Who is a Leader in the Leading Eight? Indirect Reciprocity under Private Assessment

Yuma Fujimoto [1,2,3,*] and Hisashi Ohtsuki [1,4,†]

[1] *Research Center for Integrative Evolutionary Science, SOKENDAI (The Graduate University for Advanced Studies), Shonan Village, Hayama, Kanagawa 240-0193, Japan*

[2] *Universal Biology Institute (UBI), the University of Tokyo. 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*

[3] *CyberAgent, AI Lab, Abema Towers, 40-1, Udagawa-Cho, Shibuya-Ku, Tokyo 150-0042, Japan*

[4] *Department of Evolutionary Studies of Biosystems, SOKENDAI, Shonan Village, Hayama, Kanagawa 240-0193, Japan*

(Received 9 November 2023; accepted 9 May 2024; published 28 May 2024)

Indirect reciprocity is a mechanism that explains large-scale cooperation in human societies. In indirect reciprocity, an individual chooses whether to cooperate with another based on reputation information, and others evaluate the action as good or bad. Under what evaluation rule (called "social norm") cooperation evolves has long been of central interest in the literature. It has been reported that if individuals can share their evaluations (i.e., public reputation), social norms called "leading eight" can be evolutionarily stable. On the other hand, when they cannot share their evaluations (i.e., private assessment), the evolutionary stability of cooperation is still in question. To tackle this question, we create a novel method to analyze the reputation structure in the population under private assessment. Specifically, we characterize each individual by two variables, "goodness" (what proportion of the population considers the individual as good) and "self-reputation" (whether an individual thinks of him or herself as good or bad), and analyze the stochastic process of how these two variables change over time. We discuss the evolutionary stability of each of the leading-eight social norms by studying the robustness against invasions of unconditional cooperators and defectors. We identify key pivots in those social norms for establishing a high level of cooperation or stable cooperation against mutants. Our finding gives an insight into how human cooperation is established in a real-world society.

DOI: 10.1103/PRXLife.2.023009

I. INTRODUCTION

Cooperation has been a major topic in biology, psychology, sociology, and economics [1–4]. Direct reciprocity explains cooperation between two individuals who directly and repeatedly interact with each other [1,5]. However, cooperation is also seen even in a large-scale society, such as in human societies [3,4,6,7]. This large-scale cooperation is difficult to explain because individuals frequently meet strangers and do not always interact with the same person. A key to success in large-scale cooperation is social information, such as reputations and gossip. In a real human society, individuals obtain and use the reputations of others to judge whether they cooperate. This is a core mechanism of indirect reciprocity, where those who have helped others receive help from a third party through reputations [8–10]. In fact, two-thirds of all human conversations are considered to involve reputations and gossip [11–13]. Furthermore, many experimental studies support that reputations and gossip contribute to human cooperation [14–22].

One of the difficulties in maintaining cooperation by indirect reciprocity concerns errors in choosing actions and assigning reputations. In early studies of indirect reciprocity, the social norm called "image scoring" has been discussed [9,23]. An individual with this social norm assigns a good reputation to those who cooperated and a bad reputation to those who did not. A central question is whether cooperation can be maintained by "discriminators," who cooperate with good persons while defect (i.e., not cooperate) with bad persons. In an error-free world, a discriminator cooperates with a good person, and this discriminator obtains a good reputation, which invites cooperation from a third party. However, in a world with errors in actions, a discriminator who accidentally fails to cooperate obtains a bad reputation, which triggers defection from a third party, and this third party obtains a bad reputation, and this triggers another defection, and so on, causing the collapse of cooperation. Therefore, discriminators under the image scoring social norm cannot maintain cooperation in the presence of errors [24–31].

Even under these action and assessment errors, eight social norms have been reported to maintain cooperation in the case of public reputation, where the reputation of each individual is shared among all the individuals [25,26,29]. These social norms are called the "leading eight" and have the following four properties in common [26]: (i) Maintenance of cooperation: a good person who cooperates with a good person should be evaluated as good. (ii) Identification of defectors: defection with a good person should be evaluated as bad. (iii) Justification of punishment: a good person who defects with

*fujimoto_yuma@soken.ac.jp
†ohtsuki_hisashi@soken.ac.jp

a bad person should be evaluated as good. (iv) Forgiveness: a bad person who cooperates with a good person should be evaluated as good. Here, the above image scoring satisfies the conditions of (i), (ii), and (iv) but not (iii). Many studies so far have assumed that the reputation is publicly held [32–46] because it is relatively easy to calculate distribution of reputations in the population.

In reality, however, humans do not always agree with others on the reputation of the same individual, so the assumption of public reputation usually does not hold. Instead, they can have different opinions on the same individual (i.e., private assessment). Thus, the reputation structure in the whole population is described by two-dimensional information of who evaluates whom and how. In a binary reputation case where evaluation is either good or bad, we have a matrix whose elements are good or bad, which is called an image matrix [47–53]. Given that assessment errors occur independently in individuals, this image matrix becomes very complex. This is one of the reasons why the performance of the leading-eight social norms under private assessment has so far been investigated mostly based on numerical simulations. It is known that the leading-eight social norms cannot maintain cooperation [51] in stochastic processes of invasion and fixation [54] against those who always cooperate (ALLC) and those who always defect (ALLD). On the other hand, cooperation is maintained if we additionally introduce somewhat special settings such as empathy, generosity, or spatial structure [55–64]. Although a recent study [53] has theoretically shown that one of the leading-eight norms (called "L3" or "simple standing") can be evolutionarily stable, whereas another one (called "L6" or "stern judging") is not, the remaining six of the leading-eight norms have yet to be analytically studied. This is partly because these six norms assign reputations to individuals in a more complex manner than the other two [53]. Indeed, when these six norms evaluate others, they take into account not only the reputation of the recipient but also that of the donor. This additional complexity hinders analytical treatments of those social norms. In fact, the methodology used in our preceding work [53] does not work. Therefore, we still do not have a whole picture of which of the leading-eight social norms can maintain evolutionarily stable cooperation under private assessment in the presence of errors [4,31].

This study investigates whether each of the leading-eight norms can sustain evolutionarily stable cooperation or not, by discussing its robustness against invasions of ALLC and ALLD mutants under private assessment with errors. To this end, we develop a mathematical framework to analyze stochastic processes of how the image matrix changes and derive the equilibrium state of the image matrix. Based on this theory, we calculate the payoffs of wild-types and mutants and perform invasion analyses.

II. A MODEL

We consider a population of N individuals where a binary reputation of either good (G) or bad (B) is given from each individual to each one at any given moment. Reputations can change over time. Every round, a donor and a recipient are selected at random from the population. The donor chooses an action of either cooperation (C) or defection (D) towards the recipient. If the donor cooperates, he or she pays a cost of c (>0) and gives the recipient a benefit of b ($>c$). Defection generates neither cost nor benefit. Each player has a rule to choose an action, called "action rule." This study assumes that in choosing one's action, the donor considers (1) whether the recipient is good ($X_d = G$) or bad ($X_d = B$) in the eyes of the donor, and (2) whether the donor him or herself is good ($Y_d = G$) or bad ($Y_d = B$) in the eyes of the donor [see Fig. 1(a)]. Thus, the donor's *intended* action is described by a mapping, $Action(X_d, Y_d)$, which takes a letter of either C or D. Here, we assume that, because of an action error, the donor takes the action opposite the intended one with the probability of $0 \leqslant e_1 < 1/2$. We also assume that the donor is aware of the error when he or she commits it. The situation is formalized as follows: Let $a^{X_d Y_d}$ be the probability that the donor's *actual* action, denoted by A, is cooperation (C) when the recipient's reputation in the eyes of the donor is X_d and when the donor's reputation in the eyes of the donor is Y_d. Then we have $a^{X_d Y_d} = 1 - e_1$ if $Action(X_d, Y_d) = C$, while $a^{X_d Y_d} = e_1$ if $Action(X_d, Y_d) = D$. Thus, the action rule of an individual can be characterized by a 4-dimensional vector $\boldsymbol{a} := (a^{GG}, a^{GB}, a^{BG}, a^{BB})$. See Table I for the list of symbols appearing in this paper.

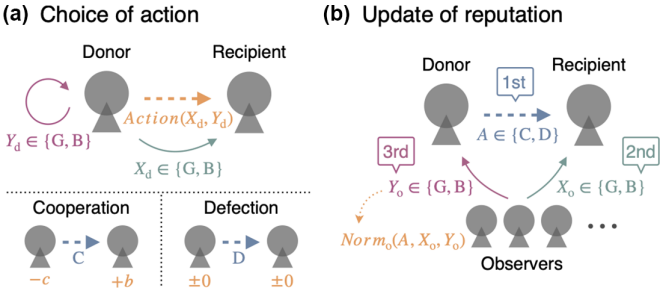

FIG. 1. An illustration of how the model of indirect reciprocity proceeds. (a) The donor chooses the action $Action(X_d, Y_d)$ depending on the recipient's (X_d) and donor's (Y_d) reputations in the eyes of the donor. If the donor cooperates ($A = C$), the donor pays the cost c, while the recipient gains the benefit of b. If the donor defects ($A = D$), neither cost nor benefit arises. (b) Each observer assigns a new reputation $Norm(A, X_o, Y_o)$ to the donor, depending on the donor's action A (called "first-order information"), the recipient's reputation X_o in the eyes of the observer (called "second-order information"), and the donor's reputation Y_o in the eyes of the observer (called "third-order information").

TABLE I. Table of notation introduced in Sec. II.

C, D	Action: cooperation, defection
G, B	Reputation: good, bad
$A \in \{C, D\}$	Variable for action
$X \in \{G, B\}$	Variable for recipient's reputation
$Y \in \{G, B\}$	Variable for donor's reputation
d, o (subscripts of X and Y)	In the eyes of donor, observer
e_1, e_2	Error rates in action and assessment
b, c	Benefit and cost of cooperation
N	Population size
$Norm(A, X, Y)$ (or $\boldsymbol{n}^A$)	Social norm
$Action(X, Y)$ (or $\boldsymbol{a}$)	Action rule

TABLE II. Social norms of L1–L8 and their optimal action rules. In the first column, $Norm(\mathrm{C}, X, Y)$ and $Norm(\mathrm{D}, X, Y)$ indicate the donor's new reputation assigned by each social norm when the donor cooperates and defects with the recipient, respectively. The column with the heading $Action(X, Y)$ indicates actions that each donor chooses. In the second row, the first and second alphabets describe the reputations of the recipient (X) and the donor (Y), respectively.

	$Norm(\mathrm{C}, X, Y)$				$Norm(\mathrm{D}, X, Y)$				$Action(X, Y)$			
XY	GG	GB	BG	BB	GG	GB	BG	BB	GG	GB	BG	BB
L1	G	G	G	G	B	B	G	B	C	C	D	C
L2	G	G	B	G	B	B	G	B	C	C	D	C
L3	G	G	G	G	B	B	G	G	C	C	D	D
L4	G	G	G	B	B	B	G	G	C	C	D	D
L5	G	G	B	G	B	B	G	G	C	C	D	D
L6	G	G	B	B	B	B	G	G	C	C	D	D
L7	G	G	G	B	B	B	G	B	C	C	D	D
L8	G	G	B	B	B	B	G	B	C	C	D	D

We assume that everyone observes this interaction as an observer, evaluates the chosen action by the donor, and updates the reputation of the donor as good or bad, independently of the other observers [see Fig. 1(b)]. We assume that all individuals adopt the same rule to update the donor's reputation; such a rule is called "social norm." In updating the donor's reputation, observers take the following three pieces of information into account: (1) whether the donor's *actual* action is cooperation ($A = \mathrm{C}$) or defection ($A = \mathrm{D}$), (2) whether the recipient is good ($X_\mathrm{o} = \mathrm{G}$) or bad ($X_\mathrm{o} = \mathrm{B}$) in the eyes of the observer, and (3) whether the donor was good ($Y_\mathrm{o} = \mathrm{G}$) or bad ($Y_\mathrm{o} = \mathrm{B}$) before the interaction in the eyes of the observer. The reputation that an observer *intends* to assign to the donor is thus represented by the function of $Norm(A, X_\mathrm{o}, Y_\mathrm{o})$, which we call *intended* reputation. Here, we also assume that each observer erroneously assigns the opposite reputation to the intended one with probability $0 < e_2 < 1/2$ independently of others. The situation is formalized as follows: We refer to the reputation that is actually assigned to the donor as *actual* reputation and distinguish it from the *intended* one. We define $n^{AX_\mathrm{o}Y_\mathrm{o}}$ as the probability that the *actual* reputation that an observer assigns to the donor is good when the donor took action A, when the recipient's reputation in the eyes of the observer is X_o, and when the donor's previous reputation in the eyes of the observer was Y_o. Then we have $n^{AX_\mathrm{o}Y_\mathrm{o}} = 1 - e_2$ if $Norm(A, X_\mathrm{o}, Y_\mathrm{o}) = \mathrm{G}$, while $n^{AX_\mathrm{o}Y_\mathrm{o}} = e_2$ if $Norm(A, X_\mathrm{o}, Y_\mathrm{o}) = \mathrm{B}$. Each social norm is characterized by two 4-dimensional vectors $\boldsymbol{n}^\mathrm{C} := (n^\mathrm{CGG}, n^\mathrm{CGB}, n^\mathrm{CBG}, n^\mathrm{CBB})$ and $\boldsymbol{n}^\mathrm{D} := (n^\mathrm{DGG}, n^\mathrm{DGB}, n^\mathrm{DBG}, n^\mathrm{DBB})$. When the reputation updates are over, the current donor-recipient pair is resolved, and we repeat the process by sampling a donor and a recipient again. We repeat this process infinitely many times and calculate the expected payoff of each player. In computer simulations for a finite population of size N, we assume that N donor-recipient interactions occur in one unit of time.

The leading-eight social norms are of particular interest in this study. We label these norms L1–L8 [30]. These norms and their corresponding action rules are described in Table II. There are several common features in the leading eight, and Ohtsuki and Iwasa [26] explained these commonalities as follows: (i) $Norm(\mathrm{C}, \mathrm{G}, \mathrm{G}) = \mathrm{G}$ and $Action(\mathrm{G}, \mathrm{G}) = \mathrm{C}$ represent "maintenance of cooperation"; a good person who cooperates with a good person should be evaluated as good, (ii) $Norm(\mathrm{D}, \mathrm{G}, *) = \mathrm{B}$ represent "identification of defectors"; defection with a good person should be evaluated as bad, (iii) $Norm(\mathrm{D}, \mathrm{B}, \mathrm{G}) = \mathrm{G}$ and $Action(\mathrm{B}, \mathrm{G}) = \mathrm{D}$ represent "punishment and justification of punishment"; a good person who defects with a bad person should be evaluated as good, and (iv) $Norm(\mathrm{C}, \mathrm{G}, \mathrm{B}) = \mathrm{G}$ and $Action(\mathrm{G}, \mathrm{B}) = \mathrm{C}$ represent "apology and forgiveness"; a bad person who cooperates with a good person should be evaluated as good. On the other hand, the other three pivots in social norms, which are $Norm(\mathrm{C}, \mathrm{B}, \mathrm{G})$, $Norm(\mathrm{C}, \mathrm{B}, \mathrm{B})$, and $Norm(\mathrm{D}, \mathrm{B}, \mathrm{B})$, were left unspecified; they can be either good or bad. This leads to $2^3 = 8$ combinations, and this is the reason why they are called leading eight [26].

TABLE III. Table of notation introduced in Sec. III

$p \in [0, 1]$	Goodness
$s \in \{0, 1\}$	Self-reputation
p', s'	Recipient's goodness and self-reputation
p'', s''	Donor's goodness and self-reputation before an update
$h^{AY}(p', s'')$	Probability that donor chooses A and assigns Y to itself
$f^A(p', p'')$	Average donor's goodness after an update when A is chosen
V	Variance of donor's goodness
$\phi(p, s)$	Frequency distribution of p and s

III. ANALYSIS OF REPUTATION STRUCTURE

A. Overview of mathematical framework

Before going into any details, let us first overview what we do in this section (see Table III for notation). This section considers the reputation structure formed in the wild-type population of one of the leading eight. First, we characterize each individual by two variables: its goodness $p \in [0, 1]$ and its self-reputation $s \in \{0, 1\}$. Here, p indicates the proportion of others who evaluate the focal individual as good, while s indicates whether the individual evaluates him or herself as good ($s = 1$) or bad ($s = 0$). Let $\phi(p, s)$ denote the frequency distribution of (p, s) for the whole population. Our goal is to calculate its equilibrium distribution, denoted by ϕ^*. For that purpose, we derive a recursive relation that ϕ^* should satisfy [Eqs. (5)], and this recursive equation is numerically solved. By using this equilibrium distribution, we derive various quantities, such as $\bar{h}^\mathrm{C}$ [Eq. (6)], which is the probability that wild-types cooperate. With these results, we suggest that the eight social norms in the leading eight can be classified into three different types.

More technically, in deriving the recursive equation for ϕ^* [Eqs. (5)], we consider microscopically a single interaction between a donor and a recipient. We assume that this recipient is characterized by (p', s'), that this donor is characterized by (p'', s''), and that the donor's reputation status is updated to (p, s). The probability with which (i) the donor's choice

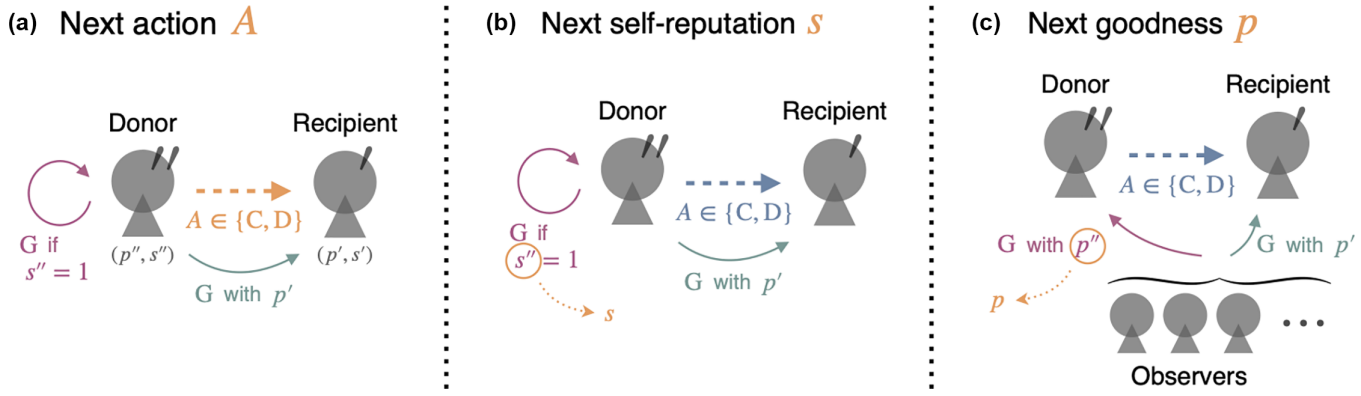

FIG. 2. An illustration of the stochastic transition of the donor's reputation state from (p'', s'') to (p, s). The recipient's icon has a single prime ($'$) while the donor's icon has double primes ($''$), which corresponds to our notations that (p', s') represents the recipient's reputation status while (p'', s'') represents donor's reputation status before an update (see Table III). (a) First, a donor, whose reputation status is (p'', s''), either cooperates (C) or defects (D) with a recipient, whose reputation status is (p', s'). The action $A \in \{C, D\}$ depends on the recipient's reputation in the eyes of the donor (which is G with probability p' because the donor is a random sample from the population) and the donor's self-reputation (given by s''). (b) The donor updates its self-reputation s'' to s, based on the donor's action A, the recipient's reputation in the eyes of the donor [the same as in panel (a)], and the donor's previous self-reputation s''. (c) The observers update the donor's reputation in their eyes. The donor's goodness p'' is updated to p, which depends on the donor's action A, the recipient's reputation in the eyes of each observer (which is G with probability p'), and the donor's previous reputation in the eyes of each observer (which is G with probability p'').

of action toward the recipient is A and (ii) the updated self-reputation of the donor is Y will be represented by h^{AY} in Eq. (1), and they depend on (p', s''). Since assessment errors occur independently of observers, the value of p (donor's new goodness) is a stochastic variable, but with the help of large N, we can approximate its distribution with a Gaussian distribution, where its mean and variance shall be given by Eqs. (3). The next section provides a detailed description of our mathematical framework, followed by the section that shows the results obtained by this framework.

B. Detailed mathematical framework

Let us consider a stochastic process describing how the reputations of a chosen donor in the eyes of itself and the others are updated. For a moment, we assume that everyone in the population adopts the same social norm and action rule and that they are one of the leading eight (see Table II). This assumption will be relaxed later.

We characterize individual's reputations by two variables (p, s), where $p \in [0, 1]$ and $s \in \{0, 1\}$. We define $p \in [0, 1]$ as the proportion of the individuals in the population except for the focal one who assigns a good reputation to the focal individual, and call it the "goodness" of the focal individual. The second variable $s \in \{0, 1\}$ is called "self-reputation" of the focal individual; $s = 1$ means that the individual considers him or herself as good, while $s = 0$ if bad. Hereafter, we call the pair (p, s) "reputation state" of an individual. For the sake of our later analysis, we introduce vector notations of these variables as $\boldsymbol{p} := (p, 1 - p)$ and $\boldsymbol{s} := (s, 1 - s)$.

Let $\phi(p, s)$ be a joint probability distribution of individuals whose reputation status is (p, s) [that is, the chance that a randomly sampled individual from the population has reputation status (p, s)]. To derive the equation that $\phi(p, s)$ satisfies, we suppose that the chosen recipient's reputation state is (p', s') and that the chosen donor's reputation state is (p'', s''). Given these, we want to calculate the transition probability that after one round of interaction, the donor's reputation state is updated from (p'', s'') to (p, s). See Fig. 2 for a schematic illustration.

For that purpose, we calculate h^{AY}, which is the probability that the donor actually takes action $A \in \{C, D\}$ and actually assigns reputation $Y \in \{G, B\}$ to him or herself. It is obtained as

$$h^{CG}(p', s'') = (\boldsymbol{p}' \otimes \boldsymbol{s}'') \cdot (\boldsymbol{a} \circ \boldsymbol{n}^{C}), \tag{1a}$$

$$h^{CB}(p', s'') = (\boldsymbol{p}' \otimes \boldsymbol{s}'') \cdot \{\boldsymbol{a} \circ (\boldsymbol{1} - \boldsymbol{n}^{C})\}, \tag{1b}$$

$$h^{DG}(p', s'') = (\boldsymbol{p}' \otimes \boldsymbol{s}'') \cdot \{(\boldsymbol{1} - \boldsymbol{a}) \circ \boldsymbol{n}^{D}\}, \tag{1c}$$

$$h^{DB}(p', s'') = (\boldsymbol{p}' \otimes \boldsymbol{s}'') \cdot \{(\boldsymbol{1} - \boldsymbol{a}) \circ (\boldsymbol{1} - \boldsymbol{n}^{D})\}. \tag{1d}$$

Here, we used tensor product $\boldsymbol{p}' \otimes \boldsymbol{s}'' := (p's'', p'(1 - s''), (1 - p')s'', (1 - p')(1 - s''))$, which generates a vector of probabilities with which the recipient's reputation in the eyes of the donor and the donor's reputation in the eyes of the donor are good-good, good-bad, bad-good, and bad-bad, respectively. The symbol $\circ$ represents the Hadamard product of two 4-dimensional vectors, which returns a 4-dimensional vector whose component is a component-wise product of the two original vectors, defined as $(x_1, x_2, x_3, x_4) \circ (y_1, y_2, y_3, y_4) := (x_1y_1, x_2y_2, x_3y_3, x_4y_4)$. For example, in Eq. (1a), the vector $\boldsymbol{a}$ represents the probabilities with which the donor cooperates with the recipient in each of the four situations above (i.e., good-good, good-bad, bad-good, and bad-bad), and the vector $\boldsymbol{n}^{C}$ represents the probabilities with which the donor assigns a good reputation to him or herself in each situation given that the donor cooperates with the recipient, so their Hadamard product $\boldsymbol{a} \circ \boldsymbol{n}^{C}$ represents the probabilities with which the donor in each of the four situations above cooperates with the recipient and assigns a good reputation to him or herself. The symbol $\boldsymbol{1}$ represents the 4-dimensional vector with ones in all components. Finally, the symbol $\cdot$ represents the inner product of two vectors.

Next, we calculate the probability that the donor's goodness changes to p after the interaction, given that the donor has chosen action $A \in \{\mathrm{C}, \mathrm{D}\}$. The vector $\boldsymbol{p}' \otimes \boldsymbol{p}'' = (p'p'', p'(1-p''), (1-p')p'', (1-p')(1-p''))$ represents the proportions of observers who thought that the recipients' reputation and the donor's reputation are good-good, good-bad, bad-good, and bad-bad before the interaction, respectively. Given that the donor has chosen action $A \in \{\mathrm{C}, \mathrm{D}\}$, each observer independently updates the donor's reputation by using the social norm, which is represented by vector $\boldsymbol{n}^A$, while an assignment error occurs independently. Thus, the number of others who update the donor's reputation to good, denoted by N_{G}^A below, is given by the sum of four stochastic variables. Each of the variables follows a binomial distribution as

$$N_{\mathrm{G}}^A := N_{\mathrm{G}}^{AGG} + N_{\mathrm{G}}^{AGB} + N_{\mathrm{G}}^{ABG} + N_{\mathrm{G}}^{ABB}, \tag{2a}$$

$$N_{\mathrm{G}}^{AGG} \sim \mathcal{B}[(N-1)p'p'', n^{AGG}], \tag{2b}$$

$$N_{\mathrm{G}}^{AGB} \sim \mathcal{B}[(N-1)p'(1-p''), n^{AGB}], \tag{2c}$$

$$N_{\mathrm{G}}^{ABG} \sim \mathcal{B}[(N-1)(1-p')p'', n^{ABG}], \tag{2d}$$

$$N_{\mathrm{G}}^{ABB} \sim \mathcal{B}[(N-1)(1-p')(1-p''), n^{ABB}], \tag{2e}$$

where $\mathcal{B}(M_{\mathrm{trial}}, P_{\mathrm{success}})$ represents the binomial distribution of parameters M_{trial} (number of trials) and P_{success} (probability of success). For a sufficiently large population ($N \gg 1$), the donor's next goodness $p = N_{\mathrm{G}}^A/(N-1)$ approximately follows the Gaussian distribution by the central limit theorem, where its mean and variance are calculated from Eqs. (2) as

$$\mathrm{E}[p] = \frac{\mathrm{E}\left[N_{\mathrm{G}}^A\right]}{N-1} = (\boldsymbol{p}' \otimes \boldsymbol{p}'') \cdot \boldsymbol{n}^A \ (=: f^A(p', p'')), \tag{3a}$$

$$\mathrm{Var}[p] = \frac{\mathrm{Var}\left[N_{\mathrm{G}}^A\right]}{(N-1)^2} = (\boldsymbol{p}' \otimes \boldsymbol{p}'') \cdot \frac{\boldsymbol{n}^A \circ (\boldsymbol{1} - \boldsymbol{n}^A)}{N-1} = \frac{e_2(1-e_2)}{N-1} \ (=: V). \tag{3b}$$

In deriving Eq. (3b), we have used the fact that each component of vector $\boldsymbol{n}^A$ is either $n^{AX_{\mathrm{o}}Y_{\mathrm{o}}} = e_2$ or $n^{AX_{\mathrm{o}}Y_{\mathrm{o}}} = 1 - e_2$, so each component-wise product in the form of $n^{AX_{\mathrm{o}}Y_{\mathrm{o}}}(1 - n^{AX_{\mathrm{o}}Y_{\mathrm{o}}})$ is always equal to $e_2(1-e_2)$.

Below, we use this Gaussian approximation. The density function of the Gaussian distribution with mean μ and variance σ^2 is denoted $g(x; \mu, \sigma^2)$:

$$g(x; \mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{4}$$

From Eqs. (1)–(3), we can formulate the dynamics of $\phi(p, s)$. The equilibrium state of $\phi(p, s)$, denoted by $\phi^*(p, s)$, satisfies

$$\phi^*(p,1) = \int_0^1 \mathrm{d}p' \sum_{s'} \int_0^1 \mathrm{d}p'' \sum_{s''} \sum_{A} h^{AG}(p', s'') g(p; f^A(p', p''), V) \phi^*(p', s') \phi^*(p'', s''), \tag{5a}$$

$$\phi^*(p,0) = \int_0^1 \mathrm{d}p' \sum_{s'} \int_0^1 \mathrm{d}p'' \sum_{s''} \sum_{A} h^{AB}(p', s'') g(p; f^A(p', p''), V) \phi^*(p', s') \phi^*(p'', s''). \tag{5b}$$

On the right-hand side of each equation, the term $\phi^*(p', s')\phi^*(p'', s'')$ represents the probability that the chosen recipient's reputation status is (p', s') and the donor's one is (p'', s'') at the equilibrium. The term $h^{AG}(p', s'')g(p; f^A(p', p''), V)$ in Eq. (5a) represents the probability the donor takes action $A \in \{\mathrm{C}, \mathrm{D}\}$ and assigns a good reputation to him or herself, times the probability density that the donor's goodness p'' is updated to p. Similarly, the term $h^{AB}(p', s'')g(p; f^A(p', p''), V)$ in Eq. (5b) corresponds to the case of donor's assigning a bad reputation to itself. Finally, the summations over s' and s'' and the integrals for p' and p'' represent all possible combinations of (p', s') and (p'', s''). Note that the Gaussian function g takes a positive value in $(-\infty, \infty)$, but we truncate it to $p \in [0, 1]$, so Eqs. (5) has an approximation error. As long as the variance of Gaussian function V is small (which is the case if e_2 is small and/or N is large), however, this error is not so large. As seen in the left-hand sides of Eqs. (5), the updated distribution should be the same as the original distribution ϕ^* because ϕ^* is at the equilibrium.

Equations (5) cannot be solved analytically in general. As for the leading-eight social norms and action rules (see Table II), the study [52] has found that, for L3 (simple standing) and L6 (stern judging), the equilibrium distribution can be solved analytically when it is approximated by a summation of Gaussian distributions, while it cannot be for the others, L1, L2, L4, L5, L7, and L8. This is mainly because social norms in L3 and L6 do not use the previous reputation of the donor in updating the donor's reputation. Such social norms are categorized as "second-order" norms [10,27,43]. In contrast, the other social norms use the reputation of the donor as well, and they are categorized as "third-order" norms [10,27,43], which are more complex than second-order norms.

Consequently, we numerically derive a solution of ϕ^*. To obtain a solution, we replace ϕ^*s on the right-hand sides of Eqs. (5) with ϕ_k and those on the left-hand sides with ϕ_{k+1} and regard it as a recursion. There is a truncation error in Eqs. (5) because the support of Gaussian function g is not [0,1] but $(-\infty, \infty)$, but we performed an error estimation and found that this error is negligibly small as long as $Ne_2 \gg 1$ (see Appendix C). As far as we explore, the choice of initial state ϕ_0 does not affect the results, either (see Appendix D for the detailed method).

For an arbitrary function $\mathcal{X}$, we denote its expected value under the equilibrium distribution ϕ^* as $\bar{\mathcal{X}}$. For example, the probability with which a recipient receives cooperation in a randomly chosen interaction, denoted by $\bar{h}^{\mathrm{C}} := \sum_{Y=\mathrm{G},\mathrm{B}} \bar{h}^{CY}$,

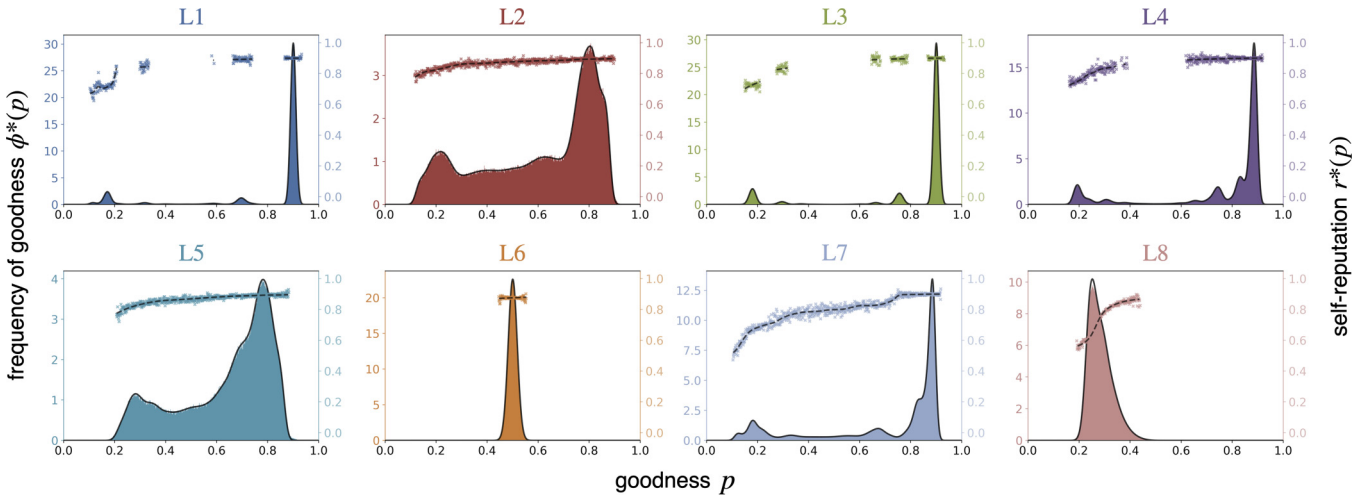

FIG. 3. The equilibrium states of $\phi^*(p, s)$ for the leading-eight norms (labeled L1–L8). In each panel, the horizontal axis indicates p, the goodness of an individual. The colored areas represent $\phi^*(p) = \phi^*(p, 0) + \phi^*(p, 1)$ in individual-based simulations, that is the marginal distribution of p, as indicated by the left axis of the panel. The colored "x" markers indicate $r^*(p) = \phi^*(p, 1)/\phi^*(p)$ in individual-based simulations, which is the conditional distribution of individuals who assign a good reputation to themselves, given their goodness is p, as indicated by the right axis of the panel. We also give the numerical solutions: the black solid lines show numerical solutions of $\phi^*(p)$, while the black broken ones show numerical solutions of $r^*(p)$. We see nearly a perfect match between individual-based simulations and numerical calculations. The individual-based simulations are based on $(e_1, e_2) = (0.03, 0.1)$, $N = 800$, and 2000 samplings from time 51 to 2050. The numerical solutions are derived for $N = 800$; we stop the iteration when the L^2 distance of $\|\phi_K(p) - \phi_{K-1}(p)\|_2 < 10^{-6}$ is achieved for the first time and regard ϕ_K as the equilibrium distribution ϕ^*, where $\|\cdot\|_2$ is the L^2 norm.

is obtained as

$$\begin{aligned}\bar{h}^{\mathrm{C}} &= \int_0^1 \mathrm{d}p' \sum_{s'} \int_0^1 \mathrm{d}p'' \sum_{s''} \underbrace{\sum_{Y} h^{\mathrm{CY}}(p', s'')}_{=(\boldsymbol{p}'\otimes \boldsymbol{s}'')\cdot \boldsymbol{a}} \\ &\quad \times \phi^*(p', s')\phi^*(p'', s'') \\ &= \left[\underbrace{\int_0^1 \mathrm{d}p' \sum_{s'} \boldsymbol{p}' \phi^*(p', s')}_{=:\bar{\boldsymbol{p}}} \otimes \underbrace{\int_0^1 \mathrm{d}p'' \sum_{s''} \boldsymbol{s}'' \phi^*(p'', s'')}_{=:\bar{\boldsymbol{s}}} \right] \cdot \boldsymbol{a} \\ &= (\bar{\boldsymbol{p}} \otimes \bar{\boldsymbol{s}}) \cdot \boldsymbol{a}, \end{aligned} \tag{6}$$

where $\bar{\boldsymbol{p}}$ and $\bar{\boldsymbol{s}}$ represent the population average of $\boldsymbol{p}$ and $\boldsymbol{s}$, respectively.

C. Three different types in the leading eight

Figure 3 shows the converged distribution by this iterative method for each of the leading-eight social norms and its corresponding action rule (see Table II), which well fits the equilibrium distribution computed from an individual-based simulation. There, the marginal distribution of p defined as $\phi^*(p) := \phi^*(p, 0) + \phi^*(p, 1)$ and its conditional distribution for $s = 1$ defined as $r^*(p) := \phi^*(p, 1)/\phi^*(p)$ are simultaneously plotted in the same figure. We remark that the computational complexity of our numerical method is independent of N, while that of the individual-based simulation is of order N^2. Thus, our numerical method is much more efficient than individual-based simulations for large N.

According to Fig. 3, we can classify the leading eight into three types, as follows:

(Type 1) L1, L3, L4, and L7. Those norms are characterized by $Norm(\mathrm{C}, \mathrm{B}, \mathrm{G}) = \mathrm{G}$ (see Table II). Under these norms, there is a sharp peak at a very high goodness (i.e., around $p = 0.9$ in Fig. 3) in the equilibrium distribution, $\phi^*(p)$, suggesting that a large majority of individuals in the equilibrium population has a good reputation in the eyes of most observers. The second highest peak is at a very low goodness, although its height is considerably lower than the first one, suggesting that some minority of individuals have a bad reputation in the eyes of most observers. The third highest peak is slightly left of the first peak, the fourth peak is slightly right of the second one, and so on. This behavior, that is, where peaks exist, has already been observed for L3 (simple standing) in our previous study [53], where we developed its analytical treatment. In short, under these four norms, most of the individuals in the population keep very high goodness in spite of errors.

(Type 2) L2 and L5. Those norms are characterized by $Norm(\mathrm{C}, \mathrm{B}, \mathrm{G}) = \mathrm{B}$ and $Norm(\mathrm{C}, \mathrm{B}, \mathrm{B}) = \mathrm{G}$ (see Table II). Under these norms, the equilibrium distribution $\phi^*(p)$ looks rather continuous, which is in contrast with Type-1 norms that admit discrete peaks in the equilibrium. The highest peak exists at a high goodness value (i.e., around $p = 0.8$ in Fig. 3), but this value is not as high as the corresponding highest peak (i.e., around $p = 0.9$ in Fig. 3) for Type-1 norms. The second highest peak is at a low goodness value. In short, under these two norms, some portion of individuals in the population keep moderately high goodness, while others have either intermediate or low goodness.

(Type 3) L6 and L8. Those norms are characterized by $Norm(\mathrm{C}, \mathrm{B}, \mathrm{G}) = \mathrm{B}$ and $Norm(\mathrm{C}, \mathrm{B}, \mathrm{B}) = \mathrm{B}$ (see Table II). Under these two norms, the equilibrium distribution $\phi^*(p)$ looks unimodal. Its position is at $p = 0.5$ for L6 ("stern-judging") and lower than that for L8. Note that this result for L6 has already been known in previous literature [47,49,51,53]. In short, these two norms fail to sustain

TABLE IV. Table of notation introduced in Sec. IV.

$\boldsymbol{a}_{\mathrm{M}}$	Action rule of mutants
$p_{\mathrm{M}} \in [0, 1]$	Goodness of mutants from wild types
$\phi_{\mathrm{M}}(p_{\mathrm{M}})$	Frequency distribution of p_{M}

individuals with a high goodness in the equilibrium population. Since corresponding action norms for L6 and L8 prescribe cooperation with good individuals and defection with bad individuals (see Table II), it follows that the equilibrium level of cooperation in the population is low, suggesting that indirect reciprocity does not sufficiently work under these norms when assessment is private.

Figure 3 further shows the conditional distribution of self-reputation $r^*(p)$ tends to increase with p in all of the leading eight. This increasing trend means a positive correlation between one's self-reputation and goodness: the one who is evaluated as good by a large proportion of others tends to think of him or herself as good. Self-reputations in L7 and L8 are relatively lower than L1–L6. This is because action rules in L1–L6 always prescribe an action that is evaluated as good under the corresponding social norm, whereas action rules in L7 and L8 do not (see Table II). More specifically, action rules in L7 and L8 choose $\mathit{Action}(\mathrm{B}, \mathrm{B}) = \mathrm{D}$ when both the donor and the recipient are bad, but this behavior is evaluated as $\mathit{Norm}(\mathrm{D}, \mathrm{B}, \mathrm{B}) = \mathrm{B}$ by the corresponding social norms (see Table II).

IV. INVASION ANALYSIS

A. Reputations of mutants

We first overview the flow of this section (see Table IV for notation). This section aims to perform an invasion analysis, where the wild type is one of the leading eight, and the rare mutant is either ALLC (who always intends to cooperate) or ALLD (who always intends to defect). The mutant's goodness is represented by p_{M}, which is the proportion of wild types who evaluate the focal mutant as good. Let $\phi^*_{\mathrm{M}}(p_{\mathrm{M}})$ denote the equilibrium frequency distribution of p_{M} among mutants. We first derive the recursive equation that ϕ^*_{M} satisfies [Eq. (8)], which is used to derive the average goodness of mutants [Eq. (9)], which is in turn used to derive the probability that mutants receive cooperation from wild-types $\bar{h}^{\mathrm{C}}_{\mathrm{M}}$ [Eq. (10)]. With these, we compare the payoff of wild types and mutants and derive the ESS parameter region of the leading eight against ALLC and ALLD (Fig. 4).

Let us start to mathematically formulate the situation where ALLC or ALLD mutants invade wild types who use one of the leading-eight social norms and their corresponding action rules. ALLC individuals are those who always intend to cooperate with the recipient. ALLD individuals are those who always intend to defect. We assume that they are susceptible to action errors. On the other hand, they do not need to possess a social norm because they choose their actions independently of others' reputations. This simple nature of ALLC and ALLD enables us to perform the following invasion analysis.

Because these mutants always intend to choose C or D, their action rules are simple enough. They are represented as $\boldsymbol{a}_{\mathrm{M}} = (a_{\mathrm{M}}, a_{\mathrm{M}}, a_{\mathrm{M}}, a_{\mathrm{M}})$, where $a_{\mathrm{M}} = 1 - e_1$ for ALLC while $a_{\mathrm{M}} = e_1$ for ALLD. Here and hereafter, symbols with subscript M represent those for mutants, and symbols without subscripts are those for wild types, unless otherwise specified. Because there are no rational reasons to expect that ALLC and ALLD are not susceptible to action errors, e_1, while strategies based on the leading eights are, assumptions $a_{\mathrm{M}} = 1 - e_1$ (ALLC) are $a_{\mathrm{M}} = e_1$ (ALLD) are quite natural. Note, however, that ALLC and ALLD are not susceptible to assessment errors, e_2, at all, while strategies based on the leading eights are. Under this setting, we would like to know the payoffs of wild types and mutants.

Since mutants are rare, the average payoff of wild types is unaffected by mutants. The probability that a wild-type individual receives cooperation as a recipient remains the same as $\bar{h}^{\mathrm{C}}$ in Eq. (6), which is the same as the probability that a wild type actually performs cooperation as a donor. Therefore, the average payoff of wild types is

$$u = (b - c)\bar{h}^{\mathrm{C}}. \tag{7}$$

As for the payoff of mutants, the probability that a mutant individual cooperates as a donor is a_{M}. What remains is to derive the probability that a mutant receives cooperation as a recipient, and here we can assume that its donor is a wild type because the chance of mutant-mutant interactions can be negligibly small.

Since ALLC or ALLD do not use its self-reputation, we do not have to consider the self-reputation of these mutants. Thus, we consider only $\phi_{\mathrm{M}}(p_{\mathrm{M}})$, which is defined as the probability density function of goodness p_{M} of mutants in the eyes of wild-type observers, in the following. Its equilibrium distribution, i.e., $\phi^*_{\mathrm{M}}(p_{\mathrm{M}})$, should satisfy

$$\begin{aligned}\phi^*_{\mathrm{M}}(p_{\mathrm{M}}) = \int_0^1 \mathrm{d}p' \sum_{s'} \int_0^1 \mathrm{d}p''_{\mathrm{M}} \{ & a_{\mathrm{M}} g(p_{\mathrm{M}}; f^{\mathrm{C}}(p', p''_{\mathrm{M}}), V) \\ & + (1 - a_{\mathrm{M}}) g(p_{\mathrm{M}}; f^{\mathrm{D}}(p', p''_{\mathrm{M}}), V)\} \phi^*(p', s') \\ & \times \phi^*_{\mathrm{M}}(p''_{\mathrm{M}}),\end{aligned} \tag{8}$$

where the equation considers how the goodness p''_{M} of a mutant donor is updated to p_{M} after an interaction with a wild-type recipient whose reputation status is (p', s'). In particular, the average goodness of the mutants in the eyes of wild types, denoted by $\bar{\boldsymbol{p}}_{\mathrm{M}}$, is calculated as

$$\begin{aligned}\bar{p}_{\mathrm{M}} &= \int_0^1 \mathrm{d}p_{\mathrm{M}} p_{\mathrm{M}} \phi^*_{\mathrm{M}}(p_{\mathrm{M}}) \\ &= \int_0^1 \mathrm{d}p' \sum_{s'} \int_0^1 \mathrm{d}p''_{\mathrm{M}} \int_0^1 \mathrm{d}p_{\mathrm{M}} p_{\mathrm{M}} \{ a_{\mathrm{M}} g(p_{\mathrm{M}}; f^{\mathrm{C}}(p', p''_{\mathrm{M}}), V) + (1 - a_{\mathrm{M}}) g(p_{\mathrm{M}}; f^{\mathrm{D}}(p', p''_{\mathrm{M}}), V)\} \phi^*(p', s') \phi^*_{\mathrm{M}}(p''_{\mathrm{M}})\end{aligned}$$

$$\simeq \int_0^1 \mathrm{d}p' \sum_{s'} \int_0^1 \mathrm{d}p''_{\mathrm{M}} \{a_{\mathrm{M}} f^{\mathrm{C}}(p', p''_{\mathrm{M}}) + (1 - a_{\mathrm{M}}) f^{\mathrm{D}}(p', p''_{\mathrm{M}})\} \phi^*(p', s') \phi^*_{\mathrm{M}}(p''_{\mathrm{M}})$$

$$\simeq a_{\mathrm{M}} f^{\mathrm{C}}(\bar{p}, \bar{p}_{\mathrm{M}}) + (1 - a_{\mathrm{M}}) f^{\mathrm{D}}(\bar{p}, \bar{p}_{\mathrm{M}}). \tag{9}$$

Here, from the first to the second line we have used Eq. (8). From the second to the third line we have computed the integral with respect to p_{M}. Note that in order to take the expectation of Gaussian distribution, we have approximated the range of integral $\int_0^1 \mathrm{d}p_{\mathrm{M}}$ as $\int_{-\infty}^{\infty} \mathrm{d}p_{\mathrm{M}}$. From the third to the fourth line we have computed the integrals with respect to p' and p''_{M}, where we have taken advantage of the fact that both f^{C} and f^{D} are multilinear functions of p' and p''_{M}. Note that we have again approximated the ranges of the integrals in the same way as above. Since the final expression in Eq. (9) is linear in $\bar{p}_{\mathrm{M}}$, we can solve Eq. (9) with respect to $\bar{p}_{\mathrm{M}}$ (see Appendix C for the estimation of errors). This approximate solution is a function of $\bar{p}$.

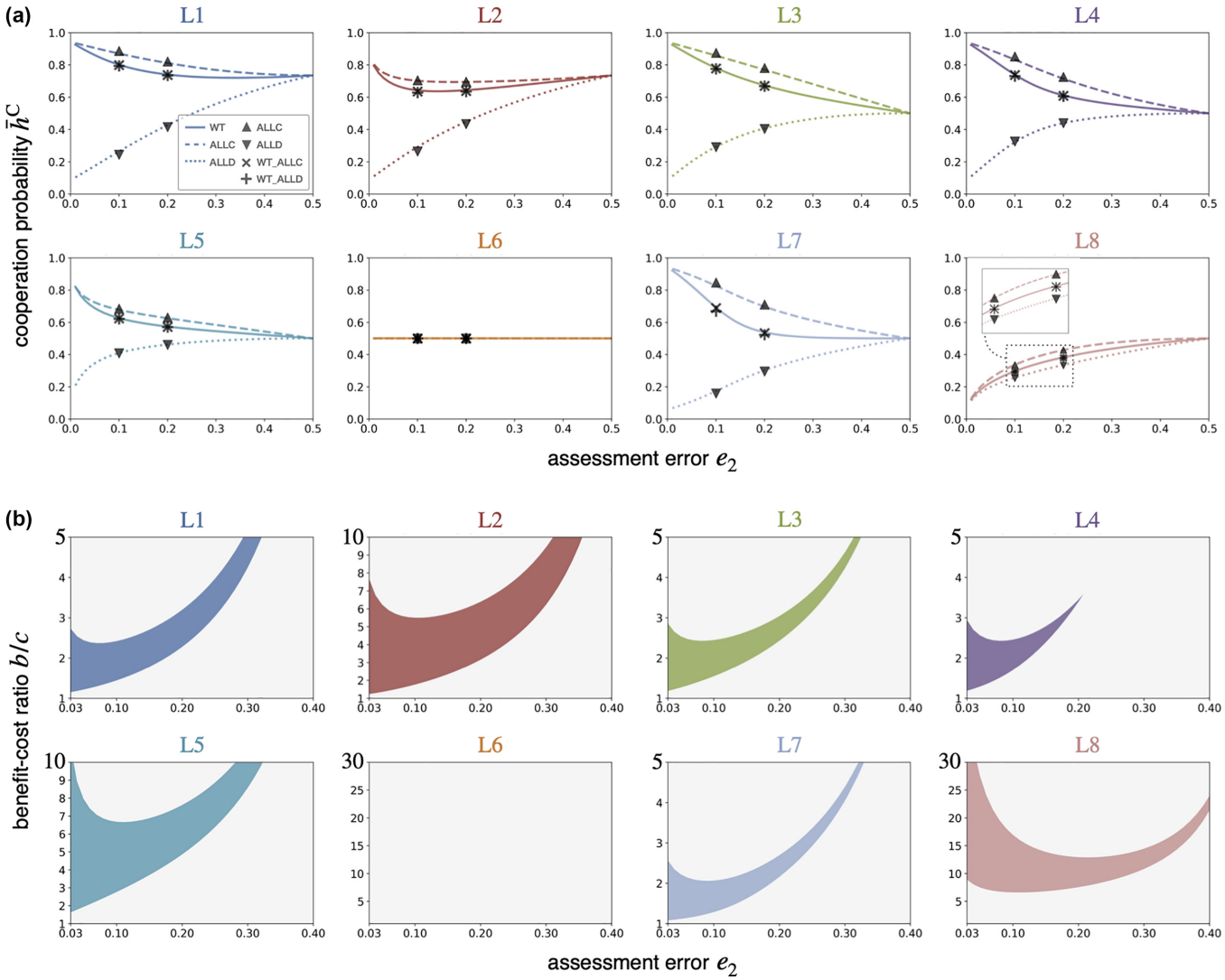

FIG. 4. (a) Cooperation probabilities of the leading eight. Colored lines in each panel show the probabilities that wild types of each of the leading eight cooperate with wild types themselves (solid), wild types cooperate with ALLC mutants (dashed), and wild types cooperate with ALLD mutants (dotted), respectively. These colored lines are obtained numerically by using Eqs. (6) and (10) for $N = 800$ and $e_1 = 0.03$. The gray markers at $e_2 = 0.1$ and $e_2 = 0.2$ show the probabilities that wild types cooperate with ALLC (triangles) and themselves (crosses) when mutants are ALLC, and the probabilities that they cooperate with ALLD (inverted triangles) and themselves (pluses) when mutants are ALLD. These markers are calculated from individual-based simulations with $N = 10\,000$ and $e_1 = 0.03$ where the proportion of mutants is set to 0.03, and averages of 1000 samples from time steps $51 \leqslant t \leqslant 1050$ are shown. (b) ESS regions of the leading eight against ALLC and ALLD mutants are shown by colored areas. These regions are obtained by the cooperation probabilities plotted in panel (a). Note that the scales of y axis are different between panels.

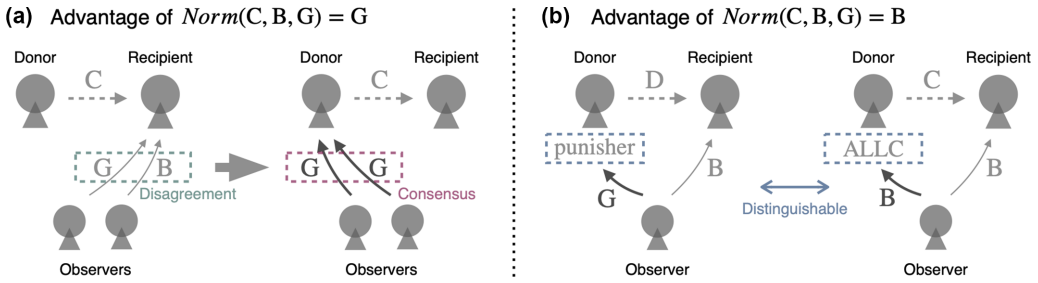

FIG. 5. (a) The advantage of *Norm*(C, B, G) = G lies in that assessment errors cannot spread. In the left panel, both the donor and recipient basically have good reputations. However, only one of two observers (i.e., the right observer) erroneously evaluates the recipient as bad. In the right panel, we show how the observers update the donor's reputation (dark-colored arrows). Despite the disagreement in their opinions on the recipient's reputation, the observers reach a consensus on the donor's reputation because *Norm*(C, G, G) = G for the left observer while *Norm*(C, B, G) = G for the right. (b) The advantage of *Norm*(C, B, G) = B lies in that unconditional cooperation is detectable. We consider a situation where the recipient is considered as bad in the eyes of all individuals. In the left panel, a good donor who adopts the corresponding action rule of the leading eight chooses defection with the recipient [*Action*(B, G) = D]. The observer updates the donor's reputation to a good one (dark-colored arrow) because *Norm*(D, B, G) = G. On the other hand, in the right panel, the donor is ALLC and unconditionally cooperates with others. The observer updates the ALLC donor's reputation to a bad one (dark-colored arrow) because *Norm*(C, B, G) = B. Thus, the leading-eight norms satisfying *Norm*(C, B, G) = B can distinguish ALLC from punishers.

With this $\bar{p}_{\rm M}$, the probability that a mutant recipient receives cooperation from a wild-type donor, denoted by $\bar{h}^{\rm C}_{\rm M}$, is calculated as

$$\begin{aligned}\bar{h}^{\rm C}_{\rm M} &= \int_0^1 {\rm d}p'_{\rm M} \int_0^1 {\rm d}p'' \sum_{s''} \underbrace{\sum_{Y} h^{CY}(p'_{\rm M}, s'')}_{=(\boldsymbol{p}'_{\rm M}\otimes \boldsymbol{s}'')\cdot \boldsymbol{a}} \\ &\quad \times \phi^*_{\rm M}(p'_{\rm M})\phi^*(p'', s'') \\ &= (\bar{\boldsymbol{p}}_{\rm M} \otimes \bar{\boldsymbol{s}}) \cdot \boldsymbol{a}. \end{aligned} \tag{10}$$

With this $\bar{h}^{\rm C}_{\rm M}$, the average payoff of mutants is calculated as

$$u_{\rm M} = b\bar{h}^{\rm C}_{\rm M} - ca_{\rm M}. \tag{11}$$

By comparing Eqs. (7) and (11), we can investigate the invasibility of ALLC and ALLD mutants.

B. Evolutionary stability against ALLC and ALLD

Using the above equations, we discuss whether mutants of ALLC or ALLD can invade wild types of each of the leading eight. Figure 4(a) shows the probability that wild types cooperate with wild types (i.e., $\bar{h}^{\rm C}$) and two probabilities that wild types cooperate with ALLC or ALLD mutants (i.e., $\bar{h}^{\rm C}_{\rm M}$ for $a_{\rm M} = 1 - e_2$ and e_2). Figure 4(b) further shows the ESS regions of the leading eight against ALLC and ALLD. The upper bound of the ESS region in each panel corresponds to the invasion condition of ALLC; that is, ALLC can invade the wild-type population if b/c ratio is above that boundary. Similarly, the lower bound of each ESS corresponds to the invasion condition of ALLD; ALLD can invade the wild-type population if b/c ratio is below that boundary.

(Type 1) L1, L3, L4, and L7. These norms can maintain a maximum cooperation level when it is dominant in the population for a small assessment error rate. Indeed, Fig. 4(a) shows $\bar{h}^{\rm C} \to 1 - e_1$ (=0.97 in this figure) in the limit of $e_2 \to 0$. Remember that these norms have *Norm*(C, B, G) = G in common (see Table II). We find that this feature contributes to a high level of cooperation. To see this, imagine two observers of an interaction between a donor and a recipient, and suppose that one of them evaluates the recipient as good, while the other evaluates the recipient as bad possibly due to an assessment error [see Fig. 5(a), left]. Suppose also that the donor with a good reputation cooperates with the recipient. This cooperation is viewed as "cooperation with a good recipient by a good donor" by the former observer, and this observer assigns the reputation *Norm*(C, G, G) = G to the donor. In contrast, this cooperation is viewed as "cooperation with a bad recipient by a good donor" by the latter observer, and this observer assigns the reputation *Norm*(C, B, G), which is G, to the donor. Therefore, even though observers disagree in their opinions on the recipient, they can reach a consensus on their opinions on the donor [see Fig. 5(a), right]. This is an intuitive reason why Type-1 norms achieve a very high level of cooperation. The result of study [53] is also explained by the same logic.

In exchange for such maximum cooperation, however, Type-1 norms have relatively narrower ESS regions compared with Type 2. Indeed, Fig. 4(b) shows that these norms are relatively more vulnerable to the invasion of ALLC for a large b/c than Type-2 norms. This is because Type-1 norms have *Norm*(C, B, G) = G in common, and hence they positively evaluate the unconditional cooperation by ALLC mutants too much. When $e_1 = 0$ and $e_2 \ll 1$, the ESS regions of Type-1 norms are analytically shown to be $1 < b/c < 2$ (see Appendix A).

(Type 2) L2 and L5. The cooperation level of these norms is high, but not as high as Type 1. When these norms are dominant in the population and when the assessment error rate is small, the cooperation level is $\bar{h}^{\rm C} \simeq 0.8$ in Fig. 4(a). Unlike Type-1 norms, these norms have the feature of *Norm*(C, B, G) = B (see Table II). Imagine again two observers described in Fig. 5(a), who disagree in the opinions of the recipient. Unlike Type-1 norms, when a good donor cooperates with the recipient, the observer who thinks of the recipient as good assigns the reputation *Norm*(C, G, G) = G to the donor, but the observer who thinks of the recipient as

bad assigns the reputation $Norm(\mathrm{C}, \mathrm{B}, \mathrm{G}) = \mathrm{B}$ to the donor. Thus, the disagreement on the recipient's reputation between the two observers will trigger a new disagreement; that is, the disagreement on the donor's reputation. Bad reputations spread in this way until another common feature of Type-2 norms, $Norm(\mathrm{C}, \mathrm{B}, \mathrm{B}) = \mathrm{G}$, eventually prevents further spread of bad reputations. This is because when a bad donor cooperates with the recipient, the feature $Norm(\mathrm{C}, \mathrm{G}, \mathrm{B}) = Norm(\mathrm{C}, \mathrm{B}, \mathrm{B}) = \mathrm{G}$ guarantees that such cooperation is always regarded as good irrespective of how the recipient is viewed from observers. Thus, Type-2 norms can keep their cooperation level high.

At the expense of their cooperation level, Type-2 norms have wider ESS regions than Type-1 norms [see Fig. 4(b)]. This is mainly because Type-2 norms are more resistant to the invasion of ALLC mutants than Type 1. Unlike Type-1 norms, Type-2 observers regard cooperation toward a bad individual as bad [$Norm(\mathrm{C}, \mathrm{B}, \mathrm{G}) = \mathrm{B}$], hence they occasionally assign bad reputations to unconditional cooperators (i.e., ALLC). Thus, Type-2 norms can selectively assign bad reputations to ALLC mutants but not to residents, leading to wider ESS regions.

(Type 3) L6 and L8. Their cooperation level is low ($\bar{h}^{\mathrm{C}} \leqslant 1/2$). This is because Type-3 norms suffer from disagreements on one's reputation among individuals. Recall that Type-3 norms are characterized by $Norm(\mathrm{C}, \mathrm{B}, \mathrm{G}) = Norm(\mathrm{C}, \mathrm{B}, \mathrm{B}) = \mathrm{B}$. Since $Norm(\mathrm{C}, \mathrm{G}, \mathrm{G}) \neq Norm(\mathrm{C}, \mathrm{B}, \mathrm{G})$, by the same logic as Type-2 norms, disagreement between observers about a recipient propagates to that about a donor. The feature $Norm(\mathrm{C}, \mathrm{G}, \mathrm{B}) \neq Norm(\mathrm{C}, \mathrm{B}, \mathrm{B})$ further worsens such disagreements. Due to the accumulation of assessment errors, individuals fail to synchronize their evaluations of the same individual at equilibrium, as suggested by the unimodal peak at an intermediate p value in Fig. 3, leading to a low cooperation level.

Furthermore, Type-3 norms are fragile against the invasion of ALLD mutants. Recall that individuals under Type-3 norms have only intermediate levels of goodness (see Fig. 3). Since $Norm(\mathrm{D}, \mathrm{B}, \mathrm{G}) = \mathrm{G}$, these norms consider ALLD mutants as good to some extent. Thus, their cooperation levels toward themselves and ALLD mutants can differ little. Indeed, Fig. 4(a) shows that L6 cannot distinguish themselves and ALLD, while L8 barely does. Figure 4(b) also shows that L6 is always invaded by ALLD, while L8 is invaded by ALLD unless the benefit-cost ratio is unrealistically high ($b/c > 5 \approx 10$).

V. CONCLUSION AND DISCUSSION

The leading-eight norms have been known to maintain cooperation under public reputation in indirect reciprocity [25,26]. Whether these norms can maintain cooperation even under private assessment (where all individuals independently evaluate others), however, has been largely an open question (but see Ref. [51]). In the present paper, we have extended our previous mathematical framework for studying second-order social norms [52,53] and developed a methodology that enables us to study third-order social norms. We have revealed that the leading-eight norms can be classified into three types based on the shape of their reputation structure and the degree of cooperation. We have also discussed the invasibility of ALLC and ALLD mutants and found that these three types of social norms are different in their resistance to invasion by those mutants. In addition, we have provided intuitive explanations of where this difference originates from.

Specifically, this study has shown that Type-1 norms have a maximal cooperation level but are relatively weak to the invasion of ALLC mutants, leading to relatively narrow ESS regions. On the other hand, Type-2 norms have a lower cooperation level than Type-1 norms, but their ESS regions are relatively wider due to their resistance to ALLC mutants. Thus, there is a trade-off between their cooperation levels and resistance to ALLC mutants. We identified that $Norm(\mathrm{C}, \mathrm{B}, \mathrm{G})$ is a key pivot and that it is the source of this trade-off. The pivot $Norm(\mathrm{C}, \mathrm{B}, \mathrm{G})$ represents how to evaluate cooperation by a good donor toward a bad recipient from the viewpoint of observers. If this evaluation is good (i.e., Type-1 norms), then an advantage arises that the spread of disagreements among individuals is avoided, with a disadvantage that individuals show generosity toward unconditional cooperation by ALLC mutants. On the other hand, if this evaluation is bad (i.e., Type-2 and -3 norms), the opposite effects appear. Whether the further spread of disagreements occurs or not is determined by another key pivot, $Norm(\mathrm{C}, \mathrm{B}, \mathrm{B})$. If this evaluation is good (i.e., Type-2 norms), cooperation is sustained, whereas if this evaluation is bad (i.e., Type-3 norms), sustaining cooperation is difficult.

Fujimoto and Ohtsuki [53] specified parameter regions where each of the second-order norms is not invaded by the other norms. The leading-eight norms contain two second-order norms, L3 and L6 (called S_{03}, S_{07} in their study). Our finding here that L3 can be evolutionarily stable but L6 cannot is consistent with their result (although detailed settings are different between these studies: for example, our study assumes ALLC and ALLD as potential mutants, while their study considered ALLG and ALLB mutants, which always assign good or bad reputations to everyone, and they studied several other mutants as well). In this sense, the current study is an extension of Fujimoto and Ohtsuki [53]; we have studied the other six leading-eight social norms as well, classified them, and investigated their property.

Hilbe *et al.* [51] performed a classification of the leading-eight norms according to their ability to recover from a single disagreement in private reputation evaluation. One of their classifications, based on the expected time until recovery (see their Proposition 3 in its Supporting Information) matches our classification here. That is, our Type-1 norms (L1, L3, L4, L7) have the quickest recovery time from a single disagreement, our Type-2 norms (L2, L5) are runners-up, and our Type-3 norms (L6, L8) are the worst. Their result is consistent with ours because a longer recovery time from a single disagreement implies that goodness p tends to decline more in an error-prone world, where errors constantly supply the sources of disagreements (see Fig. 3). In contrast, previous studies [51,65] report that the strategy based on L3, L4, L5, or L6 rarely dominates the population during the competition with ALLC and ALLD strategies, and the reason for that is explained by their common property $Norm(\mathrm{D}, \mathrm{B}, \mathrm{B}) = \mathrm{G}$, which contributes to giving a good reputation to ALLD

players. Our analysis here does not predict such patterns for L3-L6 norms simply because we studied evolutionary stability, and therefore, ALLD invaders are assumed to be rare. In particular, norm L3 is classified as a promising candidate in our analysis, while it is not in Refs. [51,65], but these seemingly inconsistent results are attributable to the different nature of the two analyses, and they have no contradiction at all. Rather, we firmly believe that our analysis here and that of Refs. [51,65] are complementary to each other, both revealing fundamental properties of the leading-eight social norms.

An example of expected future studies is to extend the analysis to more general situations. First, the current study considered only ALLC and ALLD as potential invaders. It would be interesting to study other types of mutants and see the robustness of Type-1 and -2 norms in the leading eight. This includes studying mutual invasibility among the leading-eight norms. Second, the leading-eight norms are just eight norms out of $2^8 = 256$ possible third-order norms, so exhaustively studying the property of each of the 256 third-order norms is important in order to clarify the role of third-order norms under private assessment. Third, studying higher-order norms, which use more information in assigning a new reputation such as the previous reputation of the recipient (called fourth-order information [43,45]), is an interesting direction of extension. In particular, how the norm complexity [43,45] (concerning the order of norms) is related to its ability to maintain cooperation is an open question under private assessment. Fourth, we have assumed for simplicity that everyone in the population observes every interaction in the population. It would be more realistic to consider the situation of partial observation, in which only a fraction (say, fraction $0 < \theta < 1$) of individuals can observe a given interaction and update reputations. Fifth, it would be interesting to investigate the reputation structure of the population under nonbinary reputation, such as ternary reputation [66,67], where each individual can be labeled as good, bad, or "neutral," and qualitative assessments [65], where reputations are integer scores. Sixth, our current analysis is restricted to ESS analysis, but studying full evolutionary dynamics when two or more strategies coexist with some sizable frequencies would be interesting to see the possibility of stable coexistence or evolutionary cycles.

Regarding the reality of the model, individuals that appear in our model are highly idealized, such as having a high cognitive capacity to remember reputations and achieve evaluations of others based on a third-order norm, and having opportunities to witness all social interactions in the population. Real humans are certainly less able to perform those tasks, and their behavior cannot be perfect, so the level of cooperation that would be established by those agents can be lower than the one predicted by our results due to such imperfectness. However, we believe that the value of our model lies in that it gives a good theoretical reference point as to how much cooperation can be established under the leading-eight social norms if the most favorable conditions are met. As we briefly mentioned above, relaxing those conditions to see how they affect the cooperation level will give valuable insights into real human cooperation.

In conclusion, we have successfully classified the leading-eight norms into three types according to their performance under noisy and private assessment. From the perspective of the cooperation level and resistance to ALLC and ALLD mutants, we find that Type-1 and -2 norms are promising. In particular, the pivot *Norm*(C, B, G) in the social norm determines the trade-off between cooperation level and resistance to ALLC. It is eagerly awaited to experimentally examine whether *Norm*(C, B, G) is evaluated as good or bad in real societies and how this pivot actually contributes to sustaining a high level of cooperation by indirect reciprocity.

The code that we used is available online [68].

ACKNOWLEDGMENTS

Y.F. acknowledges the support by JSPS KAKENHI Grant No. JP21J01393. H.O. acknowledges the support by JSPS KAKENHI Grants No. JP19H04431 and No. JP23K03211.

Y.F. and H.O. designed research, calculated results numerically and analytically, and wrote the paper.

APPENDIX A: RESULTS FOR $e_1 = 0$

1. Simulation results

In the main text, we have assumed $e_1 = 0.03$. To provide theoretical insights into the differences among the leading-eight norms, this section focuses on the case of no action error, $e_1 = 0$. First, Fig. 6 shows the reputation structure of the leading-eight norms in the same way as Fig. 3 but for $e_1 = 0$. This figure shows that the leading-eight norms can be again categorized into three types in their reputation structure, namely, L1, L3, L4, and L7 are in Type-1, L2 and L5 are in Type-2, and L6 and L8 are in Type-3. The only qualitative difference is that the conditional distribution for p given that $s = 1$ is constant, $r^*(p) = 1 - e_2$, for L1–L6. This is because a donor with L1–L6 is always able to take actions that are evaluated as good in its own eyes [because $e_1 = 0$ and $Norm(Action(X, Y), X, Y) = \mathrm{G}$ for any X and Y; see Table II] and actually evaluates itself as good unless an assessment error occurs. On the other hand, a donor with L7 and L8 inevitably chooses actions that are evaluated as bad in its own eyes when both the recipient and donor are bad in its own eyes [i.e., $Norm(Action(B, B), B, B) = \mathrm{B}$; see Table II]. Thus, $r^*(p) = 1 - e_2$ is not satisfied for L7 and L8.

Figure 7 shows the cooperation probabilities and ESS regions of the leading-eight norms for $e_1 = 0$ for the sake of comparison with Fig. 4. From Fig. 7, we see that the results of Fig. 4 look robust against the change of the action error rate from $e_1 = 0.03$ to $e_1 = 0$. However, panel B in Fig. 7 for Type-1 norms suggests that the ESS condition in the limit of $e_2 \to 0$ for those norms is $1 < b/c < 2$. Below, we analytically prove this result.

2. Analytical results

a. Equilibrium distribution

Let us assume $e_1 = 0$ and consider the limit, $N \to \infty$. Our temporal goal is to derive the marginal distribution, $\phi^*(p) := \phi^*(p, 1) + \phi^*(p, 0)$, at the equilibrium when there are no mutants.

First, we consider the quantity $\sum_{s''} \sum_Y h^{CY}(p', s'') \phi^*(p'', s'')$; i.e., the probability that a randomly chosen donor

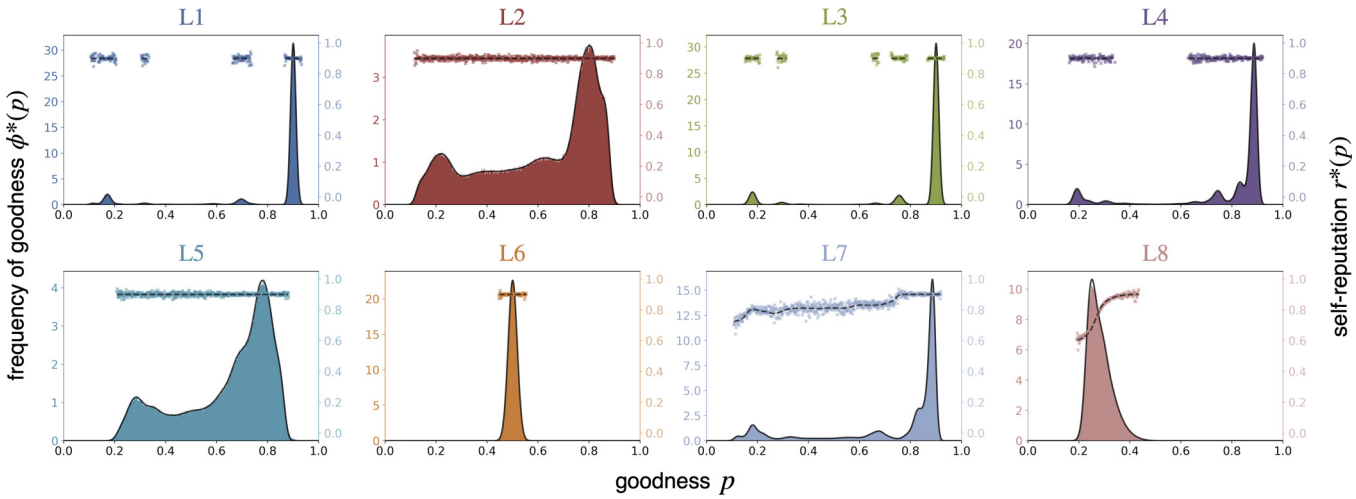

FIG. 6. The equilibrium states of $\phi^*(p, s)$. All the panels are output based on the same parameters as Fig. 3 except for $e_1 = 0$.

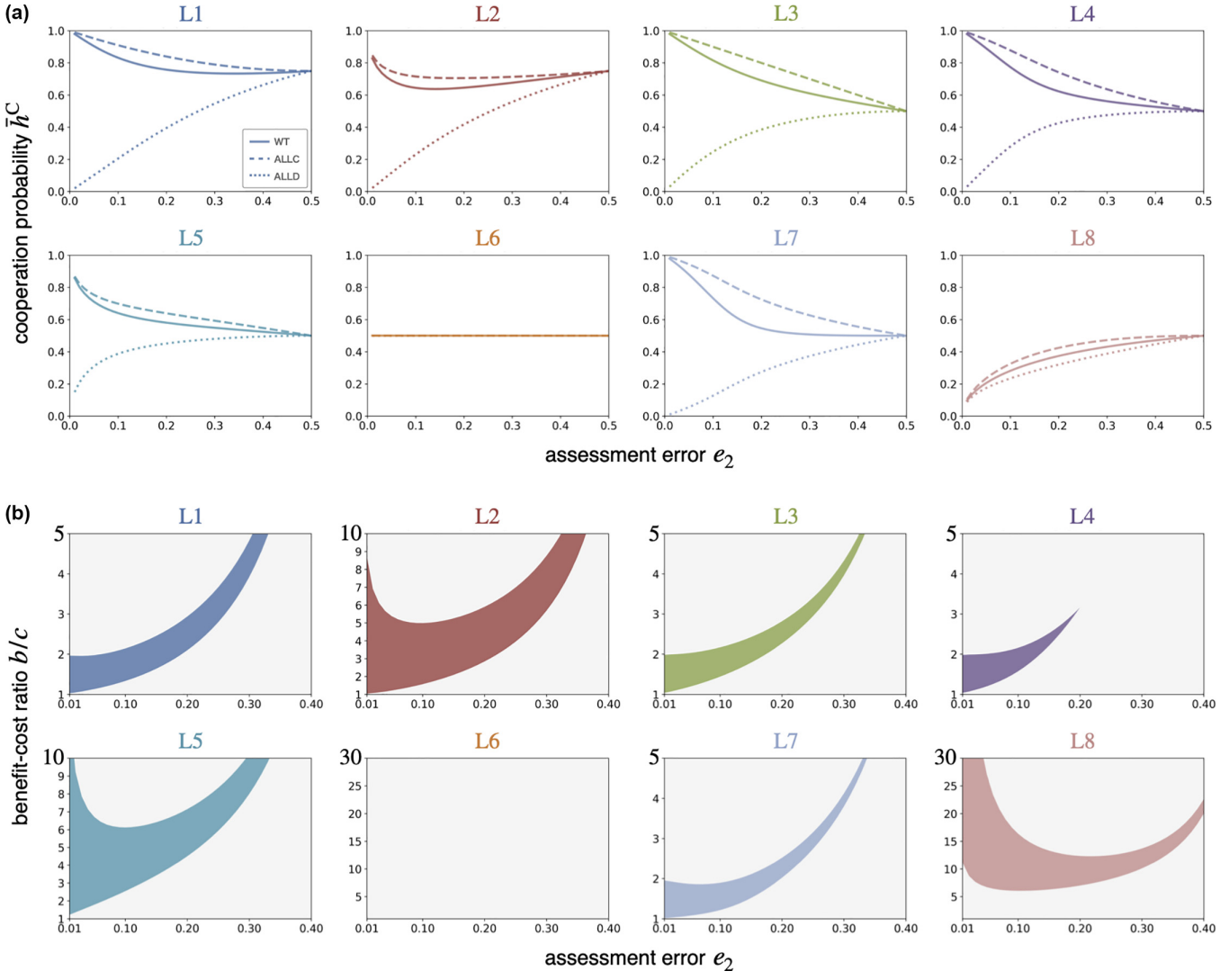

FIG. 7. (a) Cooperation probabilities of the leading-eight norms. (b) ESS regions of the leading-eight norms. Numerical calculations are based on the same parameters as Fig. 4 except for $e_1 = 0$.

has goodness p'' and it cooperates with the recipient whose goodness is p' at the equilibrium. We notice that the following properties are satisfied:

(i) For L1 and L2, since we assume $e_1 = 0$, donors always choose the action evaluated as good by themselves unless an assessment error occurs. Thus, $r^*(p) = 1 - e_2$ holds for all p, and therefore

$$\phi^*(p'', s'') = \begin{cases} \phi^*(p'')(1-e_2) & \text{if } s''=1 \\ \phi^*(p'')e_2 & \text{if } s''=0 \end{cases} \tag{A1}$$

holds. Hence we obtain

$$\begin{aligned} &\sum_{s''}\sum_{Y} h^{CY}(p', s'')\phi^*(p'', s'') \\ &= \phi^*(p'')\sum_{Y}\left\{ \underbrace{h^{CY}(p', 1)}_{=p'}(1-e_2) + \underbrace{h^{CY}(p', 0)}_{=1} e_2 \right\} \\ &= \phi^*(p''))\{p' + (1-p')e_2\}. \end{aligned} \tag{A2}$$

(ii) For L3–L8, donors do not change their actions depending on the self-reputation. So, we can obtain $\sum_Y h^{CY}(p', s'') = p'$. Thus, we obtain

$$\sum_{s''}\sum_{Y} h^{CY}(p', s'')\phi^* p'', s'') = p'\sum_{s''}\phi^*(p'', s'') = \phi^*(p'')p'. \tag{A3}$$

To summarize these properties, we define $h^C(p')$ as follows:

$$\begin{aligned} &\sum_{s''}\sum_{Y} h^{CY}(p', s'')\phi^*(p'', s'') \\ &\quad = \begin{cases} \phi^*(p'')\{p' + (1-p')e_2\} & \text{(L1–L2)} \\ \phi^*(p'')p' & \text{(L3–L8)} \end{cases} \\ &\quad = \phi^*(p'')\underbrace{\{p' + (1-p')e_2 a^{BB}\}}_{=:h^C(p')}, \end{aligned} \tag{A4}$$

because $a^{BB} = 1$ for L1 and L2 and $a^{BB} = 0$ for L3–L6. Note that $h^C(p')$ represents the probability that a random donor who faces the recipient of goodness p' cooperates with him or her at the equilibrium. We also define $h^D(p') := 1 - h^C(p')$. With these, the equilibrium marginal distribution of goodness is given, from Eqs. (5), by

$$\begin{aligned} \phi^*(p) &= \sum_s \phi^*(p, s) \\ &= \int_0^1 dp' \sum_{s'} \int_0^1 dp'' \sum_{s''}\sum_A\sum_Y h^{AY}(p', s'')\delta(p - f^A(p', p''))\phi^*(p', s')\phi^*(p'', s'') \\ &= \int_0^1 dp' \int_0^1 dp'' \sum_A \underbrace{\sum_{s''}\sum_Y h^{AY}(p', s'')\phi^*(p'', s'')}_{=h^A(p')\phi^*(p'')} \delta(p - f^A(p', p''))\phi^*(p') \\ &= \int_0^1 dp' \int_0^1 dp'' \sum_A h^A(p')\delta(p - f^A(p', p''))\phi^*(p')\phi^*(p''). \end{aligned} \tag{A5}$$

Here, we used the fact that, in the limit of $N \to \infty$, the Gaussian $g(p; f^A(p', p''), V)$ converges to the Dirac δ function $\delta(p - f^A(p', p''))$.

We now consider the case of $e_2 \ll 1$. To calculate the ESS condition of Type-1 norms (L1, L3, L4, and L7), we need to find the solution to Eq. (A5) up to the order of e_2. In other words, we neglect all the terms of $O(e_2^2)$ and higher in the following calculations. Guessing from Fig. 6, we heuristically seek the solution in the following form:

$$\phi^*(p) = \sum_{i=1}^{m_1} \kappa_i \delta_{1-k_i e_2}(p) + \sum_{i=1}^{m_2} \lambda_i \delta_{l_i e_2}(p). \tag{A6a}$$

Here, $\delta_x(p) := \delta(p - x)$ is the Dirac delta function that has a unit mass at $p = x$, and κ and λ satisfy

$$\kappa_i = O(1) \text{ or } O(e_2), \quad \lambda_i = O(e_2), \quad \sum_{i=1}^{m_1}\kappa_i + \sum_{i=1}^{m_2}\lambda_i = 1. \tag{A6b}$$

In words, Eqs. (A6) says that the equilibrium distribution ϕ^* is given as a finite sum of masses κ_i at positions $p = 1 - k_i e_2$ and masses λ_i at positions $p = l_i e_2$, where masses κ_i are either $O(1)$ or $O(e_2)$ and masses λ_j are $O(e_2)$.

In fact, calculations in Appendix B show that we can find the solution to Eq. (A5) in the form of Eqs. (A6) for each of the Type-1 norms (L1, L3, L4, and L7) separately as

$$\begin{aligned} \phi^*_{L1}(p) &= (1-2e_2)\delta_{1-e_2}(p) + e_2\delta_{1-4e_2}(p) + e_2\delta_{2e_2}(p), \\ \phi^*_{L3}(p) &= (1-2e_2)\delta_{1-e_2}(p) + e_2\delta_{1-3e_2}(p) + e_2\delta_{2e_2}(p), \\ \phi^*_{L4}(p) &= (1-3e_2)\delta_{1-e_2}(p) + e_2\delta_{1-2e_2}(p) + e_2\delta_{1-3e_2}(p) \\ &\quad + e_2\delta_{2e_2}(p), \\ \phi^*_{L7}(p) &= (1-3e_2)\delta_{1-e_2}(p) + e_2\delta_{1-2e_2}(p) + e_2\delta_{1-4e_2}(p) \\ &\quad + e_2\delta_{2e_2}(p). \end{aligned} \tag{A7}$$

Note that these solutions are correct up to $O(e_2)$. We also remark that with the same methodology, we cannot derive the equilibrium distribution for Type-2 norms (L2 and L5), because we eventually find that sums in Eq. (A6a) cannot be finite but require infinite sums, which is consistent with our

observation that the equilibrium distribution $\phi^*(p)$ for Type-2 norms look continuous, not discrete.

b. A calculation of payoffs

With the equilibrium distribution given by Eq. (A7), we can calculate various quantities for studying the invasion condition of ALLC and ALLD mutants. In the following calculation we use the facts that Type-1 norms and the corresponding action rules are given by $\boldsymbol{n}^{\mathrm{C}} = (1-e_2, 1-e_2, 1-e_2, n^{\mathrm{CBB}})$, $\boldsymbol{n}^{\mathrm{D}} = (e_2, e_2, 1-e_2, n^{\mathrm{DBB}})$, and $\boldsymbol{a} = (1, 1, 0, a^{\mathrm{BB}})$.

Expected payoff of wild types. From Eq. (A7), the average goodness $\bar{p}$ of Type-1 wild types is calculated as

$$\bar{p} = \int_0^1 p\phi^*(p)\mathrm{d}p = 1 - 2e_2 + o(e_2) \tag{A8}$$

for all Type-1 norms. The probability of giving or receiving cooperation $\bar{h}^{\mathrm{C}}$ in the monomorphic population of residents at the equilibrium is

$$\bar{h}^{\mathrm{C}} = h^{\mathrm{C}}(\bar{p}) = \bar{p} + (1-\bar{p})e_2 a^{\mathrm{BB}} = 1 - 2e_2 + o(e_2) \tag{A9}$$

for all Type-1 norms. Therefore, from Eq. (7), the expected payoff of wild types is

$$u = (1-2e_2)(b-c) + o(e_2) \tag{A10}$$

for all Type-1 norms.

Expected payoff of ALLC mutants. Consider rare ALLC mutants invading one of the Type-1 norms. By putting $a_{\mathrm{M}} = 1$ in Eq. (9), the average goodness of these mutants is

$$\begin{aligned}\bar{p}_{\mathrm{M}} &= f^{\mathrm{C}}(\bar{p}, \bar{p}_{\mathrm{M}}) \\ &= \begin{cases} 1-e_2 & \text{(for L1 and L3)} \\ 1-e_2-(1-2e_2)(1-\bar{p})(1-\bar{p}_{\mathrm{M}}) & \text{(for L4 and L7).} \end{cases}\end{aligned} \tag{A11}$$

For L4 and L7, this is further calculated as

$$\begin{aligned}\bar{p}_{\mathrm{M}} &= \frac{1-e_2-(1-2e_2)(1-\bar{p})}{1-(1-2e_2)(1-\bar{p})} \\ &= \frac{1-e_2-(1-2e_2)2e_2}{1-(1-2e_2)2e_2} + o(e_2) \\ &= 1 - e_2 + o(e_2).\end{aligned} \tag{A12}$$

From Eq. (6), the expected probability that a wild type cooperates with an ALLC mutant is

$$\bar{h}^{\mathrm{C}}_{\mathrm{M}} = h^{\mathrm{C}}(\bar{p}_{\mathrm{M}}) = \bar{p}_{\mathrm{M}} + (1-\bar{p}_{\mathrm{M}})e_2 a^{\mathrm{BB}} = 1 - e_2 + o(e_2) \tag{A13}$$

for all Type-1 wild types. By putting $a_{\mathrm{M}} = 1$ in Eq. (11) we obtain the expected payoff of ALLC mutants as

$$u_{\mathrm{M}} = (1-e_2)b - c + o(e_2) \tag{A14}$$

for all Type-1 wild types. Thus, by comparing Eq. (A10) and Eq. (A14), wild types are resistant to ALLC mutants if and only if

$$b/c < 2 \tag{A15}$$

in the limit of $e_2 \to 0$ for all Type-1 wild types.

Expected payoff of ALLD mutants. Consider rare ALLD mutants invading one of the Type-1 norms. By putting $a_{\mathrm{M}} = 0$ in Eq. (9), the average goodness of these mutants is

$$\begin{aligned}\bar{p}_{\mathrm{M}} &= f^{\mathrm{D}}(\bar{p}, \bar{p}_{\mathrm{M}}) \\ &= e_2\bar{p} + (1-e_2)(1-\bar{p})\bar{p}_{\mathrm{M}} \\ &\quad + n^{\mathrm{DBB}}(1-\bar{p})(1-\bar{p}_{\mathrm{M}}) \\ \iff \bar{p}_{\mathrm{M}} &= \frac{e_2\bar{p} + n^{\mathrm{DBB}}(1-\bar{p})}{1-(1-e_2)(1-\bar{p}) + n^{\mathrm{DBB}}(1-\bar{p})} \\ &= \frac{e_2(1+2n^{\mathrm{DBB}})}{1-2e_2(1-n^{\mathrm{DBB}})} + o(e_2).\end{aligned} \tag{A16}$$

Since $n^{\mathrm{DBB}} = e_2$ for L1 and L7, and $n^{\mathrm{DBB}} = 1-e_2$ for L3 and L4, we have

$$\bar{p}_{\mathrm{M}} = \begin{cases} e_2 + o(e_2) & \text{(for L1 and L7)} \\ 3e_2 + o(e_2) & \text{(for L3 and L4).} \end{cases} \tag{A17}$$

From Eq. (6), the expected probability that a wild type cooperates with an ALLD mutant is

$$\bar{h}^{\mathrm{C}}_{\mathrm{M}} = h^{\mathrm{C}}(\bar{p}_{\mathrm{M}}) = \bar{p}_{\mathrm{M}} + (1-\bar{p}_{\mathrm{M}})e_2 a^{\mathrm{BB}}. \tag{A18}$$

Since $a^{\mathrm{BB}} = 1$ for L1, and $a^{\mathrm{BB}} = 0$ for L3, L4, and L7, we have

$$\bar{h}^{\mathrm{C}}_{\mathrm{M}} = \begin{cases} 2e_2 + o(e_2) & \text{(for L1)} \\ 3e_2 + o(e_2) & \text{(for L3 and L4)} \\ e_2 + o(e_2) & \text{(for L7).} \end{cases} \tag{A19}$$

By putting $a_{\mathrm{M}} = 0$ in Eq. (11) we obtain the expected payoff of ALLD mutants as

$$u_{\mathrm{M}} = \begin{cases} 2e_2 b + o(e_2) & \text{(for L1)} \\ 3e_2 b + o(e_2) & \text{(for L3 and L4)} \\ e_2 b + o(e_2) & \text{(for L7).} \end{cases} \tag{A20}$$

Thus, by comparing Eq. (A10) and Eq. (A20), wild types are resistant to ALLD mutants if and only if

$$b/c > 1. \tag{A21}$$

in the limit of $e_2 \to 0$ for all Type-1 wild types.

APPENDIX B: DETAILED CALCULATION OF APPENDIX A

To solve Eq. (A5) and to obtain the equilibrium distribution $\phi^*(p)$, we iteratively solve the functional recursion

$$\begin{aligned}\phi_{t+1}(p) = \int_0^1 dp' \int_0^1 dp'' \sum_A h^A(p') \\ \times \delta(p - f^A(p', p''))\phi_t(p')\phi_t(p''),\end{aligned} \tag{B1}$$

where $t = 0, 1, 2, \ldots$ is a non-negative integer, by always assuming that the function $\phi_t(p)$ is given in the form of Eqs. (A6); i.e., a finite sum of Dirac δ functions. If we find that $\phi_{t^*+1}(p) = \phi_{t^*}(p)$ holds at some $t^* \geqslant 0$, then it is a solution to Eq. (A5).

Suppose that

$$\phi_t(p) = \sum_{i=1}^{m_1} \kappa_i \delta_{1-k_i e_2}(p) + \sum_{i=1}^{m_2} \lambda_i \delta_{l_i e_2}(p) \tag{B2}$$

holds. Then, from Eq. (B1) we have

$$\begin{aligned}\phi_{t+1}(p) = &\sum_{i=1}^{m_1}\sum_{j=1}^{m_1}\kappa_i\kappa_j\{h^{\mathrm{C}}(p')\delta_{f^{\mathrm{C}}(p',p'')}(p) + h^{\mathrm{D}}(p')\delta_{f^{\mathrm{D}}(p',p'')}(p)\}|_{p'=1-k_ie_2,p''=1-k_je_2}\\ &+\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\kappa_i\lambda_j\{h^{\mathrm{C}}(p')\delta_{f^{\mathrm{C}}(p',p'')}(p) + h^{\mathrm{D}}(p')\delta_{f^{\mathrm{D}}(p',p'')}(p)\}|_{p'=1-k_ie_2,p''=l_je_2}\\ &+\sum_{i=1}^{m_2}\sum_{j=1}^{m_1}\lambda_i\kappa_j\{h^{\mathrm{C}}(p')\delta_{f^{\mathrm{C}}(p',p'')}(p) + h^{\mathrm{D}}(p')\delta_{f^{\mathrm{D}}(p',p'')}(p)\}|_{p'=l_ie_2,p''=1-k_je_2}\\ &+\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\lambda_i\lambda_j\{h^{\mathrm{C}}(p')\delta_{f^{\mathrm{C}}(p',p'')}(p) + h^{\mathrm{D}}(p')\delta_{f^{\mathrm{D}}(p',p'')}(p)\}|_{p'=l_ie_2,p''=l_je_2}. \end{aligned} \tag{B3}$$

Each line is further calculated as follows:

Case 1. Interaction of a good recipient and good donor. We consider an event that the recipient with goodness $p' = 1 - k_ie_2$ and the donor with goodness $p'' = 1 - k_je_2$ interact. This event occurs with probability $\kappa_i\kappa_j$, which can be either $O(1)$, $O(e_2)$, or $O(e_2^2)$. We consider only terms of up to $O(e_2)$ in the following. From Eq. (A4), the donor cooperates with probability

$$\begin{aligned}h^{\mathrm{C}}(p') &= p' + (1-p')e_2a^{\mathrm{BB}}\\ &= 1 - k_ie_2 + (k_ie_2)e_2a^{\mathrm{BB}}\\ &= 1 - k_ie_2 + o(e_2).\end{aligned} \tag{B4}$$

In this case, the donor's goodness is updated to

$$\begin{aligned}f^{\mathrm{C}}(p',p'') &= (1-e_2) + \{n^{\mathrm{CBB}} - (1-e_2)\}(1-p')(1-p'')\\ &= (1-e_2) + \{n^{\mathrm{CBB}} - (1-e_2)\}k_ik_je_2^2\\ &= 1 - e_2 + o(e_2).\end{aligned} \tag{B5}$$

On the other hand, the donor defects with probability

$$h^{\mathrm{D}}(p') = 1 - h^{\mathrm{C}}(p') = k_ie_2 + o(e_2). \tag{B6}$$

In this case, the donor's goodness is updated to

$$\begin{aligned}f^{\mathrm{D}}(p',p'') &= e_2p' + (1-e_2)(1-p')p''\\ &\quad + n^{\mathrm{DBB}}(1-p')(1-p'')\\ &= e_2(1-k_ie_2) + (1-e_2)k_ie_2(1-k_je_2)\\ &\quad + n^{\mathrm{DBB}}k_ik_je_2^2\\ &= (k_i+1)e_2 + o(e_2).\end{aligned} \tag{B7}$$

Therefore, the first line of Eq. (B3) can be calculated as

$$\sum_{i=1}^{m_1}\sum_{j=1}^{m_1}\kappa_i\kappa_j\{(1-k_ie_2)\delta_{1-e_2}(p) + k_ie_2\delta_{(k_i+1)e_2}(p)\} + o(e_2). \tag{B8}$$

Case 2. Interaction of a good recipient and bad donor. We consider an event that the recipient with goodness $p' = 1 - k_ie_2$ and the donor with goodness $p'' = l_je_2$ interact. This event occurs with probability $\kappa_i\lambda_j$, which can be either $O(e_2)$ or $O(e_2^2)$. We consider only terms of up to $O(1)$ in the following, because any terms of $O(e_2)$ or higher, after the multiplication by the factor $\kappa_i\lambda_j$, become $O(e_2^2)$ or higher. From Eq. (A4), the donor cooperates with probability

$$\begin{aligned}h^{\mathrm{C}}(p') &= p' + (1-p')e_2a^{\mathrm{BB}}\\ &= 1 - k_ie_2 + (k_ie_2)e_2a^{\mathrm{BB}}\\ &= 1 + o(1).\end{aligned} \tag{B9}$$

In this case, the donor's goodness is updated to

$$\begin{aligned}f^{\mathrm{C}}(p',p'') &= (1-e_2) + \{n^{\mathrm{CBB}} - (1-e_2)\}(1-p')(1-p'')\\ &= (1-e_2) + \{n^{\mathrm{CBB}} - (1-e_2)\}k_ie_2(1-l_je_2)\\ &= 1 - \{(1-n^{\mathrm{CBB}})k_i + 1\}e_2 + o(e_2)\\ &= 1 - \{(1-\tilde{n}^{\mathrm{CBB}})k_i + 1\}e_2 + o(e_2),\end{aligned} \tag{B10}$$

where $\tilde{n}^{\mathrm{CBB}}$ represents the value of n^{CBB} evaluated at $e_2 = 0$. That is, if $n^{\mathrm{CBB}} = 1 - e_2$ then $\tilde{n}^{\mathrm{CBB}} = 1$, and if $n^{\mathrm{CBB}} = e_2$ then $\tilde{n}^{\mathrm{CBB}} = 0$. We do not have to discuss the event of defection because the event occurs with probability $h^{\mathrm{D}}(p') = 1 - h^{\mathrm{C}}(p') = o(1)$. Therefore, the second line of Eq. (B3) can be calculated as

$$\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\kappa_i\lambda_j\delta_{1-\{(1-\tilde{n}^{\mathrm{CBB}})k_i+1\}e_2}(p) + o(e_2). \tag{B11}$$

Case 3. Interaction of a bad recipient and good donor. We consider an event that the recipient with goodness $p' = l_ie_2$ and the donor with goodness $p'' = 1 - k_je_2$ interact. This event occurs with probability $\lambda_i\kappa_j$, which can be either $O(e_2)$ or $O(e_2^2)$. We consider only terms of up to $O(1)$ in the following, because any terms of $O(e_2)$ or higher, after the multiplication by the factor $\lambda_i\kappa_j$, become $O(e_2^2)$ or higher. From Eq. (A4), the donor cooperates with probability

$$\begin{aligned}h^{\mathrm{D}}(p') &= 1 - \{p' + (1-p')e_2a^{\mathrm{BB}}\}\\ &= 1 - \{l_ie_2 + (1-l_ie_2)e_2a^{\mathrm{BB}}\}\\ &= 1 + o(1).\end{aligned} \tag{B12}$$

TABLE V. How to calculate $\phi_{t+1}(p)$ from $\phi_t(p)$ for Type-1 norms.

p'	p''	Event scale	A	$h^{\mathrm{A}}(p')$	$p = f^A(p', p'')$
$1-k_ie_2$	$1-k_je_2$	$O(1)$ or $O(e_2)$	C	$1-k_ie_2$	$1-e_2$
			D	k_ie_2	$(k_i+1)e_2$
$1-k_ie_2$	l_je_2	$O(e_2)$	C	1	$1-\{(1-\tilde{n}^{\mathrm{CBB}})k_i+1\}e_2$
l_ie_2	$1-k_je_2$	$O(e_2)$	D	1	$1-\{l_i+(1-\tilde{n}^{\mathrm{DBB}})k_j+1\}e_2$

In this case, the donor's goodness is updated to

$$\begin{aligned} f^{\mathrm{D}}(p', p'') &= e_2p' + (1-e_2)(1-p')p'' \\ &\quad + n^{\mathrm{DBB}}(1-p')(1-p'') \\ &= l_ie_2^2 + (1-e_2)(1-l_ie_2)(1-k_je_2) \\ &\quad + n^{\mathrm{DBB}}(1-l_ie_2)k_je_2 \\ &= 1-\{l_i+(1-n^{\mathrm{DBB}})k_j+1\}e_2 + o(e_2) \\ &= 1-\{l_i+(1-\tilde{n}^{\mathrm{DBB}})k_j+1\}e_2 + o(e_2), \end{aligned} \tag{B13}$$

where $\tilde{n}^{\mathrm{DBB}}$ represents the value of n^{DBB} evaluated at $e_2 = 0$. That is, if $n^{\mathrm{DBB}} = 1-e_2$ then $\tilde{n}^{\mathrm{DBB}} = 1$, and if $n^{\mathrm{DBB}} = e_2$ then $\tilde{n}^{\mathrm{DBB}} = 0$. We do not have to discuss the event of cooperation because the event occurs with probability $h^{\mathrm{C}}(p') = 1-h^{\mathrm{D}}(p') = o(1)$. Therefore, the third line of Eq. (B3) can be calculated as

$$\sum_{i=1}^{m_2}\sum_{j=1}^{m_1}\lambda_i\kappa_j\delta_{1-\{l_i+(1-\tilde{n}^{\mathrm{DBB}})k_j+1\}e_2}(p) + o(e_2). \tag{B14}$$

Case 4. Interaction of a bad recipient and bad donor. We do not have to consider an event that the recipient with goodness $p' = l_ie_2$ and the donor with goodness $p'' = l_je_2$ interact, because the event occurs with the probability of $\lambda_i\lambda_j = O(e_2^2)$. Therefore, the fourth line of Eq. (B3) is simply $o(e_2)$.

The calculations in Cases 1 to 4 are summarized in Table V. By using these results, we now solve Eq. (B1) iteratively. As an initial function, we choose

$$\phi_0(p) = \delta_{1-e_2}(p). \tag{B15}$$

A direct calculation shows, up to $O(e_2)$, that

$$\begin{aligned} \phi_1(p) &= (1-e_2)\delta_{1-e_2}(p) + e_2\delta_{2e_2}(p), \\ \phi_2(p) &= (1-3e_2)\delta_{1-e_2}(p) + e_2\delta_{1-(2-\tilde{n}^{\mathrm{CBB}})e_2}(p) \\ &\quad + e_2\delta_{1-(4-\tilde{n}^{\mathrm{DBB}})e_2}(p) + e_2\delta_{2e_2}(p), \\ \phi_3(p) &= (1-3e_2)\delta_{1-e_2}(p) + e_2\delta_{1-(2-\tilde{n}^{\mathrm{CBB}})e_2}(p) \\ &\quad + e_2\delta_{1-(4-\tilde{n}^{\mathrm{DBB}})e_2}(p) + e_2\delta_{2e_2}(p) \quad (=\phi_2(p)), \end{aligned} \tag{B16}$$

and therefore the solution to Eq. (A5) is

$$\begin{aligned} \phi^*(p) &= (1-3e_2)\delta_{1-e_2}(p) + e_2\delta_{1-(2-\tilde{n}^{\mathrm{CBB}})e_2}(p) \\ &\quad + e_2\delta_{1-(4-\tilde{n}^{\mathrm{DBB}})e_2}(p) + e_2\delta_{2e_2}(p). \end{aligned} \tag{B17}$$

Substituting

$$(\tilde{n}^{\mathrm{CBB}}, \tilde{n}^{\mathrm{DBB}}) = \begin{cases} (1,0) & (\text{for L1}) \\ (1,1) & (\text{for L3}) \\ (0,1) & (\text{for L4}) \\ (0,0) & (\text{for L7}) \end{cases} \tag{B18}$$

in Eq. (B17) gives us Eq. (A7).

APPENDIX C: ESTIMATION OF ERROR IN TRANSITION OF EQUILIBRIUM DISTRIBUTION

This section evaluates the truncation error in calculating Eqs. (5) and Eq. (9). Even if ϕ^* on the right-hand side (RHS) of Eqs. (5) is properly normalized, ϕ^* on the left-hand side (LHS) is not normalized, because it has some positive value outside the interval $p \in [0,1]$. Here we estimate how much mass ϕ^* on the LHS of Eqs. (5) has in the intervals $(-\infty, 0)$ and $(1, \infty)$. A key observation for this error estimation is that f^A in Eq. (3a) always satisfies $e_2 \leqslant f^A \leqslant 1-e_2$, because each component of the vector $\boldsymbol{n}^A$ is either e_2 or $1-e_2$ and because f^A is a weighted average of these values. First, we consider total weights in the interval $(1, \infty)$. For $p > 1$, we obtain

$$\begin{aligned} \phi^*(p) &= \sum_s \phi^*(p,s) = \int_0^1 \mathrm{d}p' \sum_{s'} \int_0^1 \mathrm{d}p'' \sum_{s''} \sum_A \sum_Y h^{AY}(p', s'') \underbrace{g(p; f^A(p', p''), V)}_{\leqslant g(p; 1-e_2, V)} \phi^*(p', s')\phi^*(p'', s'') \\ &\leqslant g(p; 1-e_2, V) \int_0^1 \mathrm{d}p' \sum_{s'} \int_0^1 \mathrm{d}p'' \sum_{s''} \underbrace{\sum_A \sum_Y h^{AY}(p', s'')}_{=1} \phi^*(p', s')\phi^*(p'', s'') \\ &= g(p; 1-e_2, V) \underbrace{\int_0^1 \mathrm{d}p' \sum_{s'} \int_0^1 \mathrm{d}p'' \sum_{s''} \phi^*(p', s')\phi^*(p'', s'')}_{=1} = g(p; 1-e_2, V). \end{aligned} \tag{C1}$$

Therefore, the leak to the interval $(1, \infty)$ is evaluated as

$$\begin{aligned}
\int_1^\infty \mathrm{d}p\phi^*(p) &\leqslant \int_1^\infty \mathrm{d}p \frac{1}{\sqrt{2\pi V}} \exp\left[-\frac{[p-(1-e_2)]^2}{2V}\right] \\
&= \int_0^\infty \mathrm{d}\tilde{p} \frac{1}{\sqrt{2\pi V}} \exp\left[-\frac{(\tilde{p}+e_2)^2}{2V}\right] \\
&= \int_0^\infty \mathrm{d}\tilde{p} \frac{1}{\sqrt{2\pi V}} \underbrace{\exp\left[-\frac{\tilde{p}^2}{2V}\right]}_{\leqslant 1} \exp\left[-\frac{e_2\tilde{p}}{V}\right] \exp\left[-\frac{e_2^2}{2V}\right] \\
&\leqslant \frac{1}{\sqrt{2\pi V}} \exp\left[-\frac{e_2^2}{2V}\right] \int_0^\infty \mathrm{d}\tilde{p} \exp\left[-\frac{e_2\tilde{p}}{V}\right] \\
&= \frac{1}{\sqrt{2\pi V}} \exp\left[-\frac{e_2^2}{2V}\right] \frac{V}{e_2} \\
&\leqslant \frac{1}{\sqrt{2\pi e_2(N-1)}} \exp\left[-\frac{e_2(N-1)}{2}\right] [=:\epsilon(e_2, N)]. \qquad \text{(C2)}
\end{aligned}$$

Here, in the first line, we have used Eq. (C1). From the first to the second line, we have substituted $\tilde{p} = p - 1$. From the fifth to the sixth lines, we use $V = e_2(1-e_2)/(N-1)$.

In a similar way, we can prove that the leak to the other interval $(-\infty, 0)$ is upper-bounded by $\epsilon(e_2, N)$. Therefore, the total error is upper-bounded by $2\epsilon(e_2, N)$.

We provide some representative examples of $\epsilon(e_2, N)$. For realistic parameters $(e_2, N) = (0.1, 100)$, we have $\epsilon(e_2, N) \leqslant 10^{-3}$. For the parameter used in our simulations $(e_2, N) = (0.1, 800)$, we have $\epsilon(e_2, N) \leqslant 10^{-18}$. Thus, the truncation of the Gaussian functions is estimated to be negligibly small.

This bound of the approximation error, i.e., $\epsilon(e_2, N)$, is useful for two numerical calculations in this paper. First, when we numerically compute the equilibrium distribution ϕ^* by iterating Eqs. (5a) and (5b), $2\epsilon(e_2, N)$ gives the upper bound of errors per single iteration. Second, in the calculation of Eq. (9), the difference between its LHS and RHS is smaller than $2\epsilon(e_2, N)$.

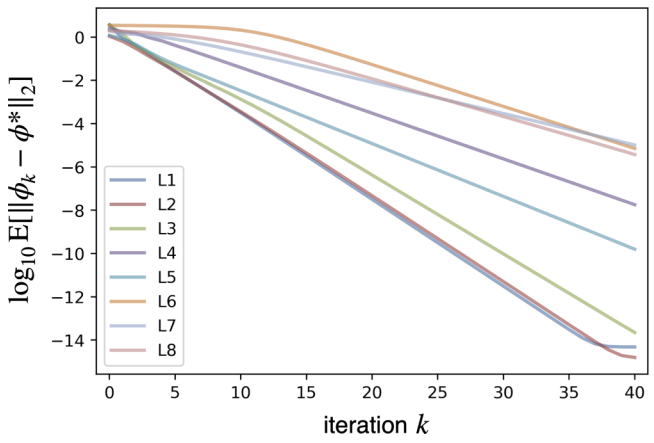

FIG. 8. The convergence to the equilibrium distribution ϕ^*. We have randomly generated N_{sample} (=100) samples. Here, the equilibrium distribution $\phi^* := \mathrm{E}[\phi_{50}]$ is estimated as the ensemble average of ϕ_{50} (the distribution after 50 iterations). For each sample, the L^2 distance between ϕ_k and ϕ^* is calculated as $\|\phi_k - \phi^*\|_2$. Finally, the indicator of convergence (i.e., $\mathrm{E}[\|\phi_k - \phi^*\|_2]$) is formulated by the ensemble average of this L^2 distance over all N_{sample} (=100) samples. The initial condition of each sample, ϕ_0, is generated by the following procedure; we have three random numbers $(r_1, r_2, r_3) \sim [0, 1]^3$, and define ϕ_0 as $\phi_0(p, 1) = 2r_1\{r_2(1-p) + (1-r_2)p\}$ and $\phi_0(p, 0) = 2(1-r_1)\{r_3(1-p) + (1-r_3)p\}$.

APPENDIX D: CONVERGENCE TO EQUILIBRIUM DISTRIBUTION

To calculate ϕ^* numerically in Eqs. (5). We replace ϕ^*s on the RHSs of Eqs. (5) with ϕ_k and those on the LHSs with ϕ_{k+1} and calculate recursions as $\phi_0 \to \phi_1 \to \phi_2 \to \cdots$ from an initial state ϕ_0. We have set ϕ_0 as the uniform distribution function, i.e., $\phi_0(p, s) = 1/2$ for all $p \in [0, 1]$ and $s \in \{0, 1\}$. We stop this iteration at Kth step if $\|\phi_K - \phi_{K-1}\|_2 \leqslant \epsilon$ is satisfied for sufficiently small $\epsilon (= 10^{-6})$. Here, $\|\cdot\|_2$ represents the L^2 norm (distance), defined as $\|\phi_K - \phi_{K-1}\|_2 := \{\int_p \mathrm{d}p \sum_s [\phi_K(p, s) - \phi_{K-1}(p, s)]^2\}^{1/2}$.

For our numerical calculation, we approximate $\phi_k(p, s)$ by a step function, which is discretized by N (=800) meshes for p. Here, however, we can take a more coarse-grained mesh independent of N. If so, the computational cost to find the equilibrium becomes small and theoretically expected to be smaller than the cost of individual-based simulation.

Although the initial state ϕ_0 is set as the uniform distribution, the choice of it is optional. Indeed, Fig. 8 demonstrates the convergence to ϕ^* from various initial distributions, which are randomly generated. We see that the ensemble-average distance from the equilibrium distribution exponentially decays over iteration steps. Interestingly, the speed of convergence differs depending on the leading-eight norms: it seems (L1, L2) > L3 > L5 > L4 > (L6, L7, L8).

[1] R. Axelrod, *The Evolution of Cooperation* (Basic Books, New York, 1984).

[2] L. A. Dugatkin, *Cooperation Among Animals: An Evolutionary Perspective* (Oxford University Press, Oxford, England, 1997).

[3] M. Tomasello, *Why We Cooperate* (MIT Press, Cambridge, MA, 2009).

[4] S. Bowles and H. Gintis, *A Cooperative Species* (Princeton University Press, Princeton, NJ, 2011).

[5] R. L. Trivers, The evolution of reciprocal altruism, Q. Rev. Biol. **46**, 35 (1971).

[6] N. Henrich and J. P. Henrich, *Why Humans Cooperate: A Cultural and Evolutionary Explanation* (Oxford University Press, Oxford, England, 2007).

[7] C. Boehm, *Moral Origins: The Evolution of Virtue, Altruism, and Shame* (Basic Books, New York City, 2012).

[8] R. D. Alexander, *The Biology of Moral Systems* (Aldine de Gruyter, New York, 1987).

[9] M. A. Nowak and K. Sigmund, Evolution of indirect reciprocity by image scoring, Nature (London) **393**, 573 (1998).

[10] M. A. Nowak and K. Sigmund, Evolution of indirect reciprocity, Nature (London) **437**, 1291 (2005).

[11] N. Emler, *Gossip, Reputation, and Social Adaptation* (University Press of Kansas, Lawrence, Kansas, 1994).

[12] R. I. M. Dunbar, *Grooming, Gossip, and the Evolution of Language* (Harvard University Press, Cambridge, MA, 1998).

[13] R. I. M. Dunbar, Gossip in evolutionary perspective, Rev. Gen. Psychol. **8**, 100 (2004).

[14] C. Wedekind and M. Milinski, Cooperation through image scoring in humans, Science **288**, 850 (2000).

[15] M. Milinski, D. Semmann, and H.-J. Krambeck, Reputation helps solve the 'tragedy of the commons', Nature (London) **415**, 424 (2002).

[16] G. E. Bolton, E. Katok, and A. Ockenfels, Cooperation among strangers with limited information about reputation, J. Public Econ. **89**, 1457 (2005).

[17] I. Seinen and A. Schram, Social status and group norms: Indirect reciprocity in a repeated helping experiment, Eur. Econ. Rev. **50**, 581 (2006).

[18] R. D. Sommerfeld, H.-J. Krambeck, D. Semmann, and M. Milinski, Gossip as an alternative for direct observation in games of indirect reciprocity, Proc. Natl. Acad. Sci. USA **104**, 17435 (2007).

[19] P. Barclay, Harnessing the power of reputation: Strengths and limits for promoting cooperative behaviors, Evol. Psychol. **10**, 868 (2012).

[20] M. Feinberg, R. Willer, and M. Schultz, Gossip and ostracism promote cooperation in groups, Psychol. Sci. **25**, 656 (2014).

[21] J. Wu, D. Balliet, and P. A. Van Lange, Reputation, gossip, and human cooperation, Soc. Pers. Psychol. Compass **10**, 350 (2016).

[22] J. van Apeldoorn and A. Schram, Indirect reciprocity; a field experiment, PLoS One **11**, e0152076 (2016).

[23] M. A. Nowak and K. Sigmund, The dynamics of indirect reciprocity, J. Theor. Biol. **194**, 561 (1998).

[24] K. Panchanathan and R. Boyd, A tale of two defectors: The importance of standing for evolution of indirect reciprocity, J. Theor. Biol. **224**, 115 (2003).

[25] H. Ohtsuki and Y. Iwasa, How should we define goodness?–reputation dynamics in indirect reciprocity, J. Theor. Biol. **231**, 107 (2004).

[26] H. Ohtsuki and Y. Iwasa, The leading eight: Social norms that can maintain cooperation by indirect reciprocity, J. Theor. Biol. **239**, 435 (2006).

[27] H. Brandt and K. Sigmund, Indirect reciprocity, image scoring, and moral hazard, Proc. Natl. Acad. Sci. USA **102**, 2666 (2005).

[28] H. Brandt and K. Sigmund, The good, the bad and the discriminator–errors in direct and indirect reciprocity, J. Theor. Biol. **239**, 183 (2006).

[29] H. Ohtsuki and Y. Iwasa, Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation, J. Theor. Biol. **244**, 518 (2007).

[30] K. Sigmund, *The Calculus of Selfishness* (Princeton University Press, Princeton, NJ, 2010).

[31] I. Okada, A review of theoretical studies on indirect reciprocity, Games **11**, 27 (2020).

[32] H. Brandt and K. Sigmund, The logic of reprobation: Assessment and action rules for indirect reciprocation, J. Theor. Biol. **231**, 475 (2004).

[33] J. M. Pacheco, F. C. Santos, and F. A. C. Chalub, Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity, PLoS Comput. Biol. **2**, e178 (2006).

[34] S. Suzuki and E. Akiyama, Evolution of indirect reciprocity in groups of various sizes and comparison with direct reciprocity, J. Theor. Biol. **245**, 539 (2007).

[35] F. C. Santos, F. A. Chalub, and J. M. Pacheco, A multi-level selection model for the emergence of social norms, in *European Conference on Artificial Life* (Springer, Berlin, Heidelberg, Germany, 2007), pp. 525–534..

[36] F. Fu, C. Hauert, M. A. Nowak, and L. Wang, Reputation-based partner choice promotes cooperation in social networks, Phys. Rev. E **78**, 026117 (2008).

[37] S. Suzuki and E. Akiyama, Evolutionary stability of first-order-information indirect reciprocity in sizable groups, Theor. Popul. Biol. **73**, 426 (2008).

[38] S. Uchida and K. Sigmund, The competition of assessment rules for indirect reciprocity, J. Theor. Biol. **263**, 13 (2010).

[39] H. Ohtsuki, Y. Iwasa, and M. A. Nowak, Reputation effects in public and private interactions, PLoS Comput. Biol. **11**, e1004527 (2015).

[40] F. P. Santos, F. C. Santos, and J. M. Pacheco, Social norms of cooperation in small-scale societies, PLoS Comput. Biol. **12**, e1004709 (2016).

[41] F. P. Santos, J. M. Pacheco, and F. C. Santos, Evolution of cooperation under indirect reciprocity and arbitrary exploration rates, Sci. Rep. **6**, 37517 (2016).

[42] T. Sasaki, I. Okada, and Y. Nakai, The evolution of conditional moral assessment in indirect reciprocity, Sci. Rep. **7**, 41870 (2017).

[43] F. P. Santos, F. C. Santos, and J. M. Pacheco, Social norm complexity and past reputations in the evolution of cooperation, Nature (London) **555**, 242 (2018).

[44] C. Xia, C. Gracia-Lázaro, and Y. Moreno, Effect of memory, intolerance, and second-order reputation on cooperation, Chaos **30**, 063122 (2020).

[45] F. P. Santos, J. M. Pacheco, and F. C. Santos, The complexity of human cooperation under indirect reciprocity, Philos. Trans. R. Soc., B **376**, 20200291 (2021).
[46] S. Podder, S. Righi, and K. Takács, Local reputation, local selection, and the leading eight norms, Sci. Rep. **11**, 16560 (2021).
[47] S. Uchida, Effect of private information on indirect reciprocity, Phys. Rev. E **82**, 036111 (2010).
[48] K. Sigmund, Moral assessment in indirect reciprocity, J. Theor. Biol. **299**, 25 (2012).
[49] S. Uchida and T. Sasaki, Effect of assessment error and private information on stern-judging in indirect reciprocity, Chaos, Solitons Fractals **56**, 175 (2013).
[50] K. Oishi, T. Shimada, and N. Ito, Group formation through indirect reciprocity, Phys. Rev. E **87**, 030801(R) (2013).
[51] C. Hilbe, L. Schmid, J. Tkadlec, K. Chatterjee, and M. A. Nowak, Indirect reciprocity with private, noisy, and incomplete information, Proc. Natl. Acad. Sci. USA **115**, 12241 (2018).
[52] Y. Fujimoto and H. Ohtsuki, Reputation structure in indirect reciprocity under noisy and private assessment, Sci. Rep. **12**, 10500 (2022).
[53] Y. Fujimoto and H. Ohtsuki, Evolutionary stability of cooperation in indirect reciprocity under noisy and private assessment, Proc. Natl. Acad. Sci. USA **120**, e2300544120 (2023).
[54] A. Traulsen, M. A. Nowak, and J. M. Pacheco, Stochastic dynamics of invasion and fixation, Phys. Rev. E **74**, 011909 (2006).
[55] E. Brush, Å. Brännström, and U. Dieckmann, Indirect reciprocity with negative assortment and limited information can promote cooperation, J. Theor. Biol. **443**, 56 (2018).
[56] R. M. Whitaker, G. B. Colombo, and D. G. Rand, Indirect reciprocity and the evolution of prejudicial groups, Sci. Rep. **8**, 13247 (2018).
[57] A. L. Radzvilavicius, A. J. Stewart, and J. B. Plotkin, Evolution of empathetic moral evaluation, eLife **8**, e44269 (2019).
[58] M. Krellner and T. A. Han, Putting oneself in everybody's shoes-pleasing enables indirect reciprocity under private assessments, in *Artificial Life Conference Proceedings 32* (MIT Press, 2020), pp. 402–410.
[59] J. Quan, X. Yang, X. Wang, J.-B. Yang, K. Wu, and Z. Dai, Withhold-judgment and punishment promote cooperation in indirect reciprocity under incomplete information, Europhys. Lett. **128**, 28001 (2020).
[60] M. Krellner and T. A. Han, Pleasing enhances indirect reciprocity-based cooperation under private assessment, Artif. Life **27**, 246 (2021).
[61] L. Schmid, P. Shati, C. Hilbe, and K. Chatterjee, The evolution of indirect reciprocity under action and assessment generosity, Sci. Rep. **11**, 17443 (2021).
[62] J. Quan, J. Nie, W. Chen, and X. Wang, Keeping or reversing social norms promote cooperation by enhancing indirect reciprocity, Chaos, Solitons Fractals **158**, 111986 (2022).
[63] P. Gu and Y. Zhang, Reputation-based rewiring promotes cooperation in complex network, in *Advances in Guidance, Navigation and Control* (Springer, Berlin, Heidelberg, Germany, 2022), pp. 1405–1415.
[64] T. A. Kessinger, C. E. Tarnita, and J. B. Plotkin, Evolution of norms for judging social behavior, Proc. Natl. Acad. Sci. USA **120**, e2219480120 (2023).
[65] L. Schmid, F. Ekbatani, C. Hilbe, and K. Chatterjee, Quantitative assessment can stabilize indirect reciprocity under imperfect information, Nat. Commun. **14**, 2086 (2023).
[66] S. Tanabe, H. Suzuki, and N. Masuda, Indirect reciprocity with trinary reputations, J. Theor. Biol. **317**, 338 (2013).
[67] Y. Murase, M. Kim, and S. K. Baek, Social norms in indirect reciprocity with ternary reputations, Sci. Rep. **12**, 455 (2022).
[68] See https://github.com/fyuuma2005/reputation_third_order.