PNAS RESEARCH ARTICLE | EVOLUTION SOCIAL SCIENCES

Evolutionary stability of cooperation in indirect reciprocity under noisy and private assessment

Yuma Fujimoto[a,b,c,1] and Hisashi Ohtsuki[a,d]

Edited by Alan Hastings, University of California, Davis, CA; received January 25, 2023; accepted April 9, 2023

Indirect reciprocity is a mechanism that explains large-scale cooperation in humans. In indirect reciprocity, individuals use reputations to choose whether or not to cooperate with a partner and update others' reputations. A major question is how the rules to choose their actions and the rules to update reputations evolve. In the public reputation case where all individuals share the evaluation of others, social norms called Simple Standing (SS) and Stern Judging (SJ) have been known to maintain cooperation. However, in the case of private assessment where individuals independently evaluate others, the mechanism of maintenance of cooperation is still largely unknown. This study theoretically shows for the first time that cooperation by indirect reciprocity can be evolutionarily stable under private assessment. Specifically, we find that SS can be stable, but SJ can never be. This is intuitive because SS can correct interpersonal discrepancies in reputations through its simplicity. On the other hand, SJ is too complicated to avoid an accumulation of errors, which leads to the collapse of cooperation. We conclude that moderate simplicity is a key to stable cooperation under the private assessment. Our result provides a theoretical basis for the evolution of human cooperation.

cooperation | indirect reciprocity | private assessment | simple standing

Significance

How large-scale cooperation, observed mainly in human societies, is maintained has long been of wide interest and studied by indirect reciprocity models, where individuals choose their behavior based on the reputation of the partner and judge observed actions of others as good or bad. Recent studies show, however, that, when individuals cannot share their evaluations of others (called private assessment), mismatches in their opinions are amplified and eventually collapse their proper judgments of good and bad individuals. This study theoretically shows that cooperation is stable even under such noisy and private assessment. We find that a key to success is moderate simplicity in the rule of reputation assignment.

Cooperation benefits others but is costly to the cooperator itself. Nevertheless, cooperation is widespread from microscopic to macroscopic scales, such as among microorganisms, animals, humans, and nations. One way to sustain cooperation is that agents conditionally cooperate with others who cooperate with them, which is realized by, for example, repeated interactions (1–3) and partner choice (4–7). Such conditional cooperation based on personal experiences is applicable only to a small population where members can interact directly and repeatedly with most of the others.

However, cooperative behavior is observed even in a large-scale society (e.g., human societies). Since individuals inevitably encounter strangers there, they need the reputations of those strangers in order not to cooperate unconditionally. The mechanism where individuals indirectly reward others via their reputations as described above is called indirect reciprocity (8–10). In reality, humans are particularly interested in reputations and gossip about themselves and others (11–13). Furthermore, many experiments have pointed out that gossips concern cooperative behaviors (14–16).

Errors that inevitably occur in actions and in assessments hinder cooperation by indirect reciprocity. Indeed, the simplest social norm called image scoring (9, 10) fails to maintain full cooperation under errors (17, 18) [a similar failure is also seen in direct reciprocity (18–20)]. This is because one erroneous defection triggers further defection. Nevertheless, previous studies have theoretically shown that cooperation can be maintained by the so-called "leading eight" social norms (21, 22) even in the presence of such errors when all individuals share the reputation of the same individual (i.e., public reputation). Public reputation cases have been thoroughly studied for about two decades (23–37). When individuals cannot share their evaluations of the same target (i.e., private assessment), however, errors cast a shadow over cooperation more crucially. In this case, a single disagreement in opinions between two individuals can lead to further disagreements (38–42). Whether cooperation is maintained under such noisy and private assessment is still largely unsolved in theory and is one of the major open problems in studies of indirect reciprocity (36, 43, 44).

Previous studies have shown that maintaining cooperation with indirect reciprocity is very difficult under noisy and private assessment. For example, Hilbe et al. (42) showed by an evolutionary simulation that the above leading eight strategies cannot succeed in cooperation under private assessment. Some studies (45–47) have demonstrated the emergence of cooperation under noisy and private assessment, but under the restrictive

Author affiliations: [a]Research Center for Integrative Evolutionary Science, SOKENDAI (The Graduate University for Advanced Studies), Hayama 240-0193, Japan; [b]Universal Biology Institute, The University of Tokyo, Bunkyo-ku 113-0033, Japan; [c]CyberAgent, Inc., Shibuya-ku 150-0042, Japan; and [d]Department of Evolutionary Studies of Biosystems, SOKENDAI, Hayama 240-0193, Japan

Author contributions: Y.F. and H.O. designed research; performed research; and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2023 the Author(s). Published by PNAS. This article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]To whom correspondence may be addressed. Email: fujimoto_yuma@soken.ac.jp.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2300544120/-/DCSupplemental.

Published May 8, 2023.

assumption that only local mutations in the strategy space are allowed, thus excluding the possibility that a fully cooperative strategy is directly invaded by free riders. Other studies have shown that a mechanism to synchronize opinions between individuals has a positive influence on cooperation in indirect reciprocity, such as empathy, generosity, spatial structure, and so on (48–57).

Most of these studies of private assessment have been performed by computer simulations (42, 45). This is because two-dimensional information of who assigns a reputation to whom [its matrix representation is called "image matrix" (38, 39, 58, 59)] becomes too complex to analyze. For example, its possible transition is illustrated in Fig. 1*A*, where a single assessment error can be amplified with time, leading to a mosaic structure in the image matrix. Previously, the authors have developed an analytical approach (60) to study the image matrix, but it was applicable only when the population is monomorphic in strategies. However, for an evolutionary analysis between wild type and mutant, we need a different method because the image matrix now includes four compartments based on different rules of reputation assignment adopted by wild-type and mutant individuals (Fig. 1*B*). In spite of these difficulties, here we report that we have successfully developed a new analytical machinery to study the image matrix, which has been obtained by qualitatively extending the previous approach. This enables us to make a general prediction of when cooperation is sustained under noisy and private assessment over the full parameter region.

In the following, we will first introduce the setting of indirect reciprocity under noisy and private assessment and explain a method to analytically calculate an expected payoff of each individual through analyzing an image matrix. Then, we will discuss which strategy can be an evolutionarily stable strategy (ESS) (61, 62) under which condition and provide intuitive reasons for the result. To our knowledge, this is the first systematic study that has analytically investigated evolutionary stability of strategies in indirect reciprocity under noisy and private assessment.

Model

We consider a model of indirect reciprocity in a well-mixed population of size N. We assume that, in every step, a binary reputation is assigned independently from everyone to everyone, either good or bad, which is summarized by image matrix $\{\beta_{ji}\}$, where $\beta_{ji} = 1$ (resp. $\beta_{ji} = 0$) if individual i assigns a good (resp. bad) reputation to individual j. The model proceeds as follows. First, a donor and a recipient are randomly chosen from this population. Next, the donor takes its action, cooperation, or defection to the recipient. When the donor cooperates, the donor incurs a cost $c(> 0)$ but gives a benefit $b(> c)$ to the recipient instead. On the other hand, when the donor defects, no change occurs in the payoff of the donor or the recipient. Here, a rule that specifies how the donor chooses its action is called "action rule". Throughout this paper, we assume that all the individuals adopt the "discriminator" action rule (9, 63), with which they choose cooperation (resp. defection) to a good (resp. bad) recipient in their own eyes; that is, donor i chooses cooperation toward recipient j if $\beta_{ji} = 1$ and chooses defection if $\beta_{ji} = 0$. We assume that the donor unintentionally takes the opposite action to the intended one with probability $0 \leq e_1 < 1/2$ (action error). All the individuals in the population observe this social interaction between the donor and the recipient and independently update the reputation of the donor in their eyes. We assume that these processes of action choice and reputation update continue sufficiently long time (i.e., continuation probability is 1).

A rule that specifies how each observer updates reputations is called its "social norm". In models of public reputation, it has often been assumed that all the individuals in the population adopt the same social norm (21, 29, 64) (but ref. 38), otherwise they cannot share the reputation of the same individual. Because

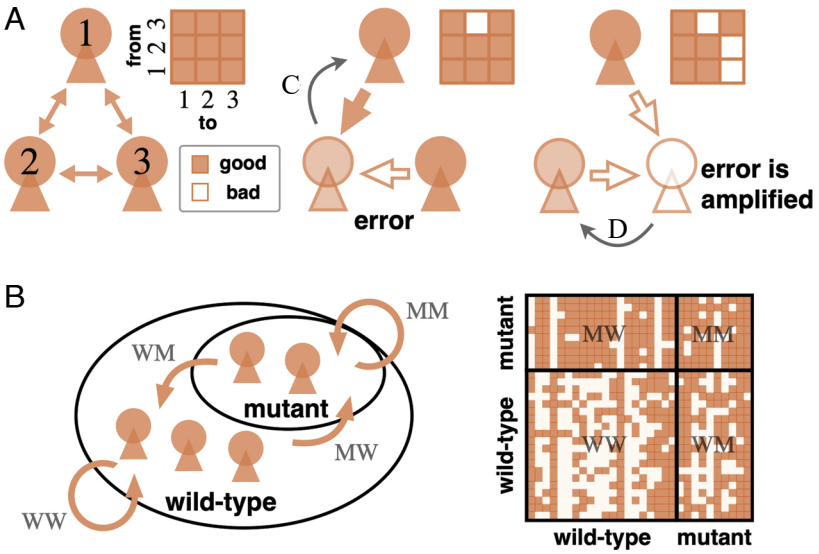

Fig. 1. (*A*) An illustration showing how an assessment error is amplified. In all the three panels, there are Persons 1, 2, and 3, and the 3 × 3 image matrices and straight arrows indicate the reputations among them. In the left panel, good reputations are assigned among all of them; hence, they achieve cooperation. In the center, Person 2 cooperates with Person 1, but Person 3 erroneously assigns a bad reputation to Person 2. In the right, Person 3 defects with Person 2 based on its bad reputation in the eyes of Person 3, but because Persons 1 and 2 believe that Person 2 is good, they assign bad reputations to Person 3. (*B*) An illustration showing the complexity of the image matrix. The *Left* panel shows that wild types and mutants are mixed in the population and that four kinds of reputations exist; WW from wild type to wild type, WM from mutant to wild type, MW from wild type to mutant, and MM from mutant to mutant. The *Right* panel shows that the image matrix is decomposed into the corresponding four components, each of which has a different reputation structure.

we consider a model of private reputation here, however, we instead assume that individuals can adopt different social norms. This study deals with a situation where each observer (say, k) refers to (i) whether the donor (say, i) cooperates (C) or defects (D) (first-order information) and (ii) whether the recipient (say, j) is good (G) or bad (B) in the eyes of the observer (second-order information, represented by β_{jk}) when this observer updates the reputation of the donor in the eyes of the observer, denoted by β_{ik}. Such social norms are called "second-order" social norms (10, 18, 33, 36). An observer who adopts a second-order social norm can face four different cases, denoted by GC("toward a Good recipient the donor Cooperates"), BC("toward a Bad recipient the donor Cooperates"), GD("toward a Good recipient the donor Defects"), and BD("toward a Bad recipient the donor Defects"), respectively, and in each case, the observer assigns either a good (G) or bad (B) reputation to the donor. Thus, a social norm is represented by a four-letter string. For example, GBBG is the social norm that assigns to the donor a good reputation in GC- and BD-cases and a bad reputation in BC- and GD-cases. There are $2^4 = 16$ such social norms in total, and we lexicographically order them with the rule that G comes first and B comes second and number them from S_{01} to S_{16}. Table 1 shows a full list of 16 social norms studied here. When updating the reputation, each observer independently commits an assessment error with probability $0 < e_2 < 1/2$, in which case he/she accidentally assigns the opposite reputation to the intended one to the donor.

Several norms are especially important in previous studies, so we explain them below. We call S_{01} ALLG and call S_{16} ALLB because these norms unconditionally assign good or bad reputations. Next, S_{03}, S_{04}, S_{07}, and S_{08} belong to G∗B∗ family. These norms share the same feature that they regard cooperation toward a good recipient as good and defection toward a good recipient as bad. They only differ when the recipient is bad in the observer's eyes. First, S_{04} is called Scoring (SC), which regards cooperation toward a bad recipient as good and defection toward a bad recipient as bad, and therefore, reputation assignment is independent of whether the recipient is good or bad in the observer's eyes (thus, categorized as a first-order norm). Next, S_{07} is called Stern Judging (SJ), which regards cooperation toward a bad recipient as bad and defection toward a bad recipient as good, as opposed to SC. Third, S_{03} is called Simple Standing (SS) and it regards any action toward a bad recipient as good, and therefore, it is the most generous norm in this family. Finally, S_{08} is called Shunning (SH) and it regards any action toward a bad recipient as bad, and therefore, it is the most intolerant one. Notably, SJ and SS are the two second-order norms that are included in the "leading eight" norms (21), which can successfully maintain cooperation under noisy and public assessment within third-order norms. In particular, SJ has long been considered promising because it is evolutionarily successful (23) and because it sustains a very high level of cooperation despite its simplicity (33, 36). SJ always suggests only one correct action to keep you good; it recommends cooperation toward good individuals and defection toward bad ones, and failure to follow this rule leads to a bad reputation. Under the noisy public reputation, SH cannot achieve full cooperation against itself but can prevent the invasion of ALLB (*SI Appendix* for detailed calculations).

Table 1. All 16 second-order social norms in this study

	Social norm	GC	BC	GD	BD
S_{01}	(ALLG)	G	G	G	G
S_{02}		G	G	G	B
S_{03}	(SS; Simple Standing)	G	G	B	G
S_{04}	(SC; Scoring)	G	G	B	B
S_{05}		G	B	G	G
S_{06}		G	B	G	B
S_{07}	(SJ; Stern Judging)	G	B	B	G
S_{08}	(SH; Shunning)	G	B	B	B
S_{09}		B	G	G	G
S_{10}		B	G	G	B
S_{11}		B	G	B	G
S_{12}		B	G	B	B
S_{13}		B	B	G	G
S_{14}		B	B	G	B
S_{15}		B	B	B	G
S_{16}	(ALLB)	B	B	B	B

Under these settings, the strategy of an individual is its social norm. For this reason, we use "strategy" and "(social) norm" interchangeably in the following. We ask which strategy is evolutionarily stable. To this end, we study invasibility of a mutant strategy against a wild-type one. A strategy is ESS if it is not invaded by any other 15 mutant strategies. To derive their payoffs, we need to analyze the image matrix, which we shall perform below.

Analysis of Reputation Structure

Let us consider a situation where individuals with mutant norm M invade the population of wild-type norm W ($\neq$ M). Here, the proportion of mutants is given by δ. By extending the methodology in Fujimoto and Ohtsuki (60), who studied private reputation structure in a monomorphic population, here, we have developed a new framework to analyze the structure of private reputations in a dimorphic population where wild types and mutants coexist (*SI Appendix* for detailed calculations). Specifically, take a focal individual whose norm is $A \in \{\mathrm{W}, \mathrm{M}\}$, and let $p_{AA'}$ (hereafter called "goodness") be the proportion of individuals among norm A' users who assign a good reputation to the focal individual, for $A' \in \{\mathrm{W}, \mathrm{M}\}$. Thus, a wild-type individual is characterized by a pair of goodnesses, $(p_{\mathrm{WW}}, p_{\mathrm{WM}})$, and we represent its joint probability distribution over all wild-type individuals by $\Phi_{\mathrm{W}}(p_{\mathrm{WW}}, p_{\mathrm{WM}})$. In the same way, a mutant is characterized by a pair of goodnesses, $(p_{\mathrm{MW}}, p_{\mathrm{MM}})$, and $\Phi_{\mathrm{M}}(p_{\mathrm{MW}}, p_{\mathrm{MM}})$ represents its distribution over all mutants. In *SI Appendix*, we derive the dynamics of Φ_{W} and Φ_{M} by formulating a stochastic transition of the donor's goodnesses under the assumption of $N \gg 1$ (the population is large), $\delta \ll 1$ (mutants are rare), and $N\delta \gg 1$ (yet, the number of mutants is sufficiently large). Since we assume that the continuation probability is 1, we derive their equilibrium distributions, Φ^*_{W} and Φ^*_{M}, in order to calculate expected payoffs of wild types and mutants, which enable us to study the invasibility condition of mutants to wild types (*SI Appendix* again).

We find that each of the two equilibrium distributions is well approximated by a weighted sum of two-dimensional Gaussian functions with zero covariance, where each Gaussian can be systematically labeled by a nonzero integer, $j \in \mathbb{Z}\backslash\{0\}$ (see an example in Fig. 2*B* and the rule of labeling in Fig. 2*C*). Hence, the number of Gaussians that appear in the sum is infinitely but countably many, but in some cases, these labels degenerate (i.e., two or more Gaussians are identical but they are given different labels) and it can be finite. Weights to Gaussians decay exponentially as j becomes large positive or large negative, so a

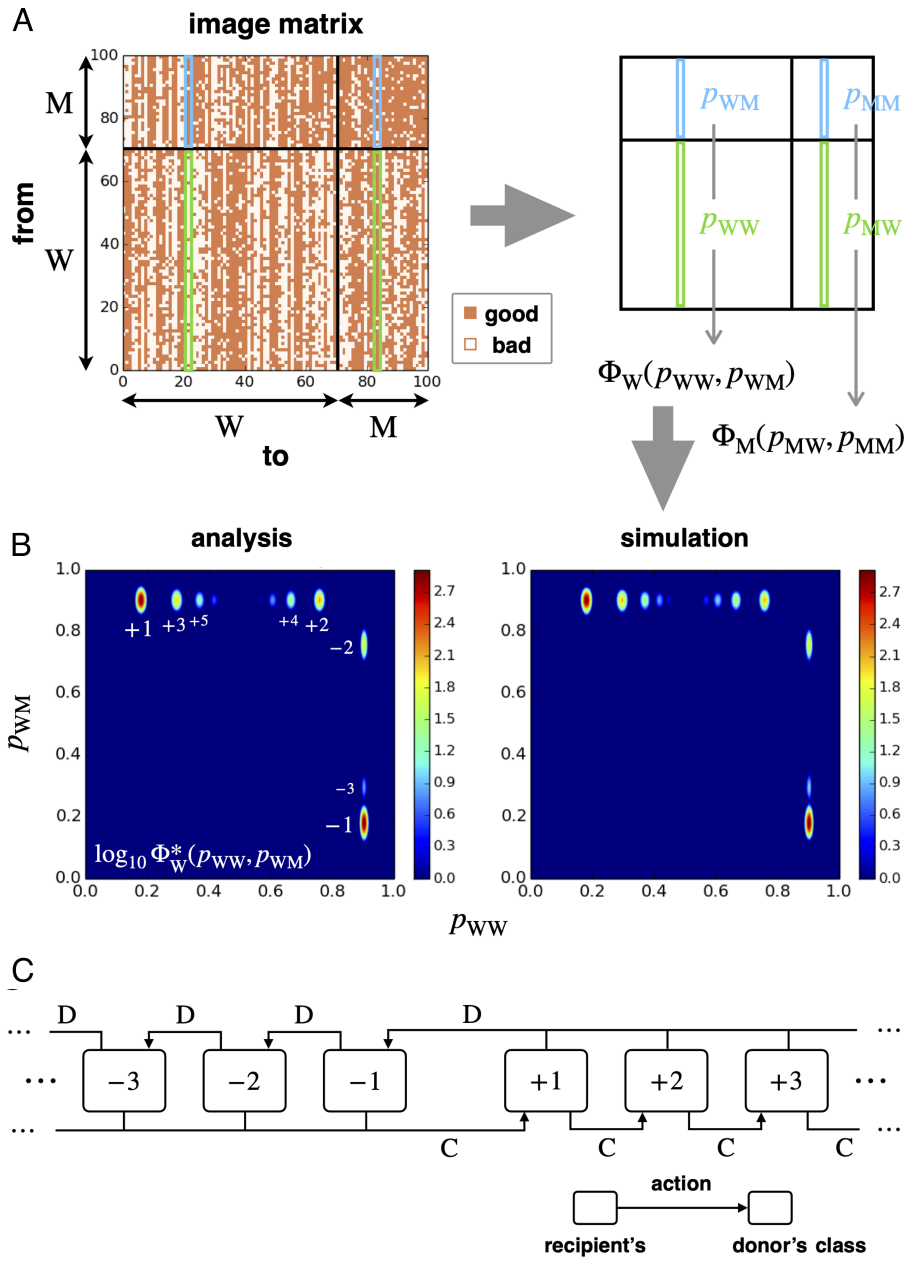

Fig. 2. Illustrations of our method to analyze reputation structure. (*A*) In the *Left* panel, an example of an image matrix is shown. We analyze this image matrix divided into four parts; the reputations from wild types (W) or mutants (M) to W/M. In the *Right* panel, a pair of goodnesses of each individual from wild types (colored green) and mutants (blue) are extracted from the image matrix. Because the pair of goodnesses correlate with each other, we consider the joint probability distribution of them, denoted by Φ_W and Φ_M. (*B*) We analytically calculated this joint probability distribution. One can see that the analytical estimates (the *Left* panel) fit the simulated one (*Right*) well. In both the panels, we assume $(W, M) = (S_{09}, S_{03})$, $N = 5{,}000$, $\delta = 0.1$, and $(e_1, e_2) = (0, 0.1)$. In the numerical simulation, we used 3,000 samples of image matrices from time $t = 51, \cdots, 3{,}050$ (a random donor's goodness is updated N times per unit time of t). On the other hand, in the theoretical analysis, we introduced the cutoff of $-100 \le j \le +100$. Each number near the heat peaks indicates the class label j. (*C*) Rules for labeling individual classes. Each class corresponds to one Gaussian distribution. Each box (labeled by $j \in \mathbb{Z}\backslash\{0\}$) indicates a class. The destination of each arrow indicates the class that the donor moves to after taking cooperation (C) or defection (D) toward the recipient that belongs to the class that the arrow originates. For example, a donor that cooperated with a class $j = -2$ recipient moves to class $j = +1$.

truncation at some finite number of terms approximates well the infinite sum for numerical calculations.

ESS Norms

Based on the analysis of the image matrix above, we have studied pairwise invasibility for all the pairs of wild-type W and mutant M. In the following, we set the action error rate as $e_1 = 0$ because this error, especially when it is small positive, does not have a qualitative impact on our results as far as we studied. Thus, the cost-benefit ratio b/c and the assessment error rate e_2 are our environmental parameters.

We first find that the four strategies, S_{06}, S_{07}(SJ), S_{10}, and S_{11}, are completely indistinguishable, both as wild type and as mutant. This is because these norms always give the goodness of $1/2$ to anyone in the population at equilibrium due to an accumulation of assessment errors and hence they appear to choose cooperation and defection in a random manner (this has been known for S_{07}(SJ); refs. 38, 39, and 42). In particular, they are neutral to each other. For these reasons, we will discuss only S_{07}(SJ)

as a representative of them and exclude the other three in the following analysis.

Our exhaustive analysis demonstrates that only three norms, S_{03}(SS), S_{08}(SH), and S_{16}(ALLB), can be ESS, and all the others cannot. As shown in Fig. 3*A*, ALLB is ESS independent of b/c and e_2 because it is the norm that assigns a bad reputation to everyone, saves the own cost, and provides no benefit to others. On the other hand, SS and SH achieve ESS for some b/c and

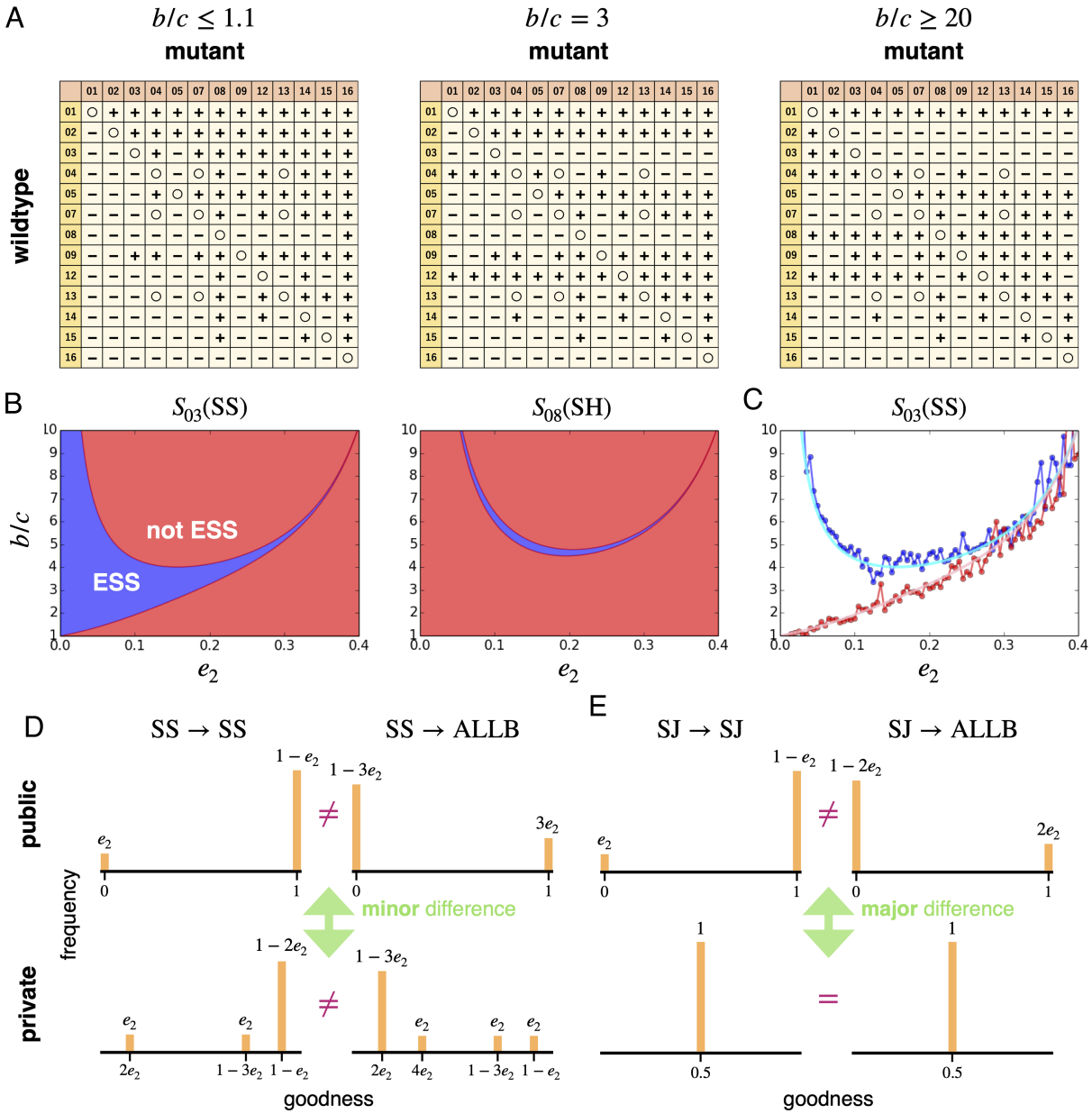

Fig. 3. Details of ESS analysis. (*A*) Invasibility between all pairs of the social norms. The *Left*, *Center*, and *Right* panels, respectively, show the cases of $1 < b/c \leq 1.1$, $b/c = 3$, and $b/c \geq 20$, for demonstrating the invasibility when b/c is close to 1 (*Left*), when S_{03}(SS) is ESS (*Center*), and when b/c is very large (*Right*). Numbers in rows (resp. columns) indicate the labels of wild-type (resp. mutant) norms. Each plus (resp. minus) mark indicates that the invasion by mutants is successful (resp. unsuccessful). Each circle mark indicates that the wild-type and mutant norms are neutral. All the panels are based on $e_2 = 0.1$ and $-10{,}000 \leq j \leq +10{,}000$. As for invasibility for the other b/c values, *SI Appendix*, Fig. S2. (*B*) The ESS parameter region for norms S_{03}(SS) (*Left*) and S_{08}(SH) (*Right*). In each panel, the horizontal (resp. vertical) axis indicates e_2 (resp. b/c). The blue (resp. red) color indicates that the norm is ESS (resp. not ESS). (*C*) Comparison between analytical and numerical calculations of the ESS region of S_{03}(SS). The horizontal and vertical axis are the same as in (*B*) The cyan (resp. pink) line indicates the theoretical upper (resp. lower) bound the same as the *Left* panel in (*B*). Blue (resp. red) dots, connected by lines, indicate the numerical estimates of the upper (resp. *Lower*) bound. Those estimates were calculated by individual-based simulations of the image matrix with $N = 10{,}000$, $\delta = 0.03$. The average of 50 samples from generations $51 \leq t \leq 100$ were used, except for in the calculation of the upper bound (blue dots) for $e_2 < 0.1$ where we instead used the average of 3,000 samples from $51 \leq t \leq 3{,}050$ to reduce errors in estimation. (*D*) A comparison between public (*Top* row) and private (*Bottom* row) assessment cases of how wild-type SS individuals evaluate other SS individuals (*Left* column) and how wild-type SS individuals evaluate mutant ALLB individuals (*Right* column). In each panel, the horizontal and vertical axes indicate individual goodness and its frequency, respectively. Positions and heights of bars are correct only up to order e_2. We see that SS gives high goodness to most of the SS individuals (*Left* column), that SS gives low goodness to most of the ALLB individuals (*Right* column), and that the difference between the *Top* and *Bottom* rows is minor (in a scale of $O(e_2)$). Thus, SS is robust against the invasion by ALLB under both public and private assessments. (*E*) A similar comparison to *D* was made for wild-type SJ and mutant ALLB. We see that SJ gives high goodness to most of the SJ individuals under public assessment (*Top Left*), that SJ gives low goodness to most of the ALLB individuals (*Top Right*), but that SJ gives the goodness of 1/2 to both SJ (*Bottom Left*) and ALLB (*Bottom Right*) individuals under private assessment. Thus, SJ is robust against the invasion by ALLB under public assessment while it is not under private assessment.

e_2; there are upper and lower bounds of b/c for them to be ESS, which depend on e_2. Below we will look at its details.

Conditions for ESS

The ESS condition of S_{03}(SS) is shown in Fig. 3*B*. When b/c exceeds the upper bound, the norm is invaded by S_{01}(ALLG) (compare the *Right* and *Center* panels of Fig. 3*A*). On the other hand, when b/c falls below the lower bound, the norm is invaded by S_{04}(SC) (compare the *Left* and *Center* panels of Fig. 3*A*). Fig. 3*C* shows that these theoretical bounds are also supported by individual-based simulations. Notably, the smaller e_2 is, the wider the ESS region of S_{03}(SS) becomes.

The ESS region of S_{08}(SH) is quite narrow in comparison to that of S_{03}(SS), as seen in Fig. 3*B*. In addition, when b/c exceeds the upper bound or falls below the lower bound, the norm is invaded by S_{04}(SC) and S_{16}(ALLB), respectively. The range of b/c-ratios that make SH evolutionarily stable is widest at an intermediate e_2 (about 0.1).

In contrast to these results, we find that S_{07}(SJ), which is known to be a successful norm when reputation is public, is invaded by norms such as S_{16}(ALLB) and S_{08}(SH) independent of the value of b/c (and also independent of e_2) and therefore that it is never an ESS. This is summarized in Fig. 3*A*.

To summarize, SS, SH, and SJ are all the ESS norms under the public reputation, but whether they remain ESS under the private assessment critically differs. This difference is clearly understood by focusing on how the reputation structure they give differs between the public and private reputation cases under a sufficiently small but positive assessment error rate, $e_2 \ll 1$. Let us consider below, for example, whether each norm can prevent the invasion of ALLB, a potential invader norm.

Success of Simple Standing. The reputation structure that S_{03}(SS) gives differs little between the public and private reputation cases (Fig. 3*D*). Under the public reputation (*SI Appendix* for the calculation), SS assigns good reputations to SS themselves (represented by the bar at goodness = 1 in the *Top-Left* panel in Fig. 3*D*), while bad reputations to ALLB (represented by the bar at goodness = 0 in the *Top-Right* panel in Fig. 3*D*). Thus, SS distinguishes between SS itself and the invader ALLB and prevents the invasion of ALLB. Even under the private reputation, SS still assigns good reputations to SS themselves (*Bottom-Left* panel in Fig. 3; high goodness of $1-e_2$ are given to the fraction $1-2e_2$ of SS individuals, for example) and assigns bad reputations to ALLB (*Bottom-Right* panel in Fig. 3; low goodness of $2e_2$ are given to the fraction $1-3e_2$ of ALLB individuals, for example). Thus, the distinction between SS and ALLB is maintained. For that reason, SS succeeds in achieving ESS even under the private assessment. The cooperation rate at this ESS is as high as $1-2e_2$ for small e_2, so it entails nearly perfect cooperation.

Failure of Stern Judging. Contrary to SS, the reputation structure that S_{07}(SJ) gives extremely differs between the public and private reputation cases (Fig. 3*E*). Under the public reputation (*SI Appendix* for the calculation), wild-type SJ gives high goodness to other SJ (*Top-Left* in Fig. 3*E*) and wild-type SJ gives low goodness to ALLB (*Top-Right* in Fig. 3*E*). Thus, SJ prevents the invasion of ALLB. Under the private assessment, however, SJ gives the goodness of 1/2 to other SJ individuals (*Bottom-Left* in Fig. 3*E*) (38, 39, 42), while SJ gives the goodness of 1/2 to ALLB individuals as well (*Bottom-Right* in Fig. 3*E*). Thus, the distinction between SJ and ALLB is lost. This is why SJ fails to be ESS under the private assessment.

Shunning Can Be ESS, but the Level of Cooperation Is Low. We can understand why S_{08}(SH) achieves ESS only in a narrow region under private reputation (*SI Appendix* for the detailed calculation and *SI Appendix*, Fig. S3 for the illustration for easy interpretation). Under the public reputation, SH gives good reputations to half of other SH wild types and bad reputations to the other half (*Top-Left* in *SI Appendix*, Fig. S3) while SH gives bad reputations to almost all ALLB (*Top-Right* in *SI Appendix*, Fig. S3). Thus, SH prevents the invasion of ALLB. Under private reputation, on the other hand, SH gives low goodness to both SH and ALLB (*Bottom-Left* and *Bottom-Right* in *SI Appendix*, Fig. S3). Here, however, SH has a slightly better chance to receive good reputations than ALLB, in the order of e_2^2. This explains why SH prevents the invasion from ALLB only in a narrow region and also explains why its ESS condition becomes more strict for a smaller assessment error rate, e_2. The cooperation rate at a realized ESS is as low as e_2 for small e_2, so we conclude that S_{08}(SH) does not contribute to cooperation.

Discussion

This study considered indirect reciprocity under noisy and private assessment. We focused on the goodness of an individual (i.e., what proportion of individuals assigns the focal individual a good reputation) between different norms and developed an analytical method to calculate the distribution of goodness at equilibrium. Using this methodology, we studied whether a mutant norm succeeds in the invasion into a wild-type norm. As far as we know, this is the first analytical study that exhaustively investigated the evolutionary stability of all possible second-order social norms under noisy and private assessment. Although both S_{03}(SS) and S_{07}(SJ) can be ESS under public reputation, we found that their evolutionary stability is totally different under private assessment. In particular, we found that S_{03}(SS) remains to be ESS under private assessment if the assessment error rate is small, while S_{07}(SJ) cannot be ESS no matter how small the error rate is (38, 39, 42).

The reason for this difference between S_{03}(SS) and S_{07}(SJ) comes from the difference in the complexity of these two norms. In the world of private assessment, errors in assessment accumulate independently among observers, which is a potential source of collapse of cooperation in the population. However, since S_{03}(SS) regards a cooperating donor as good no matter whether the recipient is good or bad, discrepancy in the opinion toward the recipient between two different observers does not produce further discrepancy; those two observers can agree that such a cooperating donor is good. In contrast, S_{07}(SJ) is more complex than S_{03}(SS) and the recipient's reputation is always decisive information (Table 1), so this complexity becomes an obstacle for correcting discrepancy between observers.

Hilbe et al. (42) studied by computer simulations whether the leading eight norms can sustain cooperation under the noisy and private assessment. They concluded that S_{03}(SS) (referred to as "L3" in their paper) and S_{07}(SJ) ("L6") fail to achieve cooperation, which is contrary to our result. This difference is partly because we studied evolutionary stability in a deterministic model, while they studied fixation probability in a stochastic model. Because those two criteria are different, drawing a general conclusion is difficult. For example, our ESS analysis cannot fully answer how robust each ESS is compared to others, although we can calculate the payoff difference between a focal

ESS (as wild type) and a mutant strategy, which indicates its robustness to some extent. For a full understanding of the system, it would be ideal to extend our current framework so that we can study a wild-type/mutant system where mutants are not necessarily rare. Such an extension will enable us to study an invasion barrier (65), that is the maximal frequency of mutants that wild types can resist, and to discuss the robustness of each ESS in a higher resolution. We leave it as a future study.

Our current study assumed that everyone adopts the discriminator action rule. This approach was taken in a previous study (45) to make the analysis tractable. One can conceive, however, three other possible action rules such as "ALLC" that always cooperates, "ALLD" that always defects, and "paradoxical discriminator" (30, 64) that cooperates with a bad individual and defects with a good one. Nevertheless, our current approach retains fairly good generality. First, studying a paradoxical discriminator with a norm S_i is mathematically equivalent to studying a discriminator with a norm S_{17-i} because these two norms are symmetric to each other with respect to G and B reputations (Table 1). Second, ALLC essentially corresponds to ALLG because ALLG always intends to assign a good reputation to others and cooperate with them, and ALLD essentially corresponds to ALLB because ALLB always intends to assign a bad reputation and defect with them. Strictly speaking, a difference arises between ALLC and ALLG and between ALLD and ALLB when there is an assessment error, but we believe that our choice of strategy space is reasonable, especially when this error rate is small.

A future direction of this study would be to examine ESS conditions of social norms when some of the assumptions are changed. For example, we have assumed second-order norms, in which individuals refer to a donor's action (first-order information) and a recipient's reputation (second-order one) when they update the donor's reputation. However, humans may use more complex norms than second-order ones. Studying the effect of higher-order information (32, 33, 36, 66), such as the previous reputation of the donor (third-order information), would further deepen our understanding. We have also assumed that all individuals simultaneously update their opinions toward the same donor. However, in a real society, the number of people who can observe a single person's behavior is limited. Thus, the effect of asynchronous updates of reputations is worth studying. Last but not least, we have assumed that game interactions last sufficiently long so that we can use equilibrium distributions of goodness for calculating payoffs. This assumption was helpful for our analysis because we can neglect any effects of the initial condition. We admit that it was the most favorable condition for cooperation. Studying a model with continuation probability being less than 1, as in refs. 9, 17, 18, and 29, is important because such a model is more realistic.

In conclusion, we have demonstrated that cooperation can be evolutionarily stable even under noisy and private assessment. Specifically, we have shown that Stern Judging, which is one of the leading norms under public reputation, cannot distinguish between cooperators and defectors under private assessment and thus fails to achieve ESS. On the other hand, we have revealed that Simple Standing can be stable in a wide range of parameters. Based on these results, we predict that Simple Standing should play a key role in sustaining cooperation by indirect reciprocity under noisy private assessment. These findings provide a rigid theoretical basis for understanding human cooperation and pave the way for future studies in biology, psychology, sociology, and economics.

Data, Materials, and Software Availability. There are no data underlying this work.

ACKNOWLEDGMENTS. Y.F. acknowledges the support by JSPS KAKENHI Grant No. JP21J01393. H.O. acknowledges the support by JSPS KAKENHI Grant No. JP19H04431.

1. R. L. Trivers, The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57 (1971).
2. R. Axelrod, W. D. Hamilton, The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
3. R. Axelrod, *The Evolution of Cooperation* (Basic, New York, 1984).
4. T. Yamagishi, N. Hayashi, N. Jin, "Prisoner's dilemma networks: Selection strategy versus action strategy" in *Social Dilemmas and Cooperation* (Springer, 1984), pp. 233–250.
5. R. Noë, P. Hammerstein, Biological markets: Supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behav. Ecol. Sociobiol.* **35**, 1–11 (1994).
6. R. Noë, P. Hammerstein, Biological markets. *Trends Ecol. Evol.* **10**, 336–339 (1995).
7. P. Barclay, Strategies for cooperation in biological markets, especially for humans. *Evol. Hum. Behav.* **34**, 164–175 (2013).
8. R. D. Alexander, *The Biology of Moral Systems* (Aldine de Gruyter, New York, 1987).
9. M. A. Nowak, K. Sigmund, Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577 (1998).
10. M. A. Nowak, K. Sigmund, Evolution of indirect reciprocity. *Nature* **437**, 1291–1298 (2005).
11. N. Emler, *Gossip, Reputation, and Social Adaptation* (University Press of Kansas, 1994).
12. R. I. M. Dunbar, *Grooming, Gossip, and the Evolution of Language* (Harvard University Press, 1998).
13. R. I. Dunbar, Gossip in evolutionary perspective. *Rev. General Psychol.* **8**, 100–110 (2004).
14. M. Feinberg, R. Willer, J. Stellar, D. Keltner, The virtues of gossip: Reputational information sharing as prosocial behavior. *J. Personality Soc. Psychol.* **102**, 1015 (2012).
15. M. Feinberg, R. Willer, M. Schultz, Gossip and ostracism promote cooperation in groups. *Psychol. Sci.* **25**, 656–664 (2014).
16. J. Wu, D. Balliet, P. A. Van Lange, Reputation, gossip, and human cooperation. *Soc. Personality Psychol. Compass* **10**, 350–364 (2016).
17. K. Panchanathan, R. Boyd, A tale of two defectors: The importance of standing for evolution of indirect reciprocity. *J. Theor. Biol.* **224**, 115–126 (2003).
18. K. Sigmund, *The Calculus of Selfishness* (Princeton University Press, 2010).
19. M. Nowak, K. Sigmund, A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature* **364**, 56–58 (1993).
20. J. Wu, R. Axelrod, How to cope with noise in the iterated prisoner's dilemma. *J. Conf. Resol.* **39**, 183–189 (1995).
21. H. Ohtsuki, Y. Iwasa, How should we define goodness? Reputation dynamics in indirect reciprocity*J. Theor. Biol.* **231**, 107–120 (2004).
22. H. Ohtsuki, Y. Iwasa, The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.* **239**, 435–444 (2006).
23. J. M. Pacheco, F. C. Santos, F. A. C. Chalub, Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity. *PLoS Comput. Biol.* **2**, e178 (2006).
24. S. Suzuki, E. Akiyama, Evolution of indirect reciprocity in groups of various sizes and comparison with direct reciprocity. *J. Theore. Biol.* **245**, 539–552 (2007).
25. F. C. Santos, F. A. Chalub, J. M. Pacheco, "A multi-level selection model for the emergence of social norms" in *European Conference on Artificial Life* (Springer, 2007), pp. 525–534.
26. F. Fu, C. Hauert, M. A. Nowak, L. Wang, Reputation-based partner choice promotes cooperation in social networks. *Phys. Rev. E* **78**, 026117 (2008).
27. S. Suzuki, E. Akiyama, Evolutionary stability of first-order-information indirect reciprocity in sizable groups. *Theore. Popul. Biol.* **73**, 426–436 (2008).
28. S. Uchida, K. Sigmund, The competition of assessment rules for indirect reciprocity. *J. Theor. Biol.* **263**, 13–19 (2010).
29. H. Ohtsuki, Y. Iwasa, M. A. Nowak, Reputation effects in public and private interactions. *PLoS Comput. Biol.* **11**, e1004527 (2015).
30. F. P. Santos, F. C. Santos, J. M. Pacheco, Social norms of cooperation in small-scale societies. *PLoS Comput. Biol.* **12**, e1004709 (2016).
31. F. P. Santos, J. M. Pacheco, F. C. Santos, Evolution of cooperation under indirect reciprocity and arbitrary exploration rates. *Sci. Rep.* **6**, 1–9 (2016).
32. T. Sasaki, I. Okada, Y. Nakai, The evolution of conditional moral assessment in indirect reciprocity. *Sci. Rep.* **7**, 1–8 (2017).
33. F. P. Santos, F. C. Santos, J. M. Pacheco, Social norm complexity and past reputations in the evolution of cooperation. *Nature* **555**, 242–245 (2018).
34. F. Santos, J. Pacheco, F. Santos, "Social norms of cooperation with costly reputation building" in *Proceedings of the AAAI Conference on Artificial Intelligence* (2018), vol. 32.
35. C. Xia, C. Gracia-Lázaro, Y. Moreno, Effect of memory, intolerance, and second-order reputation on cooperation. *Chaos: Interdiscip. J. Nonlinear Sci.* **30**, 063122 (2020).
36. F. P. Santos, J. M. Pacheco, F. C. Santos, The complexity of human cooperation under indirect reciprocity. *Philos. Trans. R. Soc. B* **376**, 20200291 (2021).
37. S. Podder, S. Righi, K. Takács, Local reputation, local selection, and the leading eight norms. *Sci. Rep.* **11**, 1–10 (2021).
38. S. Uchida, Effect of private information on indirect reciprocity. *Phys. Rev. E* **82**, 036111 (2010).
39. S. Uchida, T. Sasaki, Effect of assessment error and private information on stern-judging in indirect reciprocity. *Chaos, Solitons Fractals* **56**, 175–180 (2013).
40. I. Okada, T. Sasaki, Y. Nakai, Tolerant indirect reciprocity can boost social welfare through solidarity with unconditional cooperators in private monitoring. *Sci. Rep.* **7**, 1–11 (2017).
41. I. Okada, T. Sasaki, Y. Nakai, A solution for private assessment in indirect reciprocity using solitary observation. *J. Theor. Biol.* **455**, 7–15 (2018).
42. C. Hilbe, L. Schmid, J. Tkadlec, K. Chatterjee, M. A. Nowak, Indirect reciprocity with private, noisy, and incomplete information. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 12241–12246 (2018).

43. S. Bowles, H. Gintis, *A Cooperative Species* (Princeton University Press, 2011).
44. I. Okada, A review of theoretical studies on indirect reciprocity. *Games* **11**, 27 (2020).
45. H. Yamamoto, I. Okada, S. Uchida, T. Sasaki, A norm knockout method on indirect reciprocity to reveal indispensable norms. *Sci. Rep.* **7**, 1–7 (2017).
46. S. Lee, Y. Murase, S. K. Baek, Local stability of cooperation in a continuous model of indirect reciprocity. *Sci. Rep.* **11**, 1–13 (2021).
47. S. Lee, Y. Murase, S. K. Baek, A second-order stability analysis for the continuous model of indirect reciprocity. *J. Theor. Biol*. **548**, 111202. (2022).
48. E. Brush, Å. Brännström, U. Dieckmann, Indirect reciprocity with negative assortment and limited information can promote cooperation. *J. Theor. Biol.* **443**, 56–65 (2018).
49. R. M. Whitaker, G. B. Colombo, D. G. Rand, Indirect reciprocity and the evolution of prejudicial groups. *Sci. Rep.* **8**, 1–14 (2018).
50. A. L. Radzvilavicius, A. J. Stewart, J. B. Plotkin, Evolution of empathetic moral evaluation. *Elife* **8**, e44269 (2019).
51. M. Krellner, T. A. Han, "Putting oneself in everybody's shoes-pleasing enables indirect reciprocity under private assessments" in *Artificial Life Conference Proceedings 32* (MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA, 2020), pp. 402–410.
52. J. Quan *et al*., Withhold-judgment and punishment promote cooperation in indirect reciprocity under incomplete information. *EPL (Europhys. Lett.)* **128**, 28001 (2020).
53. M. Krellner, T. A. Han, Pleasing enhances indirect reciprocity-based cooperation under private assessment. *Artif. Life* **31**, 1–31 (2021).
54. L. Schmid, P. Shati, C. Hilbe, K. Chatterjee, The evolution of indirect reciprocity under action and assessment generosity. *Sci. Rep.* **11**, 1–14 (2021).
55. T. A. Kessinger, J. B. Plotkin, Indirect reciprocity in populations with group structure. arXiv [Preprint] (2022). http://arxiv.org/abs/2204.10811 (Accessed 25 January 2023).
56. J. Quan, J. Nie, W. Chen, X. Wang, Keeping or reversing social norms promote cooperation by enhancing indirect reciprocity. *Chaos, Solitons Fractals* **158**, 111986 (2022).
57. P. Gu, Y. Zhang, "Reputation-based rewiring promotes cooperation in complex network" in *Advances in Guidance, Navigation and Control* (Springer, 2022), pp. 1405–1415.
58. K. Sigmund, Moral assessment in indirect reciprocity. *J. Theor. Biol.* **299**, 25–30 (2012).
59. K. Oishi, T. Shimada, N. Ito, Group formation through indirect reciprocity. *Phys. Rev. E* **87**, 030801 (2013).
60. Y. Fujimoto, H. Ohtsuki, Reputation structure in indirect reciprocity under noisy and private assessment. *Sci. Rep.* **12**, 1–13 (2022).
61. J. Maynard-Smith, G. R. Price, The logic of animal conflict. *Nature* **246**, 15–18 (1973).
62. J. Maynard-Smith, *Evolution and the Theory of Games* (Cambridge University Press, 1982).
63. M. A. Nowak, K. Sigmund, The dynamics of indirect reciprocity. *J. Theor. Biol.* **194**, 561–574 (1998).
64. H. Ohtsuki, Y. Iwasa, Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. *J. Theor. Biol.* **244**, 518–531 (2007).
65. J. Hofbauer, K. Sigmund, *Evolutionary Games and Population Dynamics* (Cambridge University Press, 1998).
66. R. Sugden, *The Economics of Rights, Co-operation and Welfare* (Basil Blackwell, 1986).