PNAS RESEARCH ARTICLE | SOCIAL SCIENCES EVOLUTION

Indirect reciprocity with dual private assessment

Yukari Jessica Tham[a], Christian Hilbe[b], and Yohsuke Murase[c,d,e,1]

Edited by Marcus W. Feldman, Stanford University, Stanford, CA; received July 12, 2026; accepted July 29, 2026

People often cooperate out of concern for their reputation. The corresponding theory of indirect reciprocity predicts that such cooperation can only evolve if people's opinions of each other are sufficiently correlated. This correlation, however, can be difficult to achieve when individuals form their opinions independently from one another, as in private assessment models. In that case, errors can generate disagreements that propagate throughout a population. Here, we identify a simple mechanism that mitigates this problem. Most prior work assumes that observers revise only the reputations of donors—the individuals who choose whether to cooperate or defect. We instead allow observers to also revise the reputations of recipients—the individuals affected by the donors' choices. Using analytical calculations and simulations, we show that such dual reputation updates help synchronize opinions and facilitate cooperation. To demonstrate this approach, we focus on a particularly simple rule, which we call Recipient Image Scoring (RIS). Under RIS, recipients are assessed as good whenever donors choose to cooperate with them. We show that this rule for judging recipients effectively complements the classical leading-eight rules for judging donors. Our results establish dual assessment as a simple mechanism for cooperation that relies only on directly observable information.

evolutionary game theory | social norms | indirect reciprocity | evolution of cooperation

Significance

People's cooperative intentions are often shaped by the social norms of their community. These norms determine how people ought to act and how their actions affect reputations. However, models of evolutionary game theory suggest that such norms can only sustain cooperation when people agree on each other's reputations. Prior research has thus explored mechanisms that people use to reduce disagreements, such as gossip or empathetic perspective-taking. Here, we describe a more immediate mechanism. When observing a donor's action, people not only revise the donor's reputation but also the recipient's. We show that such dual reputation updates strongly align reputation judgments within a community. This alignment in turn facilitates the evolution of social norms previously deemed unstable.

Cooperation is at the core of many of our daily social interactions. Yet how it emerges and is maintained—especially in large and anonymous populations—remains a key question in evolutionary biology and the social sciences (1, 2). One mechanism that may explain the remarkable extent and scale of human cooperation is indirect reciprocity (3, 4). This mechanism highlights the role of social norms, which prescribe how people should act in order to maintain a good reputation (5). This literature regards good reputations as a valuable commodity. Indeed, several experiments confirm that people with a reputation of being cooperative are more likely to receive help or monetary benefits in future, even from third parties they have never interacted with (6–8). As a result, people become more cooperative when their actions are observable (9–12).

To explore the effect of reputations on cooperation formally, researchers analyze social dilemmas in evolving populations (13–15). These models assume that at each time point, two population members are selected randomly. The first member ("donor") then decides whether to cooperate with the other ("recipient"). This decision is observed by the remaining population members, who may use this information to update the donor's reputation. The updated reputation depends on the donor's action, the social norm in place, and the observability of the interaction. In the most basic setup, reputations are binary (good or bad), and interactions are observed by everyone. Observation errors, if they occur, affect all observers identically (16). For this so-called public assessment regime, it is possible to determine all social norms (of a certain complexity) that sustain cooperation (16, 17). The corresponding "leading-eight" norms have become a baseline in the field. They all have in common that cooperating with a good recipient ought to be evaluated as good, and that defecting against good recipients is bad. They disagree, however, on how to evaluate actions toward bad recipients; see Table 1. Subsequent work has explored several extensions, including when these norms are evolutionarily stable (18, 19), the role of costly punishment (20–22), and models with nonbinary reputations (23–25).

While the public-assessment assumption greatly simplifies analytical calculations, it can be unrealistic. Under this assumption, all population members are assumed to hold the same view of third parties. For example, if Alice deems Bob bad, all others must agree, including Bob himself. This remains true even in noisy public-assessment models: Errors may lead everyone to assign the wrong reputation, but they do not induce disagreements across observers. In everyday life, however, such disagreements are widespread. They naturally arise when people differ in the information they have or in how they interpret a given interaction. Models that allow for such "private assessment" suggest that, once

Author affiliations: [a]Graduate School of Humanities, Kobe University, Kobe 657-8501, Japan; [b]Interdisciplinary Transformation University, Linz 4040, Austria; [c]RIKEN Center for Interdisciplinary Theoretical and Mathematical Science, Wako 351-0198, Japan; [d]RIKEN Center for Computational Science, Kobe 650-0047, Japan; and [e]Graduate School of Science and Engineering, Saitama University, Saitama 338-8570, Japan

Author contributions: Y.J.T. and Y.M. designed research; Y.J.T. and Y.M. performed research; Y.J.T. and Y.M. analyzed data; and Y.J.T., C.H., and Y.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: yohsuke.murase@riken.jp.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2624656123/-/DCSupplemental.

Published August 27, 2026.

Table 1. The leading-eight social norms

	(G, G)			(G, B)			(B, G)			(B, B)			
	S	$R_1(C)$	$R_1(D)$	S	$R_1(C)$	$R_1(D)$	S	$R_1(C)$	$R_1(D)$	S	$R_1(C)$	$R_1(D)$	Type
L1 (standing)	C	1	0	D	**1**	1	C	1	0	**C**	**1**	**0**	I
L2 (consistent standing)	C	1	0	D	**0**	1	C	1	0	**C**	**1**	**0**	II
L3 (simple standing)	C	1	0	D	**1**	1	C	1	0	**D**	**1**	**1**	I
L4	C	1	0	D	**1**	1	C	1	0	**D**	**0**	**1**	I
L5	C	1	0	D	**0**	1	C	1	0	**D**	**1**	**1**	II
L6 (stern judging)	C	1	0	D	**0**	1	C	1	0	**D**	**0**	**1**	III
L7 (staying)	C	1	0	D	**1**	1	C	1	0	**D**	**0**	**0**	I
L8 (judging)	C	1	0	D	**0**	1	C	1	0	**D**	**0**	**0**	III

The top row (X, Y) indicates the reputations of the donor and the recipient, respectively. For example, (G, B) denotes a good donor who interacts with a bad recipient. The entries S, $R_1(C)$, and $R_1(D)$ specify the prescribed action (cooperate or defect), how a donor is evaluated when they cooperate, and how they are evaluated when they defect, respectively. An entry of 1 means that the donor is assessed as good; an entry of 0 means that the donor is assessed as bad. Columns highlighted in bold are those in which the leading-eight norms differ. In the rightmost column, we classify the norms into three types according to their evolutionary stability properties under private assessment, following ref. 32.

individuals disagree about how they view others, these disagreements tend to proliferate over time, even when the error rate is arbitrarily small (26–29). For indirect reciprocity to be effective, however, individual opinions need to be sufficiently correlated (30). Therefore, most leading-eight norms cease to be stable under private assessment (31, 32). These findings shift the focus of research on indirect reciprocity. Instead of asking "How can social norms help sustain cooperation?," the relevant question becomes "How can communities avoid disagreements?."

Prior research has described several mechanisms that increase the correlation between individuals' opinions. These mechanisms include empathy (33), gossip (34–36), pleasing (37), and public institutions (38, 39). In addition, several studies describe the positive effects of more nuanced social norms, which go beyond mere good-or-bad characterizations (40–42). The present study highlights a more elementary source of opinion synchronization. Our approach emphasizes the role of shared observations of recent interactions. We argue that after a donor decides to cooperate with or defect against a given recipient, observers may not only update their opinions of the donor, but also of the recipient. Although recipients are entirely passive, there are at least two reasons why an observer may still find it sensible to reevaluate the recipient's reputation. First, the interaction may provide observers with information about how others see the recipient. For example, if Alice defects against Bob, Charlie may infer that Alice considers Bob to be bad—a judgment Charlie may want to copy. Second, the observer may learn that a previous wrongdoing has already been sanctioned. For example, if Charlie considered Bob to be bad to begin with, he may regard Bob's reputation as restored after Alice's defection, reasoning that Bob has suffered enough for his wrongdoing. Indeed, empirical evidence suggests that how people evaluate or treat a person can be influenced by how others treat that person, even when the person has taken no action (43–45). Such "dual reputation updates" have a comparatively minor effect in public assessment models (18). Interestingly, however, we show that under private assessment, they have a major effect on opinion synchronization, making them particularly relevant for the evolution of cooperation.

Among the many possible assessment rules for updating a recipient's reputation, we identify a rule that combines strong performance with a particularly simple structure: Recipient Image Scoring (RIS). Similar to classical Image Scoring (IS) (46, 47), RIS associates cooperation with a good reputation and defection with a bad one. However, unlike IS, which is used to assess the donor, RIS is used to assess the recipient. Recipients are judged as good if they receive cooperation, and as bad otherwise. This rule can be interpreted as a form of conformity (48), in the sense that observers change their evaluations and behavior toward a recipient to match the response of the donor. We find that norms that combine the leading-eight (for assessing donors) with RIS (for assessing recipients) tend to be more cooperative and stable compared to the classical leading-eight without recipient assessment. More generally, our analysis shows that incorporating dual assessments generates a large, hitherto understudied class of social norms that naturally facilitates the evolution of cooperation.

Model and Results

Modeling Indirect Reciprocity. Our model builds on previous work on indirect reciprocity under private assessment (26–28). We consider a well-mixed population of N individuals, which we refer to as players. In each round, two players are randomly chosen: a donor and a recipient. The donor decides whether to cooperate (C) or to defect (D). Cooperation incurs a cost c for the donor and results in a benefit b for the recipient, where $b > c > 0$. Defection yields a payoff of zero for both players. In the main text, we consider the case of perfect observation: The donor's choice is observed by all other population members; see Fig. 1*A* for a schematic illustration (in *SI Appendix*, we discuss an extension to imperfect observation, where only a fraction of population members observe any given interaction). The above process is iterated for many rounds, with a new donor-recipient pair drawn each time.

During this sequence of donation games, players form private opinions about each other. The opinion that player i holds about player j is denoted by m_{ij}. Players also hold opinions about themselves. Opinions are binary, either good (G) or bad (B), and they can change over time. We refer to $M = (m_{ij})$ as the image matrix. It collects all players' views of each other at a given time point in an $N \times N$ table. Row i summarizes how player i views all others; column j summarizes how all others view player j. In previous studies of private assessment, each observer updates only their opinion about the donor. By contrast, we study dual reputation updates (18, 49). Here, observers can also revise their opinion of the recipient; see Fig. 1*B*. This approach is related to previous work where observers update their assessment of either the donor or the recipient (50). However, it differs in that observers can update both players' reputations simultaneously. As we show below, dual updates have major implications for the resulting reputation dynamics under private assessment.

The Space of Social Norms. To make their decisions in these pairwise games, players follow a social norm. The norm consists

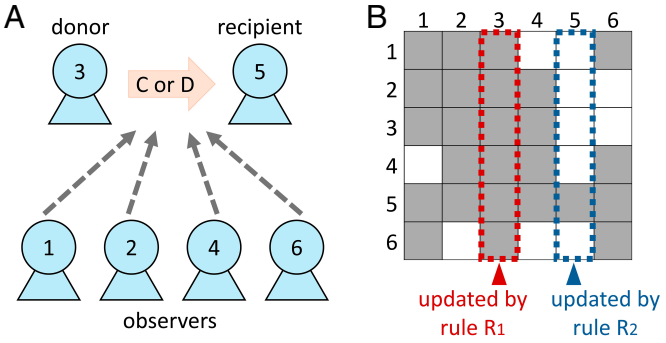

Fig. 1. Indirect reciprocity with dual private assessment. (*A*) In each round, a donor and a recipient are randomly drawn from the population. Then the donor decides whether to cooperate (*C*) or defect (*D*) based on their opinion and their action rule *S*. All other players observe this interaction and privately update their opinions about both the donor and the recipient. The updated opinion depends on their assessment rules R_1 and R_2. (*B*) After the interaction has taken place, the columns corresponding to the donor and the recipient in the image matrix are updated accordingly, depending on the revised views of the population members. Each cell is shaded if the opinion is good (*G*) and left white if the opinion is bad (*B*). When opinions are updated, assessment errors for the donor and the recipient may occur independently with probabilities μ_1 and μ_2, respectively. In the depicted case, all population members view the donor as good, whereas most of them view the recipient as bad.

of three components, (S, R_1, R_2). The first component S is the "action rule." It determines whether players cooperate when acting as a donor. The chosen action $a_{ij} \in \{C, D\}$ of donor i against recipient j may depend on their own and the recipient's reputations (from the viewpoint of the donor). Hence, the action rule is a function $a_{ij} = S(m_{ii}, m_{ij})$. It assigns to any pair of reputations in $\{G, B\} \times \{G, B\}$ a response in $\{C, D\}$.

Social norms also determine how players update their opinions over time. With dual assessments, there are two such "assessment rules"—one for the donor (R_1) and one for the recipient (R_2). To describe them, suppose observer k witnesses an interaction between donor i and recipient j. Then the observer's updated views depend on their previous view of the donor m_{ki}, their previous view of the recipient m_{kj}, and the donor's action a. We allow the updated reputations to be stochastic (18, 51). The corresponding function $R_1(m_{ki}, m_{kj}, a)$ gives the probability that the observer's updated opinion of the donor is good. Hence, it is a function from $\{G, B\} \times \{G, B\} \times \{C, D\}$ into the unit interval $[0, 1]$. Analogously, the observer assigns a good reputation to the recipient with probability $R_2(m_{ki}, m_{kj}, a)$.

It is instructive to consider two examples. First, suppose a donor defects against a recipient. Suppose further that, prior to the interaction, an observer views the donor as good, the recipient as bad, and that the observer's assessment rule satisfies $R_1(G, B, D) = 1$ and $R_2(G, B, D) = 0$. After the interaction has taken place, the observer thus again views the donor as good and the recipient as bad—both with certainty. As another example, suppose $R_2(x, y, a) = \delta_{y,G}$ for all $x, y \in \{G, B\}$ and $a \in \{C, D\}$. Here, $\delta_{x,y}$ is the Kronecker delta: It is equal to one if $x = y$, and equal to zero otherwise. This rule implies that the recipient's reputation is always left unchanged. We thus recover the classical single-update model of previous studies (16, 46).

Assessments can be noisy. We assume each assessment is flipped with probability $\mu_1 > 0$ for donors and $\mu_2 > 0$ for recipients. These errors occur independently across observers and the two targets. In addition, we allow implementation errors. If a donor intends to cooperate, they instead defect with probability $\mu_e \geq 0$ (e.g., due to a lack of resources). In line with the literature, we assume intended defections are implemented correctly. Noisy assessments are not only more realistic; they also make the reputation dynamics ergodic. Given the players' social norms, their (average) views of each other are guaranteed to converge to a stationary state, independent of their initial views.

In total, there are $2^8 = 256$ deterministic assessment rules for each of R_1 and R_2. However, in the main text, we primarily focus on those R_1 that belong to the leading-eight (Table 1), since they are evolutionarily stable under public assessment (16). For R_2, we first focus on two rules (Table 2). The first is the above-mentioned $R_2(x, y, a) = \delta_{y,G}$, where the recipient's reputation remains unchanged. We refer to the social norms incorporating this rule as "base norms." The second is the RIS rule: $R_2(x, y, a) = \delta_{a,C}$. Here, the recipient is assessed as good if the donor cooperates with the recipient, and as bad if the donor defects against the recipient. When a leading-eight R_1-rule is combined with the RIS R_2-rule, we refer to the resulting norms as L1-RIS, …, L8-RIS; when it is combined with the base R_2-rule, we refer to the resulting norms as L1-base, …, L8-base. Collectively, we denote these norms as L*x*-RIS and L*x*-base, respectively. We discuss a comprehensive analysis of the full space of possible norms (S, R_1, R_2) further below.

Exploring the Possible Reputation Dynamics. To get an impression of how dual assessments affect reputations, we start with some examples. To this end, we consider homogeneous populations in which everyone adopts the same norm. We assume that players adopt one of the four leading-eight norms L3, L5, L6, or L8; for the remaining four norms, see *SI Appendix*. We focus on these norms because they cover the three different types of leading-eight norms described in the literature (32, 52). These types differ in how they evaluate cooperation with bad recipients, and therefore in how easily disagreements spread under private assessment. Type I norms (L1, L3, L4, L7) are the most lenient, with $R_1(G, B, C) = 1$. Here, good donors who cooperate remain good, regardless of the recipient's standing. This not only helps maintain a high number of good opinions but also helps resolve occasional disagreements. Type II norms (L2, L5) instead have $R_1(G, B, C) = 0$ but $R_1(B, B, C) = 1$. These norms are more sensitive to disagreements, because the evaluation of cooperation with bad recipients depends on the donor's reputation. Finally, type III norms are the strictest. They have $R_1(G, B, C) = R_1(B, B, C) = 0$; hence, cooperation with bad recipients is judged negatively. Here, if a donor and an observer disagree about the recipient's reputation, the donor's action appears justified to the donor but unjustified to the observer. The observer may then assign the donor a bad reputation, which not only increases the number of bad opinions but also allows an initial disagreement about one individual to spread to another. Consequently, populations of type III show comparatively low cooperation rates, and they tend to be evolutionarily unstable (26–28).

These differences among the three types can also be seen in Fig. 2*A*. Here, the left column shows snapshots of the population's image matrix $M = (m_{ij})$ after a sufficiently long time (for L*x*-base). For type I and type II norms, we observe that most individuals are viewed as good. Moreover, opinions are largely correlated (each column of the matrix is either mostly colored or mostly white). In contrast, the image matrices for type III norms show a different pattern. As the *Top Left* panel of Fig. 2*A* shows, "Stern Judging" (L6, 53–56), is particularly prone to disagreements (26–29). Here, the fraction of good opinions approaches one half, and opinions are completely independent. This independence is problematic as it renders

Table 2. Representative rules for recipient assessment

	(G,G)		(G,B)		(B,G)		(B,B)	
	$R_2(C)$	$R_2(D)$	$R_2(C)$	$R_2(D)$	$R_2(C)$	$R_2(D)$	$R_2(C)$	$R_2(D)$
Base	1	1	0	0	1	1	0	0
Recipient image scoring (RIS)	1	0	1	0	1	0	1	0
Good-donor trusting (GDT)	1	0	1	0	1	1	0	0
Always good (ALLG)	1	1	1	1	1	1	1	1

The top row (X,Y) indicates the original reputations of the donor and the recipient, respectively. The entries $R_2(C)$ and $R_2(D)$ give the updated assessment of the recipient when the donor cooperates (C) and when the donor defects (D), respectively. An entry of 1 means that the recipient is assessed as good, whereas an entry of 0 means that the recipient is assessed as bad. The base rule corresponds to the classical single-update model; here, the recipient's reputation remains unaffected.

conditional cooperation unstable: Even if Alice views Bob as good, cooperating with him does not help her to improve her reputation, because her view is not indicative of how others view him. We also observe largely uncorrelated opinions in the image matrix for the other type III norm, "Judging" (L8); but here, the proportion of bad entries is even larger.

The situation changes markedly once the leading-eight norms are combined with RIS (*Right* panels in Fig. 2*A*). For example, while opinions under L6-base had been completely uncorrelated, they become highly correlated under L6-RIS and most players regard each other as good. Only occasionally is a player regarded as bad by almost everyone, corresponding to the white vertical stripes. These stripes arise from rare disagreements (27): If Alice mistakenly views Bob as bad and hence defects against him, the other players, who view Bob as good, evaluate Alice's defection as bad. However, these white stripes need not render cooperation unstable, as long as opinions remain sufficiently correlated (30). A similar conclusion holds for L8. While L8-base results in many bad and largely uncorrelated opinions, L8-RIS leads to a large proportion of good, correlated opinions (see the second row in Fig. 2*A*). For the other two norms L5 and L3 as well, the RIS regime yields more good and more correlated opinions than the base regime, although the difference is now smaller (because already in the base setup, opinions tended to be good and correlated).

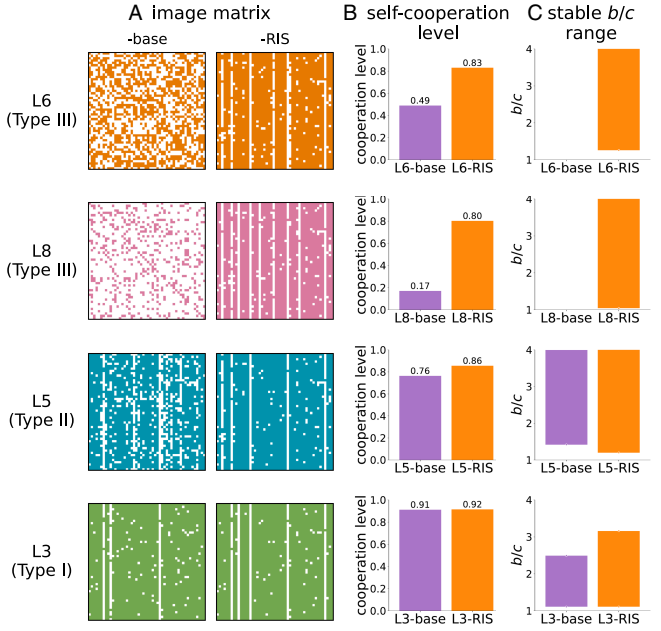

Fig. 2. Comparison of the base and RIS norms. From top to bottom, we consider four different donor-assessment rules, $R_1 \in \{L6, L8, L5, L3\}$. In addition, we consider two different recipient-assessment rules: base and RIS. (*A*) For each combination of rules, we consider a homogeneous population ($N = 50$) and simulate its reputation dynamics. Assessment and implementation errors occur at rate $\mu_1 = \mu_2 = \mu_e = 0.02$. The panels depict a snapshot of the image matrix after a sufficiently long time. Each cell is colored (G) or left white (B) according to the opinion m_{ij}. The first variable i (corresponding to a row of the matrix) represents the player holding the opinion; the second index j (corresponding to a column) represents the player being judged. (*B*) Whereas the previous image matrices are random snapshots, which may not be fully indicative of the average goodness of players, here we display how often players cooperate on average. (*C*) In addition, we depict for which b/c values the respective resident norm is stable against ALLC and ALLD mutants. Since L6-base and L8-base do not have any stable b/c region, the respective areas are left blank.

After discussing these special cases, a natural question is how RIS compares to other assessment rules for recipients. Fig. 3 shows simulation results for all possible deterministic R_2 rules, with the previous four R_1 rules fixed. In each case, we depict a norm's self-cooperation level (on the x-axis) and the minimum benefit-to-cost ratio required for the norm to be stable against ALLD (on the y-axis). Well-performing norms are those toward the *Bottom Right* corner of each panel. In each case, we find $R_2 =$ RIS (orange squares) to be among the best rules. Consistent with previous observations, the effect of the RIS rule is most pronounced for L6 and L8, where the base R_2 rule performs poorly (and is thus invisible in the figure).

To further highlight the effectiveness of Lx-RIS, we also show results for another recipient assessment rule. The "good-donor trusting" (GDT) rule is a hybrid: It adopts the RIS rule when the donor is perceived as good, and the base rule when the donor is perceived as bad;

$$R_2(x,y,a) = \begin{cases} \delta_{a,C} & \text{if } x = G \\ \delta_{y,G} & \text{if } x = B \end{cases} \quad [1]$$

In Fig. 3, the Lx-GDT norm is indicated by purple triangles. For L6 and L8, this rule does not improve performance compared to the base rule. In fact, in both cases, the Lx-GDT norm does not allow for evolutionarily stable cooperation at all. For L5 and

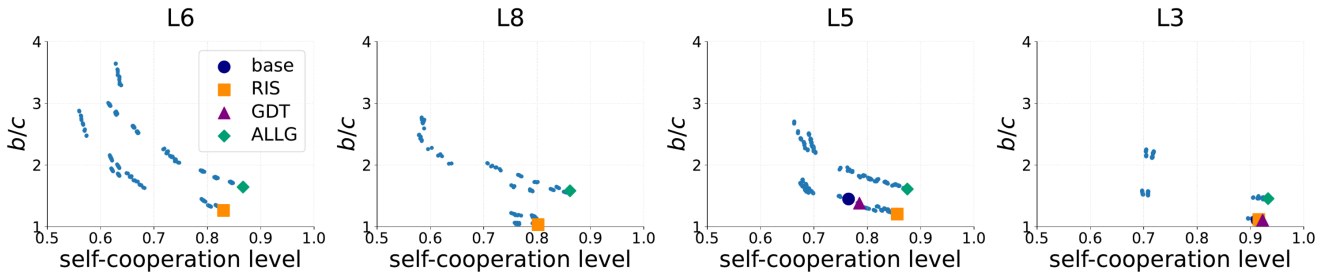

Fig. 3. Cooperation levels and stability against ALLD across recipient-assessment rules. We pair each donor-assessment rule, $R_1 \in$ {L6, L8, L5, L3}, with all possible deterministic recipient-assessment rules, R_2. For each pair, we consider a homogeneous population, in which everyone applies the respective norm. For that population, we compute its self-cooperation level (x-axis) and the critical b/c ratio above which the norm is stable against ALLD (y-axis). Points toward the *Bottom Right* corner of each panel correspond to norms that perform well with respect to both dimensions. We highlight four particular recipient-assessment rules: base, RIS, GDT, and ALLG. For L6 and L8, there is no b/c region for which the respective base or GDT norm is stable; hence, those points are not shown. The population size is $N = 50$, and the assessment and implementation error rates are $\mu_1 = \mu_2 = \mu_e = 0.02$.

L3, too, the performance of Lx-GDT is similar to that of the base norm. Overall, the GDT rule yields no advantage, despite its intuitive appeal. Another interesting R_2 rule is the ALLG rule, under which the recipient is always regarded as good. Although this rule yields the highest self-cooperation level, it requires larger b/c ratios to be stable against ALLD mutants than the RIS rule. Overall, RIS strikes a good balance between being generous enough to allow for cooperation and being strict enough to prevent exploitation by ALLD.

Recovery Analysis of Opinion Disagreements. To gain some analytical insight into why RIS leads to more aligned opinions compared to the base rule, we perform a recovery analysis (27). To this end, we consider a homogeneous population where everyone adopts the same norm. Additionally, we assume errors to be rare, such that $\mu \to 0$ for all error rates. Initially, we suppose everyone regards everyone as good, with a single exception: Player 1 regards player 2 as bad. Given this initial state, we analyze how long the population takes on average to revert to a state in which everyone is regarded as good, depending on the norm in place. This kind of analysis provides insights into how robust different norms are with respect to initial disagreements.

For the standard case when donors are assessed according to a leading-eight rule, and recipients according to the base rule, this recovery analysis was carried out in ref. 27. That study reports two results. *i*) For four norms, recovery is not even guaranteed (L4, L6, L7, L8). The recovery probability is lowest for L6 and L8, for which it is $1 - 1/N$. *ii*) The expected time until recovery (conditional on recovery occurring) is

$$\begin{cases} O(N) & \text{for type I norms (L1, L3, L4, L7),} \\ O(N \log N) & \text{for type II norms (L2, L5),} \\ N \cdot H_N - N & \text{for type III norms (L6, L8).} \end{cases} \quad [2]$$

Here, $O(\cdot)$ is the Big-O notation and $H_N = \sum_{k=1}^{N} 1/k$ is the N-th harmonic number. In particular, recovery is fastest for type I norms, followed by type II norms.

In the following, we sketch how to perform a similar recovery analysis for L6-RIS (see *SI Appendix* for details). Instead of considering all possible 2^{N^2} configurations of the image matrix, it turns out that the state space can be simplified considerably. To this end, we first note that at any given point in time, there can be disagreement about at most one player's reputation. This is because after any interaction, all players agree on their opinion of the recipient (due to the assumption of rare errors). Therefore, the number of disagreements cannot increase. In the following, we distinguish between player 1 and the "remaining players." For L6-RIS, one can furthermore show that the remaining players always agree on how they assess any given player.

As a result of these considerations, we can characterize the state of the image matrix by a triplet (α, β, γ). Here, $\alpha \in \{0, 1\}$ indicates whether the remaining players consider player 1 as good. The next entry $\beta \in \{0, 1, \ldots, N-1\}$ indicates how many remaining players are considered good, from the viewpoint of the remaining players. Finally, $\gamma \in \{0, 1, -, +\}$ indicates where the disagreement exists (if there is any). If $\gamma = 0$, there is no disagreement in the opinions of player 1 and the remaining players. If $\gamma = 1$, they disagree about player 1: Either player 1 considers player 1 to be good whereas the others do not, or vice versa. If $\gamma = -$, they disagree about one of the remaining players, say player j. Here, while player 1 considers j as bad, the remaining players consider j as good. Finally, $\gamma = +$ corresponds to the opposite case: While player 1 considers j to be good, everyone else considers j to be bad. Interestingly, this final case never occurs during recovery.

For this small set of states, we can explicitly compute the transition probabilities from any state to any other. For sufficiently small N, this allows us to compute the recovery time exactly (without simulations). Moreover, for large N, we prove that the expected recovery time is of order $O(N)$, and therefore faster than that of L6-base.

Fig. 4 shows simulation results for the recovery time of all Lx-base and Lx-RIS norms, as a function of population size N. We observe four distinct curves, with the following regularities. *i*) For each leading-eight norm, an Lx-base population takes longer to recover than an Lx-RIS population. *ii*) For each R_2 rule, type I norms recover faster than type II norms, which in turn recover faster than type III norms. The only exception occurs for type II-RIS and type III-RIS; they recover equally fast. Moreover, the analytical calculation for L6-RIS matches the numerical results, which is reassuring. Overall, we find that compared to the base rule, RIS improves the performance of all leading-eight norms, as disagreements are resolved more quickly.

For the above analysis, we have studied recovery from a state with a single disagreement. However, a similar recovery analysis is feasible for more complicated initial configurations. In *SI Appendix*, we explore recovery from an image matrix in which initial opinions are assigned randomly, taking Stern-Judging (L6) as an example. For L6-base, we find that populations typically do not recover at all, even when there are no further errors ($\mu \to 0$). For L6-RIS, on the other hand, recovery is guaranteed, and it happens relatively quickly (*SI Appendix*, Fig. S6).

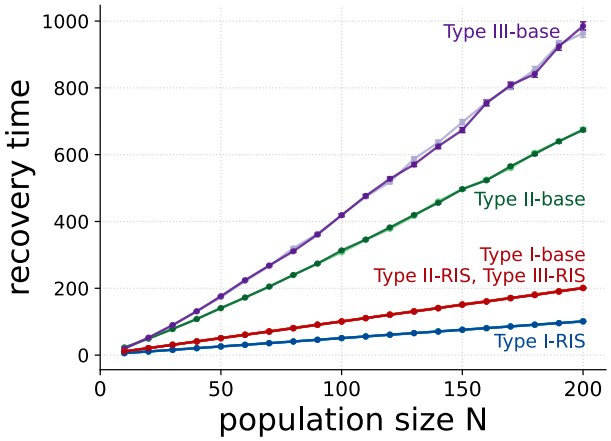

Fig. 4. Recovery time as a function of population size. We explore how quickly a homogeneous population recovers from a single disagreement, in the limit of rare errors. Initially, all players consider everyone good, except for player 1, who considers player 2 bad. The expected time it takes to reach a state where all players are considered good by everyone depends on the adopted social norm. Simulations suggest that among the sixteen norms considered, there are four distinct cases. For L6-RIS, we also compute recovery times analytically. The respective results are indistinguishable from the numerically derived values.

What Makes RIS Particularly Effective. To a large extent, the previous results are driven by one of the key properties of the RIS rule: Assigned reputations only depend on the donor's action. This property facilitates opinion synchronization because observers typically agree on which action was chosen, even if they disagree on the prior reputations of the donor or the recipient. However, RIS is not the only rule with this property. Unconditional rules or other rules depending only on the observed action might be equally effective (i.e., zeroth- or first-order norms). For instance, the above-mentioned ALLG with $R_2(x,y,a) = 1$, Always Bad (ALLB) with $R_2(x,y,a) = 0$, and Anti-RIS with $R_2(x,y,a) = \delta_{a,D}$ would achieve similar levels of opinion synchronization. Still, these alternative rules are inferior to the RIS rule, as we argue below.

To understand why, consider an idealized scenario where all players share the same opinion about each other: the public assessment model. In that case, it is possible to explicitly derive a necessary and sufficient condition for a dual assessment norm to be a cooperative ESS (18). When R_1 is a leading-eight rule, the corresponding condition reads as follows (see *SI Appendix* for details):

$$\begin{cases} R_2(G,G,C) = 1 \\ R_2(G,B,D) + R_2(B,G,C) > 0 \\ b/c > 1 + R_2(G,B,D). \end{cases} \quad [3]$$

Among the unconditional or first-order rules discussed above, only the RIS rule satisfies these conditions for $b/c > 1$. The other rules fail because they either reduce self-cooperation (ALLB or Anti-RIS), or because they require a higher b/c ratio to be stable (ALLG requires $b/c > 2$). Thus, RIS is particularly effective because it balances simplicity, which facilitates opinion synchronization, with the requirements for mutual cooperation and stability.

Evolutionary Dynamics of Social Norms. In our analysis so far, we have taken the players' social norms as given and fixed. This has allowed us to explore how a social norm affects the resulting reputation dynamics. This analysis, however, tells us little about the dynamics when players differ in their applied social norms, or when they can choose which norm to apply. To explore such a setup, in the following, we let players' social norms evolve.

To this end, we study a pairwise comparison process (61). We assume players update their social norms on a slower time scale than they play their games. In each time step of the updating process, one player, say i, is randomly chosen. With probability ϵ, player i's norm is updated by a "mutation." In that case, player i randomly adopts one of three strategies: ALLC, ALLD, or the focal norm. With the remaining probability $1-\epsilon$, player i draws a random role model j from the population. Player i then compares its own long-time average payoff π_i with the role model's payoff π_j. Depending on the payoff difference, player i adopts player j's norm with probability given by a Fermi function (62),

$$\phi = \frac{1}{1+\exp[-\sigma(\pi_j - \pi_i)]}. \quad [4]$$

The parameter $\sigma \geq 0$ is the intensity of selection. It regulates how sensitive players are to payoffs. When $\sigma = 0$, payoffs are irrelevant and imitation occurs randomly. For larger values of σ, more profitable strategies are increasingly likely to be imitated.

In the following, we study this evolutionary process when mutations are rare (63–66). In that case, the population is typically monomorphic. Only occasionally is a mutant norm introduced. This mutant norm either reaches fixation or goes extinct by the time the next mutation happens. This allows us to describe the evolutionary process as a Markov chain between monomorphic populations, as described in *Materials and Methods*.

Fig. 5 shows the results. In each panel, players can choose among ALLC, ALLD, and one of the previously studied norms of the form Lx-base and Lx-RIS. As reported earlier, none of the base norms is particularly successful (27). L6-base and L8-base prove ineffective as they are quickly invaded by ALLD. L3-base is vulnerable to invasion by ALLC, which in turn facilitates the subsequent invasion by ALLD. L5-base is more robust than the other three. However, once the population moves toward ALLD, this norm has difficulty reinvading; since $R_1(B,B,D) = 1$, L5-base players frequently assign a good reputation to defectors, which renders this norm ineffective in a pairwise competition.

In contrast, all four Lx-RIS norms considered are stable against ALLD and most of them evolve naturally (Fig. 5, *Bottom* row). Only L3-RIS hardly evolves, because ALLC mutants can invade through neutral drift. A more comprehensive analysis including all deterministic R_2 rules is shown in Fig. 6. Here, the x-axis shows a norm's self-cooperation level if adopted by everyone. The y-axis shows how often the norm is adopted in the long run, when competing with ALLC and ALLD. Therefore, in each panel the *Top Right* corner contains those norms that are successful both in terms of cooperation and evolutionary performance. When donors are assessed according to the norm L6, L8, or L5, again RIS is among the most successful norms. Only for L3 does there seem to be a trade-off. Here, social norms that yield high self-cooperation levels are less stable, as they are more easily invaded by ALLC.

We discuss the remaining leading-eight norms in *SI Appendix*. Interestingly, for some of these norms (L1 and L2), RIS is no longer superior. Their action rules prescribe cooperation when a bad donor interacts with a bad recipient, so under RIS, even ALLD players may be assigned a good reputation. But also in these cases, one can identify recipient assessment rules for which the resulting norm outperforms the base norm in terms of self-cooperation, evolutionary performance, or both. Moreover, if we consider variants of L1 and L2 in which the action rule is a

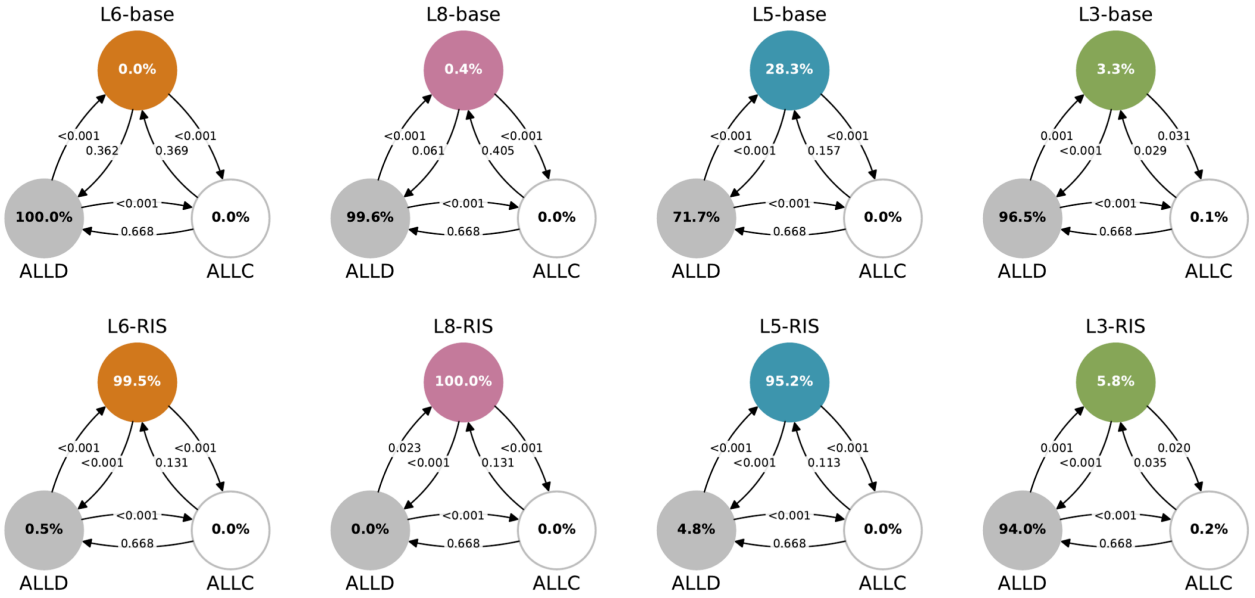

Fig. 5. Evolutionary competition among the focal norm, ALLC, and ALLD. We compare the base and RIS rules for recipient assessment across four donor-assessment rules: L6, L8, L5, and L3. The numbers near the arrows indicate the evolutionary transition probabilities, and the equilibrium fractions are denoted in the circles. Parameters: Population size $N = 50$, assessment and implementation error rates $\mu_1 = \mu_2 = \mu_e = 0.02$, intensity of selection $\sigma = 1$, and benefit-to-cost ratio $b/c = 5$.

simple discriminator, RIS again becomes superior, underscoring the robustness of our main findings.

Beyond the Leading-Eight. In the analyses above, we have varied the updating rule R_2 for recipients (e.g., Figs. 3 and 6), but we have assumed that donors are always assessed according to a leading-eight rule. We provide a more comprehensive analysis in *SI Appendix*. There, we first explore the case when donors are assessed according to the "secondary-sixteen" rules, another class of social norms that are evolutionarily stable under public information (18). Compared to the leading-eight, we find that the secondary-sixteen are generally less effective, since they require a larger b/c ratio to resist invasion by ALLD (*SI Appendix*, Fig. S4).

Second, we used a supercomputer to explore the space of all social norms (S, R_1, R_2) where the action rule is fixed to S = DISC (i.e., donors cooperate with good recipients and defect against bad ones). For each of the $256 \times 256 = 65{,}536$ possible combinations, we recorded three quantities: *i*) the norm's self-cooperation level, *ii*) how often the norm is adopted in an evolutionary competition with ALLC and ALLD, and *iii*) whether the norm is stable with respect to all possible deviations in the norm's action rule (for $b/c = 2$). We find that even within this larger space, L*x*-RIS norms are among the best-performing norms in terms of cooperation level and evolutionary stability (especially when R_1 is either L2, L5, L6, or L8, see *SI Appendix*, Fig. S9).

Finally, we repeated this analysis for all (S, R_1, R_2) norms (now also allowing for different action rules S). For all $16 \times 256 \times 256 = 1{,}048{,}576$ norms, we recorded the same three quantities as before. Based on these computations, we compiled a table (*SI Appendix*, Table S3) that contains all norms *i*) with a self-cooperation level greater than 0.8, *ii*) with an evolutionary abundance when competing with ALLC and ALLD greater than 0.95, and *iii*) that are stable against all mutations in the norm's action rule (up to a small numerical tolerance threshold). We found 61 norms that meet all criteria. They exhibit several interesting patterns. First, all of them have leading-eight R_1 rules (either L2, L5, L6, or L8) and DISC-like action rules; most use the DISC rule, whereas several use its variant with $S(B, B) = C$. Second, all of them require some form of dual reputation updating. Their R_2 rules are either RIS, combinations of the RIS and base rules (i.e., if there is any dual reputation update, it is based on RIS), or minor variants of those combinations. Third, RIS is the only R_2 rule that satisfies all three criteria with all four

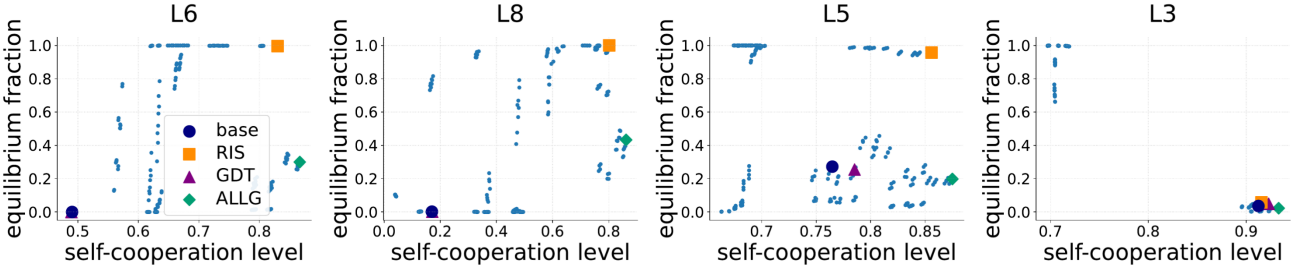

Fig. 6. Cooperation levels and evolutionary abundance across recipient-assessment rules. We sweep over all possible deterministic R_2 rules while fixing R_1 and S to be the same as L6, L8, L5, and L3, from the *Left* to the *Right* panels, respectively. The horizontal axis shows the self-cooperation levels in monomorphic populations. The vertical axis shows the equilibrium fraction of the focal norm in the evolutionary competition among the focal norm, ALLC, and ALLD. Points toward the *Top Right* corner of each panel correspond to norms that perform well with respect to both dimensions. The symbols for norms with R_2 = base, RIS, GDT, and ALLG are the same as in Fig. 3. Parameters: Population size $N = 50$, assessment and implementation error rates $\mu_1 = \mu_2 = \mu_e = 0.02$, intensity of selection $\sigma = 1$, and benefit-to-cost ratio $b/c = 5$.

leading-eight R_1 rules, L2, L5, L6, and L8. These results provide some ex-post justification for our focus on L*x*-RIS norms in the main text, while also showing that dual reputation updates can give rise to a broader class of cooperative and evolutionarily stable norms.

Discussion

Indirect reciprocity explores how individuals cooperate when their actions affect their social image. This mechanism for cooperation seems particularly relevant in large communities, in which stable pairwise relationships are less relevant (67). However, as communities grow larger, individuals are more likely to disagree about the reputations they assign to others. As shown by several models, such disagreements tend to proliferate (26–29). In this way, disagreements are a major obstacle to the evolution of indirect reciprocity: The less individuals agree on each other's reputations, the weaker their incentives to cooperate (30).

Past work has thus highlighted the importance of mechanisms that increase the correlation between individuals' opinions. To be effective, these mechanisms often rely on the acquisition of second-order knowledge: In addition to forming their own beliefs about a target, individuals need to know, or at least infer with sufficient accuracy, what third parties think of that target. For example, empathic assessment requires observers to know a donor's private view of a recipient's reputation (33). The model of pleasing involves the process of a donor acquiring information about how observers evaluate the donor and the recipient (37), and models of gossip assume that individuals acquire information about how others evaluate a focal individual (34–36). Similarly, in the model of public institutions, forming the public reputation of a focal individual requires acquiring the private opinions of a subset of the population (38).

In stark contrast, dual reputation updates offer a much simpler mechanism that does not require such second-order knowledge, which is often difficult to obtain accurately. Instead, individuals make use of a different source of readily available information: the donor's action toward the recipient. For example, if Alice sees Bob defecting against Charlie, she may revise not only her opinion of Bob, as in classical models of indirect reciprocity, but also her opinion of Charlie. Through this process, Alice may make a relatively complex inference about Bob's mental state, relying on her "theory of mind" [the ability to impute mental states, such as beliefs and intentions, to other individuals (68)]. For example, she may infer that Bob defected because he perceives Charlie as bad, or even because he expects observers to perceive Charlie as bad. Crucially, however, Alice need not know what Bob or other observers think of Charlie; all she needs is directly observable information about what Bob did to Charlie.

Interestingly, dual reputation updates have a limited effect in models of public assessment, where opinions are perfectly aligned by definition (18). But here we show that they can dramatically enhance opinion synchronization under private assessment. Importantly, we do not argue that dual assessments render any of the other mechanisms for opinion synchronization obsolete. Rather, we believe that these different mechanisms might have evolved in tandem, each contributing to the reduction of disagreements.

Among the many rules that can be used to update the reputation of passive recipients, we find that one rule is both simple and effective when combined with leading-eight norms: Recipient Image Scoring (RIS). This rule mirrors the classical IS rule, a rule previously used to update the reputation of active donors (46, 47). IS has been found to be unstable (69, 70), as it undermines a donor's ability to engage in "justified punishment" (withholding cooperation from recipients with a bad reputation). This problem occurs because, under IS, any act of punishment costs the donor their own good reputation. In contrast, this problem does not arise under RIS, because recipients do not choose whether to punish others. The combination of leading-eight norms for donors and RIS for recipients thus allows for justified punishment while minimizing disagreements. While there are other rules that are simple enough to facilitate opinion synchronization under leading-eight norms, RIS strikes a good balance between being generous enough to allow for cooperation and being strict enough to prevent exploitation.

Our study points toward several interesting theoretical questions. First, our study considered two particular sources of noise: assessment errors, in which observers mistakenly assign the wrong reputation, and implementation errors, in which donors who intend to cooperate instead defect. Another relevant source of noise is perception errors, in which observers misperceive the donor's action (21, 27). Dual assessment norms may be particularly vulnerable to this type of noise, as it simultaneously affects the updated reputation of the donor and the recipient. Similar effects may arise when players have limited observation opportunities (71), a possibility we briefly discuss in *SI Appendix*. A more systematic analysis would be an interesting direction for future work.

Second, understanding the evolution of dual assessment norms represents another important challenge. Our study investigated evolution in a limited context. We mainly focused on settings in which a focal norm competes with two extreme strategies, ALLC and ALLD, following the main approach in the previous literature (e.g. refs. 27 and 57–60). In *SI Appendix*, we also briefly explore how a focal norm competes with various mutants with different action rules. However, many more norms could be considered as competitors. For single reputation updates, previous work (52) has investigated evolution among all third-order social norms. Although expanding this framework to include dual reputation updates is a natural next step, it will be challenging due to the large number of norms.

Our study also opens several avenues for empirical research. Previous psychological and economic research has identified common knowledge, or even sufficiently high "common *p*-belief," as a crucial proximate factor of cooperation (72–74). Since dual reputation updates facilitate opinion synchronization, they may also help establish common knowledge. Exploring how dual reputation updates occur, and how they contribute to common knowledge, would deepen our understanding of the mechanisms underlying cooperative behavior. Understanding how passive recipients are evaluated may also shed light on negative aspects of human social behavior. For example, if people consciously or unconsciously apply an RIS-like rule, unfavorable treatment of a recipient may lead observers to perceive that recipient negatively, even if the recipient has done nothing wrong. As a result, people may withhold help from someone in need simply because others have failed to help them, as in the bystander effect (43). Similarly, people may become more willing to punish a target when others do so, as in conditional punishment (44), which may even escalate into ostracism (75). Conversely, favorable treatment may lead observers to mistakenly perceive a bad recipient as good, thereby inflating the recipient's reputation. These possibilities suggest that dual reputation updates may provide a useful framework not only for studying the evolution of cooperation, but also for understanding broader patterns of social perception and group dynamics.

Materials and Methods

Here, we describe how we simulate a population's reputation dynamics (e.g., Figs. 2 and 3) and its evolutionary dynamics (e.g., Figs. 5 and 6). In *SI Appendix*, we show further simulation results, derive Eq. **3**, and analyze the recovery time of L6-RIS in detail.

Simulating the Reputation Dynamics. To calculate average reputations, self-cooperation levels, and the evolutionary stability of a given social norm, we use Monte Carlo simulations. To this end, we consider a well-mixed population of size $N = 50$; results are qualitatively unchanged for larger N. Players engage in donation games, and they form opinions of each other under private assessment with dual reputation updates. In each round, a donor and a recipient are randomly selected. The donor cooperates or defects based on their opinion about the recipient and the prescribed action rule. All players then update their opinions about both the donor and the recipient according to their assessment rules and prior opinions. Errors occur at three stages: Assessment errors for the donor and the recipient occur with probabilities μ_1 and μ_2, respectively. Implementation errors occur with probability μ_e when a donor intends to cooperate. We set $\mu_1 = \mu_2 = \mu_e = 0.02$. Results remain qualitatively similar for smaller error rates. Each simulation runs for 5×10^5 rounds, but to reduce the effect of players' initial reputations, we only use the second half of the simulation to compute averages.

For Figs. 2 and 3, we calculate the self-cooperation level $p_c^{\text{res}\to\text{res}}$ as the fraction of cooperative actions during the data collection period. For these simulations, we consider a monomorphic population where all N players adopt the same norm. To calculate evolutionary stability, we introduce a single mutant adopting either ALLC or ALLD into a resident population of $N-1$ individuals with the focal norm. We track cooperation probabilities $p_c^{\text{res}\to\text{mut}}$ and $p_c^{\text{mut}\to\text{res}}$. Based on these probabilities, we compute payoffs

$$\pi_{\text{res}} = (b-c)\, p_c^{\text{res}\to\text{res}} \quad [5]$$

$$\pi_{\text{mut}} = b p_c^{\text{res}\to\text{mut}} - c p_c^{\text{mut}\to\text{res}}. \quad [6]$$

The resident norm is stable against the mutant when $\pi_{\text{res}} > \pi_{\text{mut}}$. When $p_c^{\text{res}\to\text{res}} > p_c^{\text{res}\to\text{mut}}$, this condition is equivalent to

$$\frac{b}{c} > \frac{p_c^{\text{res}\to\text{res}} - p_c^{\text{mut}\to\text{res}}}{p_c^{\text{res}\to\text{res}} - p_c^{\text{res}\to\text{mut}}}. \quad [7]$$

Conversely, when $p_c^{\text{res}\to\text{res}} < p_c^{\text{res}\to\text{mut}}$, stability requires

$$\frac{b}{c} < \frac{p_c^{\text{res}\to\text{res}} - p_c^{\text{mut}\to\text{res}}}{p_c^{\text{res}\to\text{res}} - p_c^{\text{res}\to\text{mut}}}. \quad [8]$$

We calculate the ranges of the b/c ratio for which the resident norm is stable against ALLC and ALLD mutants.

Details of Evolutionary Simulations. To explore how players adopt new social norms over time, we simulate a pairwise comparison process. At each (evolutionary) time step, one player i is randomly selected. With probability ϵ, this player randomly mutates to either ALLC, ALLD, or the focal norm, all with equal probability. With probability $1-\epsilon$, the player selects a random role model j. Then i imitates j's norm with probability ϕ, as given by Eq. **4**. This probability is monotonically increasing in the payoff difference $\pi_j - \pi_i$. Therefore, more successful strategies are more likely to spread. The parameter σ controls the strength of selection. When $\sigma = 0$, we recover neutral drift. Larger values of σ increase the advantage of social norms with a high payoff. Following previous work (27), we use $\sigma = 1$.

We run our simulations in the low-mutation rate limit ($\epsilon \to 0$). As a result, the population is typically monomorphic. Mutations either reach fixation or go extinct, causing transitions between monomorphic populations. The fixation probability of a single mutant with norm $\mathbf{p}$ in a resident population with norm $\mathbf{q}$ can be computed explicitly. It is given by the formula (63):

$$\rho_{\mathbf{p}\to\mathbf{q}} = \frac{1}{1 + \sum_{l'=1}^{N-1} \prod_{l=1}^{l'} \exp\left(-\sigma\left[\pi_{\mathbf{q}}(l) - \pi_{\mathbf{p}}(l)\right]\right)}. \quad [9]$$

Here, $\pi_{\mathbf{p}}(l)$ and $\pi_{\mathbf{q}}(l)$ are the average payoffs of mutants and residents given the population contains l mutants (and hence $N-l$ residents), respectively. To compute the selection-mutation equilibrium (as in Fig. 5), we consider a Markov chain with three possible states, corresponding to the three monomorphic populations (64). The transition probability from monomorphic state $\mathbf{p}$ to state $\mathbf{q} \neq \mathbf{p}$ is $T_{\mathbf{p}\to\mathbf{q}} = \rho_{\mathbf{p}\to\mathbf{q}}/3$. The stationary distribution of this ergodic Markov chain can be obtained by computing the principal eigenvector of the transition matrix.

Data, Materials, and Software Availability. The source code used to numerically verify the results and generate the figures in this paper is available at: https://github.com/yohm/sim_indirect_dual_priv (76).

ACKNOWLEDGMENTS. Y.J.T. acknowledges support by the Program for the Development of Next-Generation Leading Scientists with Global Insight, sponsored by the Ministry of Education, Culture, Sports, Science and Technology, Japan. C.H. acknowledges generous support by the European Research Council Starting Grant 850529: E-DIRECT. Y.M. acknowledges support by Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research (KAKENHI) Grant Number JP25K07145 and from RIKEN Pioneering Project "Planetary Resilience Science for Safeguarding the Global Commons." This research used computational resources of the supercomputer Fugaku provided by the RIKEN Center for Computational Science.

1. K. Sigmund, *The Calculus of Selfishness* (Princeton Univ. Press, Princeton, 2010).
2. M. Broom, J. Rychtář, *Game-Theoretical Models in Biology* (Chapman and Hall/CRC, 2013).
3. M. A. Nowak, Five rules for the evolution of cooperation. *Science* **314**, 1560–1563 (2006).
4. M. Frean, S. Marsland, Score-mediated mutual consent and indirect reciprocity. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2302107120 (2023).
5. D. G. Rand, M. A. Nowak, The evolution of antisocial punishment in optional public goods games. *Nat. Commun.* **2**, 434 (2011).
6. C. Wedekind, M. Milinski, Cooperation through image scoring in humans. *Science* **288**, 850–852 (2000).
7. M. Milinski, D. Semmann, H. J. Krambeck, Reputation helps solve the "tragedy of the commons". *Nature* **415**, 424–426 (2002).
8. D. Engelmann, U. Fischbacher, Indirect reciprocity and strategic reputation building in an experimental helping game. *Games Econ. Behav.* **67**, 399–407 (2009).
9. E. Yoeli, M. Hoffman, D. G. Rand, M. A. Nowak, Powering up with indirect reciprocity in a large-scale field experiment. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 10424–10429 (2013).
10. D. G. Rand, E. Yoeli, M. Hoffman, Harnessing reciprocity to promote cooperation and the provisioning of public goods. *Policy Insights Behav. Brain Sci.* **1**, 263–269 (2014).
11. J. Wu, D. Balliet, P. A. Van Lange, Reputation, gossip, and human cooperation. *Soc. Pers. Psychol. Compass* **10**, 350–364 (2016).
12. A. Bradley, C. Lawrence, E. Ferguson, Does observability affect prosociality? *Proc. R. Soc. B Biol. Sci.* **285**, 20180116 (2018).
13. M. A. Nowak, K. Sigmund, Evolution of indirect reciprocity. *Nature* **437**, 1291–1298 (2005).
14. I. Okada, T. Sasaki, Y. Nakai, A solution for private assessment in indirect reciprocity using solitary observation. *J. Theor. Biol.* **455**, 7–15 (2018).
15. I. Okada, A review of theoretical studies on indirect reciprocity. *Games* **11**, 27 (2020).
16. H. Ohtsuki, Y. Iwasa, How should we define goodness? - Reputation dynamics in indirect reciprocity. *J. Theor. Biol.* **231**, 107–120 (2004).
17. H. Ohtsuki, Y. Iwasa, The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.* **239**, 435–444 (2006).
18. Y. Murase, C. Hilbe, Indirect reciprocity with stochastic and dual reputation updates. *PLoS Comput. Biol.* **19**, e1011271 (2023).
19. N. E. Glynatsi, C. Hilbe, Y. Murase, Exact conditions for evolutionary stability in indirect reciprocity under noise. *PLoS Comput. Biol.* **21**, e1013584 (2025).
20. H. Ohtsuki, Y. Iwasa, M. A. Nowak, Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature* **457**, 79 (2009).
21. Y. Murase, Costly punishment sustains indirect reciprocity under low defection detectability. *J. Theor. Biol.* **600**, 112043 (2025).
22. H. Kim, Y. Murase, Incomplete reputation information and punishment in indirect reciprocity. *Sci. Rep.* **16**, 12773 (2026).
23. S. Tanabe, H. Suzuki, N. Masuda, Indirect reciprocity with trinary reputations. *J. Theor. Biol.* **317**, 338–347 (2013).
24. S. Lee, Y. Murase, S. K. Baek, Local stability of cooperation in a continuous model of indirect reciprocity. *Sci. Rep.* **11**, 14225 (2021).
25. Y. Murase, M. Kim, S. K. Baek, Social norms in indirect reciprocity with ternary reputations. *Sci. Rep.* **12**, 455 (2022).
26. S. Uchida, Effect of private information on indirect reciprocity. *Phys. Rev. E* **82**, 036111 (2010).
27. 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).

28. Y. Fujimoto, H. Ohtsuki, Reputation structure in indirect reciprocity under noisy and private assessment. *Sci. Rep.* **12**, 10500 (2022).
29. K. Oishi, T. Shimada, N. Ito, Group formation through indirect reciprocity. *Phys. Rev. E* **87**, 030801 (2013).
30. Y. Murase, C. Hilbe, Indirect reciprocity under opinion synchronization. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2418364121 (2024).
31. Y. Fujimoto, H. Ohtsuki, Evolutionary stability of cooperation in indirect reciprocity under noisy and private assessment. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2300544120 (2023).
32. Y. Fujimoto, H. Ohtsuki, Who is a leader in the leading eight? Indirect reciprocity under private assessment. *PRX Life* **2**, 023009 (2024).
33. A. L. Radzvilavicius, A. J. Stewart, J. B. Plotkin, Evolution of empathetic moral evaluation. *eLife* **8**, e44269 (2019).
34. M. Seki, M. Nakamaru, A model for gossip-mediated evolution of altruism with various types of false information by speakers and assessment by listeners. *J. Theor. Biol.* **407**, 90–105 (2016).
35. X. Pan, V. Hsiao, D. S. Nau, M. Gelfand, Explaining the evolution of gossip. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2214160121 (2024).
36. M. Kawakatsu, T. A. Kessinger, J. B. Plotkin, A mechanistic model of gossip, reputations, and cooperation. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2400689121 (2024).
37. M. Krellner, T. A. Han, Pleasing enhances indirect reciprocity-based cooperation under private assessment. *Artif. Life* **27**, 246–276 (2022).
38. A. L. Radzvilavicius, T. A. Kessinger, J. B. Plotkin, Adherence to public institutions that foster cooperation. *Nat. Commun.* **12**, 3567 (2021).
39. T. A. Kessinger, J. B. Plotkin, Institutions of public judgment established by social contract and taxation. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2506537122 (2025).
40. U. Berger, A. Grüne, On the stability of cooperation under indirect reciprocity with first-order information. *Games Econ. Behav.* **98**, 19–33 (2016).
41. L. Schmid, F. Ekbatani, C. Hilbe, K. Chatterjee, Quantitative assessment can stabilize indirect reciprocity under imperfect information. *Nat. Commun.* **14**, 2086 (2023).
42. S. Michel-Mata *et al*., The evolution of private reputations in information-abundant landscapes. *Nature* **634**, 883–889 (2024).
43. J. M. Darley, B. Latané, Bystander intervention in emergencies: Diffusion of responsibility. *J. Pers. Soc. Psychol.* **8**, 377 (1968).
44. L. Molleman, F. Kölle, C. Starmer, S. Gächter, People prefer coordinated punishment in cooperative interactions. *Nat. Hum. Behav.* **3**, 1145–1153 (2019).
45. J. J. Jordan, M. Kouchaki, Virtuous victims. *Sci. Adv.* **7**, eabg5902 (2021).
46. M. A. Nowak, K. Sigmund, Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573 (1998).
47. H. Brandt, K. Sigmund, Indirect reciprocity, image scoring, and moral hazard. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 2666–2670 (2005).
48. R. B. Cialdini, N. J. Goldstein, Social influence: Compliance and conformity. *Annu. Rev. Psychol.* **55**, 591–621 (2004).
49. A. Glaubitz, F. Fu, The other side of the coin: Recipient norms and their impact on indirect reciprocity and cooperation. *Appl. Math. Comput.* **513**, 129790 (2026).
50. I. Okada, H. De Silva, Norms prioritizing positive assessments are likely to maintain cooperation in private indirect reciprocity. *Sci. Rep.* **14**, 17264 (2024).
51. L. Schmid, P. Shati, C. Hilbe, K. Chatterjee, The evolution of indirect reciprocity under action and assessment generosity. *Sci. Rep.* **11**, 17443 (2021).
52. Y. Murase, C. Hilbe, Computational evolution of social norms in well-mixed and group-structured populations. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2406885121 (2024).
53. 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).
54. F. A. Chalub, F. C. Santos, J. M. Pacheco, The evolution of norms. *J. Theor. Biol.* **241**, 233–240 (2006).
55. F. P. Santos, F. C. Santos, J. M. Pacheco, Social norms of cooperation in small-scale societies. *PLoS Comput. Biol.* **12**, e1004709 (2016).
56. 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).
57. 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).
58. T. Sasaki, I. Okada, Y. Nakai, The evolution of conditional moral assessment in indirect reciprocity. *Sci. Rep.* **7**, 41870 (2017).
59. T. A. Kessinger, C. E. Tarnita, J. B. Plotkin, Evolution of norms for judging social behavior. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2219480120 (2023).
60. H. Brandt, K. Sigmund, The good, the bad and the discriminator–errors in direct and indirect reciprocity. *J. Theor. Biol.* **239**, 183–194 (2006).
61. A. Traulsen, M. A. Nowak, J. M. Pacheco, Stochastic dynamics of invasion and fixation. *Phys. Rev. E* **74**, 011909 (2006).
62. G. Szabó, C. Tőke, Evolutionary Prisoner's Dilemma game on a square lattice. *Phys. Rev. E* **58**, 69–73 (1998).
63. M. A. Nowak, A. Sasaki, C. Taylor, D. Fudenberg, Emergence of cooperation and evolutionary stability in finite populations. *Nature* **428**, 646–650 (2004).
64. D. Fudenberg, L. A. Imhof, Imitation processes with small mutations. *J. Econ. Theory* **131**, 251–262 (2006).
65. B. Wu, C. S. Gokhale, L. Wang, A. Traulsen, How small are small mutation rates? *J. Math. Biol.* **64**, 803–827 (2012).
66. A. McAvoy, Comment on "Imitation processes with small mutations". *J. Econ. Theory* **159**, 66–69 (2015).
67. D. G. Rand, M. A. Nowak, Human cooperation. *Trends Cogn. Sci.* **117**, 413–425 (2012).
68. D. Premack, G. Woodruff, Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* **1**, 515–526 (1978).
69. O. Leimar, P. Hammerstein, Evolution of cooperation through indirect reciprocity. *Proc. R. Soc. B* **268**, 745–753 (2001).
70. 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).
71. M. Nakamura, N. Masuda, Indirect reciprocity under incomplete observation. *PLoS Comput. Biol.* **7**, e1002113 (2011).
72. N. A. Dalkiran, M. Hoffman, R. Paturi, D. Ricketts, A. Vattani, "Common knowledge and state-dependent equilibria" in *International Symposium on Algorithmic Game Theory*, M. Serna, Ed. (Springer, 2012), pp. 84–95.
73. K. A. Thomas, P. DeScioli, O. S. Haque, S. Pinker, The psychology of coordination and common knowledge. *J. Pers. Soc. Psychol.* **107**, 657 (2014).
74. P. Deutchman, D. Amir, M. R. Jordan, K. McAuliffe, Common knowledge promotes cooperation in the threshold public goods game by reducing uncertainty. *Evol. Hum. Behav.* **43**, 155–167 (2022).
75. J. Henrich, M. Muthukrishna, The origins and psychology of human cooperation. *Annu. Rev. Psychol.* **72**, 207–240 (2021).
76. Y. J. Tham, C. Hilbe, Y. Murase, Source code (2026). GitHub. https://github.com/yohm/sim_indirect_dual_priv. Deposited 2 July 2026.