e_2-6e_1e_2-2e_2^2+4e_1e_2^2)}{1+e_1-2e_1e_2}$	$1-e_2$	$3e_2-5e_2^2+2e_2^3$
S_{04}	$\frac{1}{2}$	$>$	$e_1+2e_2-4e_1e_2-2e_2^2+4e_1e_2^2$	$\frac{1}{2}$	$2e_2-2e_2^2$
S_{05}	$\frac{1-e_1-e_2+2e_1e_2}{1-e_1+2e_1e_2}$	$>$	$\frac{1-e_1-e_2+2e_1e_2-e_2^2+2e_1e_2^2+2e_2^3-4e_1e_2^3}{1-e_1+2e_1e_2}$	$1-e_2$	$1-e_2-e_2^2+2e_2^3$
S_{06}	$\frac{1}{2}$	$=$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
S_{07}	$1-e_1-e_2+2e_1e_2$	$>$	$2e_1+3e_2-2e_1^2-12e_1e_2-6e_2^2+12e_1^2e_2+24e_1e_2^2+4e_2^3-24e_1^2e_2^2-16e_1e_2^3+16e_1^2e_2^3$	$1-e_2$	$3e_2-6e_2^2+4e_2^3$
S_{08}	$\frac{e_2}{e_1+2e_2-2e_1e_2}$	$>$	$\frac{e_2(2e_1+3e_2-6e_1e_2-2e_2^2+4e_1e_2^2)}{e_1+2e_2-2e_1e_2}$	$\frac{1}{2}$	$\frac{3}{2}e_2-e_2^2$
S_{09}	$\frac{1-e_2}{2-e_1-2e_2+2e_1e_2}$	$<$	$\frac{(1-e_2)(2-2e_1-3e_2+6e_1e_2+2e_2^2-4e_1e_2^2)}{2-e_1-2e_2+2e_1e_2}$	$\frac{1}{2}$	$1-\frac{3}{2}e_2+e_2^2$
S_{10}	$e_1+e_2-2e_1e_2$	$<$	$2e_1+3e_2-2e_1^2-12e_1e_2-6e_2^2+12e_1^2e_2+24e_1e_2^2+4e_2^3-24e_1^2e_2^2-16e_1e_2^3+16e_1^2e_2^3$	e_2	$3e_2-6e_2^2+4e_2^3$
S_{11}	$\frac{1}{2}$	$=$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
S_{12}	$\frac{e_1+e_2-2e_1e_2}{1+e_1-2e_1e_2}$	$<$	$\frac{e_1+2e_2-4e_1e_2-3e_2^2+6e_1e_2^2+2e_2^3-4e_1e_2^3}{1+e_1-2e_1e_2}$	e_2	$2e_2-3e_2^2+2e_2^3$
S_{13}	$\frac{1}{2}$	$<$	$1-e_1-2e_2+4e_1e_2+2e_2^2-4e_1e_2^2$	$\frac{1}{2}$	$1-2e_2+2e_2^2$
S_{14}	$\frac{e_2}{1-e_1+2e_1e_2}$	$<$	$\frac{e_2(2-2e_1-3e_2+6e_1e_2+2e_2^2-4e_1e_2^2)}{1-e_1+2e_1e_2}$	e_2	$2e_2-3e_2^2+2e_2^3$
S_{15}	$\frac{1-e_1-e_2+2e_1e_2}{2-e_1-2e_2+2e_1e_2}$	$>$	$\frac{1-e_1-2e_2+4e_1e_2+3e_2^2-6e_1e_2^2-2e_2^3+4e_1e_2^3}{2-e_1-2e_2+2e_1e_2}$	$\frac{1}{2}$	$\frac{1}{2}-\frac{1}{2}e_2+e_2^2$
S_{16}	e_2	$=$	e_2	e_2	e_2

Table S2. Analytical solution of reputation structure under public assessment. Here, the equality (i.e., $=$) and inequality (i.e., $>$ or $<$) signs show the relations between $\bar{p}_{\mathrm{WW}}$ and $\bar{p}_{\mathrm{MW}}$ for all of $0 \leq e_1 < 1/2$ and $0 < e_2 < 1/2$.

Based on Table S2, we can see how the reputation structure differs between the public and private reputation cases. The reputation structure for norms S_{03}(SS) and S_{07}(SJ) are illustrated in Fig. 3-D and E in the main manuscript, while that of S_{08}(SH) is in Fig. S3.

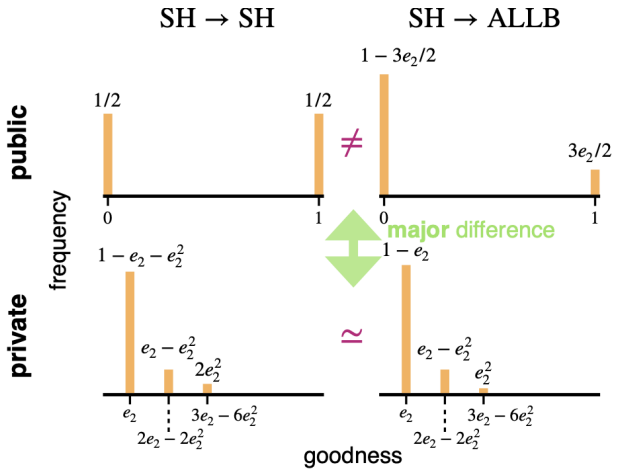

Fig. S3. Illustration of how the wild-type SH gives reputations to the self and mutant ALLB norms. In each panel, the horizontal and vertical axes indicate the goodness and its frequency, respectively. Positions and heights of bars are correct only up to order e_2^2. By comparing the upper panels with the lower ones, we can see that the reputation from SH differs significantly between the public and private reputation cases. In the private reputation case, SH still manages to distinguish the self norm with ALLB, but only with the difference of the order of e_2^2.

5. Numerical algorithm and error estimate

This section provides how to computationally calculate Eq. (31)-Eq. (34) and Eq. (35)-Eq. (38) with sufficient accuracy.

Instead of Eq. (31)-Eq. (34), we aim to compute

$$Q_{\mathrm{W}+1} := 1, \tag{44}$$
$$Q_{\mathrm{W}+j} := h(\mu_{+(j-1),\mathrm{W}})Q_{\mathrm{W}+(j-1)} \quad (j = 2, \cdots, \infty), \tag{45}$$
$$Q_{\mathrm{W}-1} := \sum_{j=1}^{\infty}(1 - h(\mu_{+j,\mathrm{W}}))Q_{\mathrm{W}+j}, \tag{46}$$
$$Q_{\mathrm{W}-j} := (1 - h(\mu_{-(j-1),\mathrm{W}}))Q_{\mathrm{W}-(j-1)} \quad (j = 2, \cdots, \infty), \tag{47}$$

(see Fig. S4 for the illustration of this computation). Via these equations, we obtain $q_{\mathrm{W}j}$ by rescaling $Q_{\mathrm{W}j}$ as

$$q_{\mathrm{W}j} = \frac{Q_{\mathrm{W}j}}{\sum_{k=\pm1}^{\pm\infty} Q_{\mathrm{W}k}}, \tag{48}$$

which satisfies Eq. (31)-Eq. (34). We should also obtain average goodnesses

$$\bar{p}_{\mathrm{W}A} = \frac{\sum_{j=\pm1}^{\pm\infty} Q_{\mathrm{W}j}\mu_{j,A}}{\sum_{j=\pm1}^{\pm\infty} Q_{\mathrm{W}j}}, \tag{49}$$

in order to obtain Fig. S2.

In a practical computer simulation, we approximate Eq. (44)-Eq. (47) by

$$\hat{Q}_{\mathrm{W}+1} := 1, \tag{50}$$
$$\hat{Q}_{\mathrm{W}+j} := h(\mu_{+(j-1),\mathrm{W}})\hat{Q}_{\mathrm{W}+(j-1)} \quad (j = 2, \cdots, j_{\max}), \tag{51}$$
$$\hat{Q}_{\mathrm{W}+j} := 0 \quad (j = j_{\max} + 1, \cdots, \infty), \tag{52}$$
$$\hat{Q}_{\mathrm{W}-1} := \sum_{j=1}^{\infty}(1 - h(\mu_{+j,\mathrm{W}}))\hat{Q}_{\mathrm{W}+j} = \sum_{j=1}^{j_{\max}}(1 - h(\mu_{+j,\mathrm{W}}))\hat{Q}_{\mathrm{W}+j}, \tag{53}$$
$$\hat{Q}_{\mathrm{W}-j} := (1 - h(\mu_{-(j-1),\mathrm{W}}))\hat{Q}_{\mathrm{W}-(j-1)} \quad (j = 2, \cdots, j_{\max}), \tag{54}$$
$$\hat{Q}_{\mathrm{W}-j} := 0 \quad (j = j_{\max} + 1, \cdots, \infty), \tag{55}$$

with sufficient large $j_{\max}(=10^4)$ (see Fig. S4 for the illustration of this computation). We will show below that these computationally obtained $\hat{Q}_{\mathrm{W}j}$ well approximate $Q_{\mathrm{W}j}$. Note that in the following calculations we use the fact that

$$e_2 \leq \mu_{j,A} \leq 1 - e_2 \tag{56}$$

holds for all $j = \pm 1, \cdots, \pm\infty$ and A.

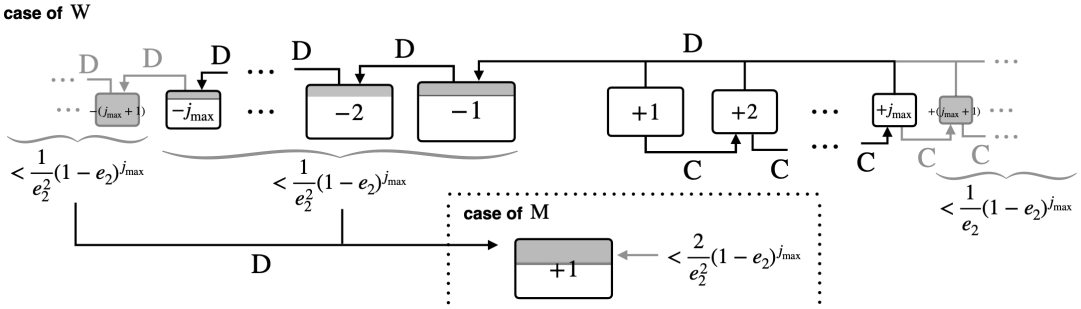

Fig. S4. An illustration of numerical algorithm and error estimation of the masses. The black and gray arrows show how theoretical calculations of Eq. (44)-Eq. (47) are performed, whereas only black arrows are relevant in the computation of Eq. (50)-Eq. (55). Each box shows the size of $Q_{\mathrm{W}j}$. The gray part in each box shows the size of approximation error, $Q_{\mathrm{W}j} - \hat{Q}_{\mathrm{W}j}$. Apart from $Q_{\mathrm{W}j}$, the area surrounded by dots show the calculation of $Q_{\mathrm{M}+1}$.

From the definition, we obtain

$$Q_{\mathrm{W}+j} - \hat{Q}_{\mathrm{W}+j} = 0 \quad (j = 1, \cdots, j_{\max}). \tag{57}$$

Then, we obtain

$$Q_{\mathrm{W}+j} = Q_{\mathrm{W}+1} \prod_{k=1}^{j-1} \mu_{+k,\mathrm{W}} \leq (1-e_2)^{j-1}, \tag{58}$$

$$\Rightarrow \sum_{j=j_{\max}+1}^{\infty} (Q_{\mathrm{W}+j} - \hat{Q}_{\mathrm{W}+j}) = \sum_{j=j_{\max}+1}^{\infty} Q_{\mathrm{W}+j} \leq \sum_{j=j_{\max}+1}^{\infty} (1-e_2)^{j-1} = \frac{1}{e_2}(1-e_2)^{j_{\max}}. \tag{59}$$

Then, we have

$$Q_{\mathrm{W}-1} = \sum_{j=1}^{\infty} Q_{\mathrm{W}+j}(1-\mu_{+j,\mathrm{W}}) = \underbrace{\sum_{j=1}^{j_{\max}} Q_{\mathrm{W}+j}(1-\mu_{+j,\mathrm{W}})}_{=\hat{Q}_{\mathrm{W}-1}} + \sum_{j=j_{\max}+1}^{\infty} Q_{\mathrm{W}+j}(1-\mu_{+j,\mathrm{W}}) \leq \hat{Q}_{\mathrm{W}-1} + \frac{1}{e_2}(1-e_2)^{j_{\max}}, \tag{60}$$

$$\Rightarrow Q_{\mathrm{W}-j} = Q_{\mathrm{W}-1} \prod_{k=1}^{j-1}(1-\mu_{-k,\mathrm{W}}) = \underbrace{\hat{Q}_{\mathrm{W}-1} \prod_{k=1}^{j-1}(1-\mu_{-k,\mathrm{W}})}_{=\hat{Q}_{\mathrm{W}-j}} + (Q_{\mathrm{W}-1} - \hat{Q}_{\mathrm{W}-1}) \prod_{k=1}^{j-1}(1-\mu_{-k,\mathrm{W}}) \leq \hat{Q}_{\mathrm{W}-j} + \frac{1}{e_2}(1-e_2)^{j_{\max}+j-1}, \tag{61}$$

$$\Rightarrow \sum_{j=1}^{j_{\max}} (Q_{\mathrm{W}-j} - \hat{Q}_{\mathrm{W}-j}) \leq \frac{1}{e_2^2}(1-e_2)^{j_{\max}}, \tag{62}$$

We also obtain

$$Q_{\mathrm{W}-1} = \sum_{j=1}^{\infty} Q_{\mathrm{W}+j}(1-\mu_{+j,\mathrm{W}}) \leq \sum_{j=1}^{\infty} Q_{\mathrm{W}+j} \leq \frac{1}{e_2}, \tag{63}$$

$$\Rightarrow Q_{\mathrm{W}-j} = Q_{\mathrm{W}-1} \prod_{k=1}^{j-1}(1-\mu_{-k,\mathrm{W}}) \leq \frac{1}{e_2}(1-e_2)^{j-1}, \tag{64}$$

$$\Rightarrow \sum_{j=j_{\max}+1}^{\infty} (Q_{\mathrm{W}-j} - \hat{Q}_{\mathrm{W}-j}) = \sum_{j=j_{\max}+1}^{\infty} Q_{\mathrm{W}-j} \leq \sum_{j=j_{\max}+1}^{\infty} \frac{1}{e_2}(1-e_2)^{j-1} = \frac{1}{e_2^2}(1-e_2)^{j_{\max}}. \tag{65}$$

From the above error estimations, we can obtain upper and lower bounds of Eq. (48) and Eq. (49) as

$$\frac{\hat{Q}_{\mathrm{W}j}}{\sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{W}j} + \frac{3}{e_2^2}(1-e_2)^{j_{\max}}} \leq q_{\mathrm{W}j} \leq \frac{\hat{Q}_{\mathrm{W}j} + \frac{1}{e_2}(1-e_2)^{j_{\max}}}{\sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{W}j}}. \tag{66}$$

$$\frac{\sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{W}j}\mu_{j,A}}{\sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{W}j} + \frac{3}{e_2^2}(1-e_2)^{j_{\max}}} \leq \bar{p}_{\mathrm{W}A} \leq \frac{\sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{W}j}\mu_{j,A} + \frac{3}{e_2^2}(1-e_2)^{j_{\max}}}{\sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{W}j}}. \tag{67}$$

Here, we used

$$\hat{Q}_{\mathrm{W}j} \leq Q_{\mathrm{W}j} = \hat{Q}_{\mathrm{W}j} + (Q_{\mathrm{W}j} - \hat{Q}_{\mathrm{W}j}) \tag{68}$$

$$\leq \hat{Q}_{\mathrm{W}j} + \underbrace{\max_j (Q_{\mathrm{W}j} - \hat{Q}_{\mathrm{W}j})}_{=Q_{\mathrm{W}-1} - \hat{Q}_{\mathrm{W}-1} \leq \frac{1}{e_2}(1-e_2)^{j_{\max}}} \leq \hat{Q}_{\mathrm{W}j} + \frac{1}{e_2}(1-e_2)^{j_{\max}}, \tag{69}$$

$$\sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{W}j} \leq \sum_{j=\pm 1}^{\pm \infty} Q_{\mathrm{W}j} = \sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{W}j} + \sum_{j=\pm 1}^{\pm \infty} (Q_{\mathrm{W}j} - \hat{Q}_{\mathrm{W}j}) \tag{70}$$

$$= \sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{W}j} + \underbrace{\sum_{j=1}^{j_{\max}+1} (Q_{\mathrm{W}j} - \hat{Q}_{\mathrm{W}-j})}_{\leq \frac{1}{e_2^2}(1-e_2)^{j_{\max}}} + \underbrace{\sum_{j=j_{\max}+1}^{\infty} (Q_{\mathrm{W}+j} - \hat{Q}_{\mathrm{W}+j})}_{\leq \frac{1}{e_2}(1-e_2)^{j_{\max}}} + \underbrace{\sum_{j=j_{\max}+1}^{\infty} (Q_{\mathrm{W}-j} - \hat{Q}_{\mathrm{W}-j})}_{\leq \frac{1}{e_2^2}(1-e_2)^{j_{\max}}} \tag{71}$$

$$\leq \sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{W}j} + \frac{3}{e_2^2}(1-e_2)^{j_{\max}}, \tag{72}$$

$$\sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{W}j}\mu_{j,A} \leq \sum_{j=\pm 1}^{\pm \infty} Q_{\mathrm{W}j}\mu_{j,A} \leq \sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{W}j}\mu_{j,A} + \sum_{j=\pm 1}^{\pm \infty} (Q_{\mathrm{W}j} - \hat{Q}_{\mathrm{W}j}) \tag{73}$$

$$\leq \sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{W}j}\mu_{j,A} + \frac{3}{e_2^2}(1-e_2)^{j_{\max}}. \tag{74}$$

In the same way, let us consider Eq. (35)-Eq. (38) and compute

$$Q_{\mathrm{M}+1} := \sum_{j=1}^{\infty} h(\mu_{-j,\mathrm{M}}) Q_{\mathrm{W}-j}, \tag{75}$$

$$Q_{\mathrm{M}+j} := h(\mu_{+(j-1),\mathrm{M}}) Q_{\mathrm{W}+(j-1)} \quad (j = 2, \cdots, \infty), \tag{76}$$

$$Q_{\mathrm{M}-1} := \sum_{j=1}^{\infty} (1 - h(\mu_{+j,\mathrm{M}})) Q_{\mathrm{W}+j}, \tag{77}$$

$$Q_{\mathrm{M}-j} := (1 - h(\mu_{-(j-1),\mathrm{M}})) Q_{\mathrm{W}-(j-1)} \quad (j = 2, \cdots, \infty). \tag{78}$$

Via these equations, we obtain $q_{\mathrm{W}j}$ by rescaling $Q_{\mathrm{M}j}$ as

$$q_{\mathrm{M}j} = \frac{Q_{\mathrm{M}j}}{\sum_{k=\pm 1}^{\pm \infty} Q_{\mathrm{M}k}}, \tag{79}$$

which satisfies Eq. (35)-Eq. (38). We also need to obtain average goodnesses

$$\bar{p}_{\mathrm{M}A} = \frac{\sum_{j=\pm 1}^{\pm \infty} Q_{\mathrm{M}j}\mu_{j,A}}{\sum_{j=\pm 1}^{\pm \infty} Q_{\mathrm{M}j}}, \tag{80}$$

in order to obtain Fig. S2.

In a practical computer simulation, we approximate Eq. (75)-Eq. (78) by

$$\hat{Q}_{\mathrm{M}+1} := \sum_{j=1}^{\infty} h(\mu_{-j,\mathrm{M}})\hat{Q}_{\mathrm{W}-j} = \sum_{j=1}^{j_{\max}} h(\mu_{-j,\mathrm{M}})\hat{Q}_{\mathrm{W}-j}, \tag{81}$$

$$\hat{Q}_{\mathrm{M}+j} := h(\mu_{+(j-1),\mathrm{M}})\hat{Q}_{\mathrm{W}+(j-1)} \quad (j = 2, \cdots, j_{\max}), \tag{82}$$

$$\hat{Q}_{\mathrm{M}+j} := 0 \quad (j = j_{\max}+1, \cdots, \infty), \tag{83}$$

$$\hat{Q}_{\mathrm{M}-1} := \sum_{j=1}^{\infty}(1-h(\mu_{+j,\mathrm{M}}))\hat{Q}_{\mathrm{W}+j} = \sum_{j=1}^{j_{\max}}(1-h(\mu_{+j,\mathrm{M}}))\hat{Q}_{\mathrm{W}+j}, \tag{84}$$

$$\hat{Q}_{\mathrm{M}-j} := (1-h(\mu_{-(j-1),\mathrm{M}}))\hat{Q}_{\mathrm{W}-(j-1)} \quad (j = 2, \cdots, j_{\max}), \tag{85}$$

$$\hat{Q}_{\mathrm{M}-j} := 0 \quad (j = j_{\max}+1, \cdots, \infty), \tag{86}$$

with sufficient large $j_{\max}(= 10^4)$.

By exactly similar calculations, we obtain

$$Q_{\mathrm{M}+j} - \hat{Q}_{\mathrm{M}+j} = 0 \quad (j = 2, \cdots, j_{\max}), \tag{87}$$

$$\sum_{j=j_{\max}+1}^{\infty} (Q_{\mathrm{M}+j} - \hat{Q}_{\mathrm{M}+j}) \leq \frac{1}{e_2}(1-e_2)^{j_{\max}}, \tag{88}$$

$$\sum_{j=1}^{j_{\max}} (Q_{\mathrm{M}-j} - \hat{Q}_{\mathrm{M}-j}) \leq \frac{1}{e_2^2}(1-e_2)^{j_{\max}}, \tag{89}$$

$$\sum_{j=j_{\max}+1}^{\infty} (Q_{\mathrm{M}-j} - \hat{Q}_{\mathrm{M}-j}) \leq \frac{1}{e_2^2}(1-e_2)^{j_{\max}}. \tag{90}$$

The difference from $Q_{\mathrm{W}j}$ exists only in $j = +1$, as

$$Q_{\mathrm{M}+1} = \sum_{j=1}^{\infty} Q_{\mathrm{W}-j}\mu_{-j,\mathrm{M}} = \underbrace{\sum_{j=1}^{j_{\max}} \hat{Q}_{\mathrm{W}-j}\mu_{-j,\mathrm{M}}}_{=\hat{q}_{\mathrm{M}+1}} + \underbrace{\sum_{j=1}^{j_{\max}} (Q_{\mathrm{W}-j} - \hat{Q}_{\mathrm{W}-j})\mu_{-j,\mathrm{M}}}_{\leq \frac{1}{e_2^2}(1-e_2)^{j_{\max}}} + \underbrace{\sum_{j=j_{\max}+1}^{\infty} \hat{Q}_{\mathrm{W}-j}\mu_{-j,\mathrm{M}}}_{\leq \frac{1}{e_2^2}(1-e_2)^{j_{\max}}} \tag{91}$$

$$\leq \hat{Q}_{\mathrm{M}+1} + \frac{2}{e_2^2}(1-e_2)^{j_{\max}}, \tag{92}$$

(see the area surrounded by dots in Fig. S4 for the illustration of this computation).

From the above error estimations, we can obtain upper and lower bounds of Eq. (79) and Eq. (80) as

$$\frac{\hat{Q}_{\mathrm{M}j}}{\sum_{j=\pm1}^{\pm j_{\max}} \hat{Q}_{\mathrm{M}j} + \frac{5}{e_2^2}(1-e_2)^{j_{\max}}} \leq q_{\mathrm{M}j} \leq \frac{\hat{Q}_{\mathrm{M}j} + \frac{2}{e_2^2}(1-e_2)^{j_{\max}}}{\sum_{j=\pm1}^{\pm j_{\max}} \hat{Q}_{\mathrm{M}j}}, \tag{93}$$

$$\frac{\sum_{j=\pm1}^{\pm j_{\max}} \hat{Q}_{\mathrm{M}j}\mu_{j,A}}{\sum_{j=\pm1}^{\pm j_{\max}} \hat{Q}_{\mathrm{M}j} + \frac{5}{e_2^2}(1-e_2)^{j_{\max}}} \leq \bar{p}_{\mathrm{M}A} \leq \frac{\sum_{j=\pm1}^{\pm j_{\max}} \hat{Q}_{\mathrm{M}j}\mu_{j,A} + \frac{5}{e_2^2}(1-e_2)^{j_{\max}}}{\sum_{j=\pm1}^{\pm j_{\max}} \hat{Q}_{\mathrm{M}j}}. \tag{94}$$

Here, we used

$$\hat{Q}_{\mathrm{M}j} \leq Q_{\mathrm{M}j} = \hat{Q}_{\mathrm{M}j} + (Q_{\mathrm{M}j} - \hat{Q}_{\mathrm{M}j}) \quad [95]$$

$$\leq \hat{Q}_{\mathrm{M}j} + \underbrace{\max_j (Q_{\mathrm{M}j} - \hat{Q}_{\mathrm{M}j})}_{=Q_{\mathrm{M}+1} - \hat{Q}_{\mathrm{M}+1} \leq \frac{1}{e_2^2}(1-e_2)^{j_{\max}}} \leq \hat{Q}_{\mathrm{M}j} + \frac{1}{e_2^2}(1-e_2)^{j_{\max}}, \quad [96]$$

$$\sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{M}j} \leq \sum_{j=\pm 1}^{\pm \infty} Q_{\mathrm{M}j} = \sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{M}j} + \underbrace{(Q_{\mathrm{M}+1} - \hat{Q}_{\mathrm{M}+1})}_{\leq \frac{2}{e_2^2}(1-e_2)^{j_{\max}}} + \underbrace{\sum_{j=1}^{j_{\max}} (Q_{\mathrm{M}-j} - \hat{Q}_{\mathrm{M}-j})}_{\leq \frac{1}{e_2^2}(1-e_2)^{j_{\max}}} \quad [97]$$

$$+ \underbrace{\sum_{j=j_{\max}+1}^{\infty} (Q_{\mathrm{M}+j} - \hat{Q}_{\mathrm{M}+j})}_{\leq \frac{1}{e_2}(1-e_2)^{j_{\max}}} + \underbrace{\sum_{j=j_{\max}+1}^{\infty} (Q_{\mathrm{M}-j} - \hat{Q}_{\mathrm{M}-j})}_{\leq \frac{1}{e_2^2}(1-e_2)^{j_{\max}}} \quad [98]$$

$$\leq \sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{M}j} + \frac{5}{e_2^2}(1-e_2)^{j_{\max}}, \quad [99]$$

$$\sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{M}j}\mu_{j,A} \leq \sum_{j=\pm 1}^{\pm \infty} Q_{\mathrm{M}j}\mu_{j,A} \leq \sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{M}j}\mu_{j,A} + \sum_{j=\pm 1}^{\pm \infty} (Q_{\mathrm{M}j} - \hat{Q}_{\mathrm{M}j}) \quad [100]$$

$$\leq \sum_{j=\pm 1}^{\pm j_{\max}} \hat{Q}_{\mathrm{M}j}\mu_{j,A} + \frac{5}{e_2^2}(1-e_2)^{j_{\max}}. \quad [101]$$

References

1. S Uchida, K Sigmund, The competition of assessment rules for indirect reciprocity. *J. Theor. Biol.* **263**, 13–19 (2010).