Indirect reciprocity under opinion synchronization

Yohsuke Murase[a,1] (ID) and Christian Hilbe[b] (ID)

Indirect reciprocity is a key explanation for the exceptional magnitude of cooperation among humans. This literature suggests that a large proportion of human cooperation is driven by social norms and individuals' incentives to maintain a good reputation. This intuition has been formalized with two types of models. In public assessment models, all community members are assumed to agree on each others' reputations; in private assessment models, people may have disagreements. Both types of models aim to understand the interplay of social norms and cooperation. Yet their results can be vastly different. Public assessment models argue that cooperation can evolve easily and that the most effective norms tend to be stern. Private assessment models often find cooperation to be unstable, and successful norms show some leniency. Here, we propose a model that can organize these differing results within a single framework. We show that the stability of cooperation depends on a single quantity: the extent to which individual opinions turn out to be correlated. This correlation is determined by a group's norms and the structure of social interactions. In particular, we prove that no cooperative norm is evolutionarily stable when individual opinions are statistically independent. These results have important implications for our understanding of cooperation, conformity, and polarization.

cooperation | indirect reciprocity | social norms | evolutionary game theory | conformity

Indirect reciprocity can explain why unrelated individuals—even complete strangers— might cooperate with each other (1–3). This explanation suggests that people cooperate because they wish to maintain a positive reputation within their community. There are a number of empirical patterns that align with this view. For example, humans act more prosocially when their actions are widely observable (4–6); they seek information to gauge the social standing of their interaction partners (7, 8); and they are more likely to help those with a positive reputation (9–11).

To better understand these empirical patterns, theoretical studies work with two types of models. The first type, the public assessment model (12–24), assumes that all community members agree on each other's reputations. In particular, if one member thinks highly of some third party, then so does everyone else. Such an assumption may appear as rather extreme. Yet it has been hugely successful, mostly because it drastically simplifies a model's mathematical complexity. Based on this assumption, Ohtsuki and Iwasa (12, 13) were able to identify eight social norms that can stabilize cooperation. These norms, known as the "leading eight," have been widely studied since, even though there are other evolutionarily stable norms that equally support cooperation (24).

The second type, the private assessment model (25–43), recognizes that individuals may differ in how they view others. A mathematical analysis of this type of model is more complex. Private assessment models need to keep track of how each population member thinks of everyone else. The situation can be represented by an "image matrix"; see Fig. 1. Each row of this matrix represents an individual who evaluates the reputations of other group members. Each column represents whose reputation is evaluated. The entries of this matrix correspond to the assigned reputations (in Fig. 1, they are black or white, i.e., "bad" or "good"). This image matrix can change in time, depending on whether individuals cooperate, how observable their actions are, and on the social norm in place. With respect to the observability of individual actions, one can further distinguish three types of models: simultaneous observation models (28–35), solitary observation models (36–38), and models incorporating communication (39–43).

Each of these model types is well established. However, they have typically been studied in isolation. Moreover, the various private assessment and public assessment models often yield conflicting findings. For instance, public assessment models often find that a particular leading-eight norm, "Stern Judging," is most favorable for the evolution of cooperation (14, 16). In contrast, in most private assessment models, the very same social norm proves to be highly inefficient (25, 28). These discrepancies make

Significance

People are more cooperative when their reputation is at stake. This increased tendency to cooperate can be explained with models of indirect reciprocity. However, results in this literature often depend on subtle details. They depend on which social norm is in place, whether interactions are commonly observed, and whether people share their opinions with each other. Here, we show that the differing results can be organized with a simple ordering principle. A given social norm and social environment can only promote cooperation when they lead individuals to have opinions that are sufficiently correlated. These results suggest a selective advantage of conformity. Groups can maintain cooperation most easily when individual opinions tend to be synchronized.

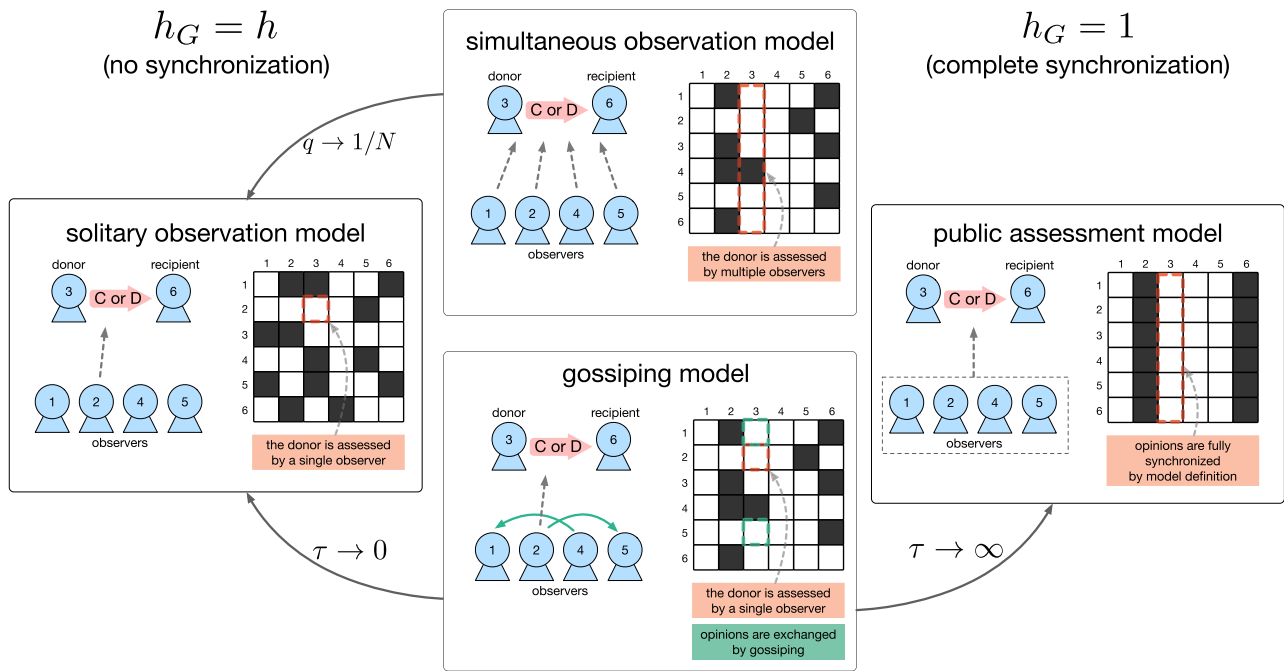https://doi.org/10.1073/pnas.2418364121   **1 of 9**

Fig. 1. Schematic illustration of the models considered in this paper. We consider a population of N players (here, $N = 6$). At each time step, randomly chosen donor-recipient pairs play the donation game. In each game, the donor decides whether to cooperate (C) or to defect (D) according to its action rule P. Other players may observe the donor's decision and update its reputation according to their assessment rule R. The resulting reputations can be represented by an image matrix. It records how each player assesses every other at a given time. Here, we represent the image matrix by a square with six rows and six columns. Black entries indicate that the respective row player thinks negatively of the column player. White entries indicate positive opinions. In the following, we revisit four classical models that differ in how these image matrices are updated. On the *Left*, there is the solitary observation model. Here, each interaction is observed by a single player. As a result, the rows of the image matrix turn out to be independent. In the *Top-Middle*, there is the simultaneous observation model, where each action is observed by any given player with probability q. As the observation probability q approaches $1/N$, the model simplifies to the solitary observation model. In the *Bottom-Middle*, there is the gossiping model by Kawakatsu et al. (39). In this model, players share their opinions with other population members. When the gossip duration τ approaches zero, the model becomes equivalent to the solitary observation model. When $\tau \to \infty$, the model approaches the public assessment model depicted on the *Right*. In that model, opinions are perfectly synchronized.

it difficult to assess whether indirect reciprocity can sustain cooperation at all, and which social norms are most effective.

Here, we propose a general framework to understand the literature through the lens of opinion synchronization. Our framework contains the previous models as special cases, and it systematically reproduces their results. The two key quantities in our model are the variables h and h_G. The first variable h describes how often, on average, individuals assess others as good. The second variable h_G captures to which extent opinions are synchronized. For example, consider three distinct individuals, Alice, Bob, and Charlie. Then h_G corresponds to the conditional probability that Charlie views Bob as good, given that Alice does. The solitary observation model (36–38) corresponds to the case $h_G = h$. Here, opinions are statistically independent. At the other extreme, the public assessment model assumes $h_G = 1$. Here, opinions are perfectly correlated. The simultaneous observation model and models that allow for communication are in between these two extremes. They satisfy $h \leq h_G \leq 1$ (Fig. 1). One important finding is that stable cooperation requires opinions to be sufficiently synchronized. In particular, when opinions are statistically independent, cooperative social norms are either unstable or at most neutrally stable. As opinions become more synchronized, say because of shared experiences or gossip, cooperation can be established more easily.

Our results make an interesting connection to the literature on conformity (44–46). In the context of evolutionary game theory, this literature often argues that conformity may enhance the chance people cooperate, even though cooperators are at a slight disadvantage (47, 48). Our model offers a slightly different

perspective. In models of indirect reciprocity, a certain kind of conformity is what allows cooperation to be in everybody's interest. To this end, we interpret conformity as the degree to which individual opinions (about others) are synchronized. This synchronicity is an emerging trait, which depends on the social norm in place and on the structure of social interactions. In particular, it depends on how publicly observable interactions are, and to which extent individuals exchange gossip. We find that only when opinions are sufficiently synchronized, the mechanism of indirect reciprocity can be effective.

Model

We consider a model of indirect reciprocity that interpolates between public and private assessment models. There is a population of size $N \gg 1$. The members of this population (referred to as "players") engage in a sequence of donation games. Each round, one player is randomly selected to act as a donor. Another player is randomly selected to be the recipient. Donors can either cooperate (C) or defect (D). A cooperating donor pays a cost c to provide a benefit $b > c$ to the recipient. A defecting donor pays no cost and provides no benefit. This elementary process is repeated for many rounds, with changing donors and recipients. Over the course of these games, the players accumulate their payoffs.

During this sequence of the donation games, players form opinions about each other. The opinion player i holds about j is denoted as m_{ij}. The corresponding matrix $M = (m_{ij})$ is referred to as an image matrix. Following the convention of the previous

literature, we assume opinions are binary, either good (*G*) or bad (*B*). These opinions can change over time. Moreover, one player's opinion about another player is not necessarily shared by all other population members. In terms of the image matrix, this means that the entries in any given column do not need to be the same.

A strategy of player *i* is a combination (P_i, R_i) of an action rule and an assessment rule (in line with the indirect reciprocity literature, we use the terms "strategy" and "social norm" synonymously). Herein, we consider stochastic second-order strategies (24). The action rule $P_i(m_{ij})$ determines the cooperation probability of donor *i*, given the donor's opinion m_{ij} about recipient *j*. Let us denote the realized action as $A_{ij} \in \{C, D\}$. An observer *k* assesses the donor *i* based on the donor's action A_{ij} and the observer's opinion about the recipient m_{kj}. The donor is assessed as good with probability $R_k(m_{kj}, A_{ij})$. Otherwise, the donor is assessed as bad. Table 1 gives a few examples of well-known assessment rules. Note that because neither the action rule nor the assessment rule depend on a player's self-image, the diagonal entries m_{ii} of the image matrix are irrelevant.

Assessments can be subject to errors. With probability μ_a, observers assign the opposite reputation to a donor, compared to the assignment prescribed by the assessment rule. As a result, the effective assessment rule becomes

$$\tilde{R}(A, X) = (1 - \mu_a) R(A, X) + \mu_a [1 - R(A, X)]. \qquad [1]$$

In the following, we refer to $\tilde{R}$ as *R* for simplicity. Moreover, we limit ourselves to the case $0 < \mu_a \leq 1/2$ (if the error probability was greater than 1/2, the meaning of "good" and "bad" would be flipped). Because the error probability μ_a is positive, effective assessments are stochastic, $0 < R < 1$. This stochasticity ensures that the reputation dynamics is ergodic. As a result, the time average of the image matrix reaches a unique stationary state that is independent of the initial conditions (for details, see *SI Appendix*).

Based on this general framework, we consider four different models to update the image matrix. First, we consider the solitary observation model (36–38), as depicted on the left hand side of Fig. 1. At each donation game, a single observer is randomly chosen to update their opinion about the donor. Since only a single element of the image matrix is updated each round, the elements in each column are statistically independent. This simplification allows for a fully analytical treatment. Here, we only need to keep track of the fraction *h* of good entries in the image matrix.

Second, we consider the simultaneous observation model, depicted in the *Top* of Fig. 1. This model allows each population member to observe the donor's action with some fixed observation probability *q*. In particular, several observers may witness the same interaction simultaneously. As *q* becomes small (compared to the population size), the model becomes equivalent to the solitary observation model. However, for general *q*, an analytical treatment is more challenging. Except for a few special cases, this model often requires numerical simulations.

Third, we consider the gossiping model by Kawakatsu et al. (39), as depicted in the *Bottom* of Fig. 1. This model allows for communication among the players. Donation games are played between all pairs in the population. Each player updates its opinion about each of the other players based on their action as the donor in a randomly chosen game. Thus, on average, each action is observed by a single observer, as in the solitary observation model. The assessment phase is followed by a gossip phase, where players exchange opinions. This phase consists of

Table 1. Examples of the assessment rules

| Recipient's reputation | G | | B | |
Donor's action	C	D	C	D
Simple standing (L3)	1	0	1	1
Stern judging (L6)	1	0	0	1
Image scoring	1	0	1	0
Shunning	1	0	0	0

We represent four well-known assessment rules. In each case, the probability $R(X, A)$ of assessing the donor as good is shown for each combination of the donor's action *A* and the recipient's reputation *X*. Among these four rules, only the leading-eight strategies L3 and L6 promote evolutionarily stable cooperation in the public assessment model.

several gossiping events. During each event, a randomly chosen pair of players exchange their opinions about a randomly chosen population member. As a result, a randomly chosen entry of the image matrix replaces another randomly chosen entry in the same column. The gossiping events occur repeatedly for a certain number of times, characterized by the gossip duration τ. The parameter τ controls the degree of synchronization of opinions. When $\tau = 0$, the model is equivalent to the solitary observation model. As $\tau \to \infty$, we recover the public assessment model.

The public assessment model is depicted on the *Right* hand side of Fig. 1. This model assumes that all players always have the same opinion about each coplayer. This assumption implies that all entries in any given column of the image matrix are the same. The public assessment model represents the most extreme case of opinion synchronization. Because it allows for a comfortable analytical treatment, it is often used as a benchmark.

Analysis of the Solitary Observation Model

To illustrate our approach, we first analyze the solitary observation model. Here, opinions of different individuals turn out to be statistically independent, which simplifies the analysis. In subsequent sections, we generalize this approach to allow for arbitrary correlations between opinions.

For the solitary observation model, we show that evolutionarily stable cooperation is impossible. More specifically, we show that for any second-order resident strategy, either unconditional cooperation (ALLC) or unconditional defection (ALLD) is always a best response. In the main text, we present an outline of the analysis; all details are in the *SI Appendix*.

We consider a monomorphic resident population in which everyone uses the strategy (P, R). Because of the assumption of solitary observations, the entries of the image matrix are statistically independent. Moreover, due to the ergodicity of the process, the fraction *h* of good entries in the image matrix converges to a unique stationary value after a sufficiently long time. This value satisfies the equation

$$h = h^2 R_P(G, G) + h(1 - h)[R_P(G, B) + R_P(B, G)] \\ + (1 - h)^2 R_P(B, B). \qquad [2]$$

Here, the variable

$$R_P(X, Y) \equiv P(X)R(Y, C) + [1 - P(X)]R(Y, D). \qquad [3]$$

denotes the probability that an observer *k* assesses the donor *i* as good, given their initial opinions about recipient *j*, $m_{ij} = X$ and $m_{kj} = Y$. Eq. **2** provides an implicit formula for the average fraction *h* of well-assessed individuals in a homogeneous resident population.

Next, we consider a small minority of players who deviate toward a different strategy. For brevity, we refer to the deviating players as mutants. To compute whether mutants can invade, we need to compute their payoffs, which in turn depends on how often residents cooperate with a mutant. To do this computation, it turns out to be useful to consider the average cooperation rate of a mutant toward the residents, $p_{\text{mut}\to\text{res}}$. Importantly, here we do not have to define the specific form of the mutant's action and assessment rule. For the following calculations, only the mutants' actions matter, irrespective of how complex their underlying strategies are. Let H denote the average probability that a resident considers a mutant to be good. After a sufficiently long time, H converges to a unique stationary fixed point, defined by

$$H = p_{\text{mut}\to\text{res}} \left[hR(G,C) + (1-h)R(B,C) \right]$$
$$+ (1 - p_{\text{mut}\to\text{res}}) \left[hR(G,D) + (1-h)R(B,D) \right]. \quad [4]$$

In particular, we obtain the following formula for how likely residents are to cooperate with the mutant,

$$p_{\text{res}\to\text{mut}} = HP(G) + (1-H)P(B)$$
$$= p_{\text{mut}\to\text{res}} P_b + P_0. \quad [5]$$

The coefficient P_0 is independent of the mutant's strategy; its exact form is given in *SI Appendix*. The other coefficient P_b can be interpreted as the expected net reward for cooperation,

$$P_b \equiv [P(G) - P(B)] \{ h[R(G,C) - R(G,D)]$$
$$+ (1-h)[R(B,C) - R(B,D)] \}. \quad [6]$$

The factor $[P(G) - P(B)]$ reflects how valuable a good reputation is. The larger this factor, the more likely a good player receives cooperation compared to a bad player. The second factor $\{ h[R(G,C)-R(G,D)] + (1-h)[R(B,C)-R(B,D)] \}$ indicates how much more likely the mutant gets a good reputation by cooperating.

Crucially, Eq. **5** indicates that $p_{\text{res}\to\text{mut}}$ is a linear function of $p_{\text{mut}\to\text{res}}$. In Fig. 2, we illustrate this linear relationship with numerical simulations. For these simulations, we consider four different resident strategies (the same as in Table 1). The resident strategy is adopted by $N-1$ population members. The remaining mutant player either cooperates unconditionally with a certain probability, or adopts a deterministic second-order strategy. Given this population composition, we simulate the game dynamics described in *Model*. Over the course of the simulation, we record how often the mutant cooperates with the residents, and conversely how often residents cooperate with the mutant. As predicted by Eq. **5**, we find a perfect linear relationship between these cooperation rates. This relationship is independent of the complexity of the mutant strategy.

Because cooperation rates obey a linear relationship, also the mutant's payoff can be written as a linear function,

$$\pi_{\text{mut}} = b p_{\text{res}\to\text{mut}} - c p_{\text{mut}\to\text{res}}$$
$$= (bP_b - c) p_{\text{mut}\to\text{res}} + bP_0. \quad [7]$$

This representation of the mutant's payoff is useful because linear functions are comparably easy to analyze. In particular, they typically attain a unique maximum on the boundary of the domain (in this case, for $p_{\text{mut}\to\text{res}} \in \{0, 1\}$). For Eq. 7, we conclude that the mutant maximizes its payoff when

$$p_{\text{mut}\to\text{res}} = \begin{cases} 1 & \text{when} & bP_b > c \\ \text{any} & \text{when} & bP_b = c \\ 0 & \text{when} & bP_b < c \end{cases} \quad [8]$$
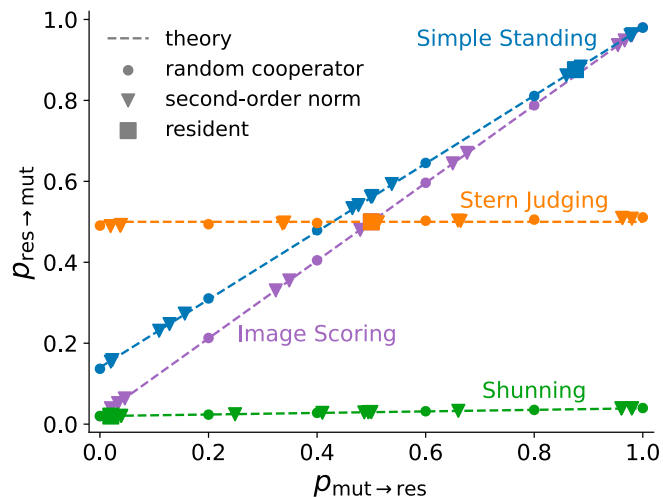

Fig. 2. Relationship of the cooperation levels between residents and mutants. As resident norms, we consider Simple Standing, Stern Judging, Image Scoring, and Shunning, as defined in Table 1. Dashed lines are theoretical predictions obtained from Eq. **5**. Points are obtained from numerical simulations for $N = 100$ and $\mu_a = 0.02$ (see *Materials and Methods, Numerical Simulations* for details). Circles indicate results for unconditionally cooperating mutants with cooperation probability $p_{\text{mut}\to\text{res}} \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, respectively. Triangles are the results for deterministic second-order mutants. We observe that regardless of the mutant strategy, a resident's average cooperation rate toward the mutant, $p_{\text{res}\to\text{mut}}$, is a linear function of the mutant's cooperation rate $p_{\text{mut}\to\text{res}}$.

Thus, a mutant can always maximize its payoff by playing ALLC or ALLD. In contrast, conditional cooperation is generally not optimal. The only exception arises when $bP_b = c$, which corresponds to the Generous Scoring norm (30). But even here, any mutant strategy obtains the same payoff as the residents; thus any mutant can invade by neutral drift. In this sense, any conditionally cooperative strategy is unstable.

For the above result, the assumption of statistically independent opinions is crucial. For an intuition, consider the best action for the donor Alice, toward the recipient Bob, when observed by Charlie. If Alice wants to maximize her long-term payoff, and if opinions are statistically independent, Alice does not need to change her actions depending on her own opinion of Bob. For her, only the opinion of Charlie, who is a potential future interaction partner, matters. It would be best for Alice if she could condition her action based on Charlie's opinion. However, because of the independence assumption, her own view of Bob provides Alice with no information about Charlie's opinion. Conversely, from Charlie's perspective, Alice looks as if she is randomly cooperating with a certain probability, even as Alice changes her actions depending on her own opinion. Thus, if the long-term benefit of cooperation exceeds the immediate cooperation costs, $bP_b > c$, Alice should always cooperate. If the long-term benefit is below this threshold, Alice should always defect. In *SI Appendix*, we derive an analogous conclusion for second-order norms with nonbinary reputations.

The above results assume mutants to be infinitesimally rare. A similar argument, however, also applies when the fraction of mutants is strictly positive: no resident strategy can prevent the neutral invasion by randomly cooperating mutants. To see why, let p_r be the cooperation level of the pure resident population. From a resident's viewpoint, other residents act as if they are randomly cooperating with probability p_r. Thus, a resident cannot distinguish another resident from a mutant who unconditionally cooperates with probability p_r. This neutrality

remains even after the share of mutants in the population increases. Hence, the residents are subject to neutral invasion for any mixed population of residents and mutants, as shown in *SI Appendix,* Fig. S2. This conclusion only relies on the statistical independence of opinions. In particular, the same conclusion applies to more complex strategies, such as strategies with nonbinary opinions (23, 31, 35), higher-order strategies (16), and strategies with dual-reputation updates (24).

The above arguments, however, require that there are at most two strategies in the population. Hence, this result does not rule out the possibility that a mixture of several strategies forms stable cooperation. For instance, previous research suggests that Simple Standing (L3) and ALLC can stably coexist when these strategies compete with ALLD (36–38, 40–42). We do not have a general argument on the stability of such a mixture. It is unknown whether the mixture is stable against a more diverse set of strategies, or invadable by a certain mutant such as random cooperators.

Analysis on Models with Correlated Opinions

In a next step, we generalize the above approach to allow for correlated opinions. In that case, we show there are evolutionarily stable norms that sustain cooperation. Again, in the following, we present an outline of the analysis; *SI Appendix* contains all details.

Consider three players, Alice, Bob, and Charlie, whose opinions may be correlated. The first player, Charlie, considers Bob as good with probability h. Let h_G denote the conditional probability that Alice assigns a good reputation to Bob as well, given that Charlie does. Similarly, let h_B denote the conditional probability that Alice assigns a good reputation to Bob given that Charlie does not. The three probabilities h, h_G, and h_B are not independent of each other. After all, the probability that two randomly chosen players have opposite opinions is equal to both $h(1-h_G)$ and $(1-h)h_B$. Therefore, $h(1-h_G) = (1-h)h_B$ needs to hold. In particular, when $h_G = h$, $h_B = h$ as well. This corresponds to the special case of independent opinions. If Alice's opinion is positively correlated with Charlie's, then $h_G > h > h_B$; Alice is more likely to consider Bob as good when Charlie does. In the perfectly correlated case (i.e., for the public assessment model), $h_G = 1$ and $h_B = 0$.

At this moment, we do not make any further assumptions on how exactly opinions get correlated. For example, Alice and Charlie could have both witnessed Bob's behavior simultaneously (28, 29, 31–34, 49). Alternatively, they could have exchanged opinions by gossiping (39, 40, 50), or obtained relevant information from some public institution (42). Our results are independent of how correlations are achieved.

Consider a monomorphic population in which all players use the norm $\{P, R\}$. As before, we can derive an equation that needs to be satisfied in the stationary state,

$$
\begin{aligned}
h = {} & hh_G R_P(G, G) \\
& + h(1-h_G)[R_P(G, B) + R_P(B, G)] \\
& + (1-h)(1-h_B) R_P(B, B).
\end{aligned}
\tag{9}
$$

In particular, there is only one degree of freedom: given the strategy $\{P, R\}$, h is determined once h_G is fixed (and vice versa). The exact value of h_G depends on the considered model. Once we define a model (such as solitary observation, or the simultaneous observation model), all three quantities h, h_G, h_B are uniquely determined. For the solitary observation model, $h = h_G = h_B$, and Eq. **9** simplifies to Eq. **2**. For other models, it may not be

possible to obtain analytical expressions for h and h_G. In that case, we need simulations; see *Materials and Methods* for details.

Next, we consider a mutant having a different action rule P' but the same assessment rule R. We are going to show that the optimal action rule is either ALLC, ALLD, or conditional cooperation with $P'(G) = 1$ and $P'(B) = 0$. We derive under which conditions conditional cooperation is optimal for given $\{P, R\}$, h, and h_G. To this end, let H be the average probability that a mutant is assessed as good. Again, H converges to a unique stationary value after a sufficiently long time. Since H can be written as a function of h and h_G, the probability that a resident cooperates with a mutant becomes

$$
\begin{aligned}
p_{\text{res}\to\text{mut}} = {} & HP(G) + (1-H)P(B) \\
= {} & [hh_G\Delta_G + h(1-h_G)\Delta_B]P'(G) \\
& + [h(1-h_G)\Delta_G + (1-2h+hh_G)\Delta_B]P'(B) \\
& + P_1.
\end{aligned}
\tag{10}
$$

Here, Δ_G and Δ_B are defined as

$$
\begin{aligned}
\Delta_G &\equiv [R(G, C) - R(G, D)][P(G) - P(B)] \\
\Delta_B &\equiv [R(B, C) - R(B, D)][P(G) - P(B)].
\end{aligned}
\tag{11}
$$

Moreover, P_1 is a constant term that does not depend on P'. Similarly, the probability that a mutant cooperates with a resident is

$$
p_{\text{mut}\to\text{res}} = hP'(G) + (1-h)P'(B).
\tag{12}
$$

Therefore, the mutant's payoff is

$$
\begin{aligned}
\pi_{\text{mut}} &= bp_{\text{res}\to\text{mut}} - cp_{\text{mut}\to\text{res}} \\
&= \alpha_G P'(G) + \alpha_B P'(B) + bP_1,
\end{aligned}
\tag{13}
$$

where we defined

$$
\begin{aligned}
\alpha_G &\equiv hh_G(b\Delta_G - c) + h(1-h_G)(b\Delta_B - c) \\
\alpha_B &\equiv h(1-h_G)(b\Delta_G - c) + (1-2h+hh_G)(b\Delta_B - c).
\end{aligned}
\tag{14}
$$

Since π_{mut} is a linear function of $P'(G)$ and $P'(B)$, the best action rule $\hat{P}'$ for the mutant is summarized as follows:

$$
\hat{P}'(G) = \begin{cases} 1 & \text{when } \alpha_G > 0 \\ \text{any} & \text{when } \alpha_G = 0 \\ 0 & \text{when } \alpha_G < 0 \end{cases}
$$
$$
\hat{P}'(B) = \begin{cases} 1 & \text{when } \alpha_B > 0 \\ \text{any} & \text{when } \alpha_B = 0 \\ 0 & \text{when } \alpha_B < 0 \end{cases}.
\tag{15}
$$

From this equation, we conclude that the best action rule $\hat{P}'$ needs to be deterministic, except for the special cases of $\alpha_G = 0$ or $\alpha_B = 0$. When α_G and α_B have the same sign, unconditional cooperation or defection is the best action. Conditional cooperation is optimal when $\alpha_G > 0$ and $\alpha_B < 0$. In that case, the average cooperation rate of the population coincides with the fraction h of individuals with a good reputation. Note that when $h = h_G$, α_G and α_B have the same sign, reproducing the conclusion of the previous section.

In Fig. 3, we illustrate these results for the norms Stern Judging (L6) and Simple Standing (L3). The *Top* panels, Fig. 3 *A* and *B*, show the functional relationship between h and h_G. In each case, h increases as h_G increases: the cooperation level rises as opinions are
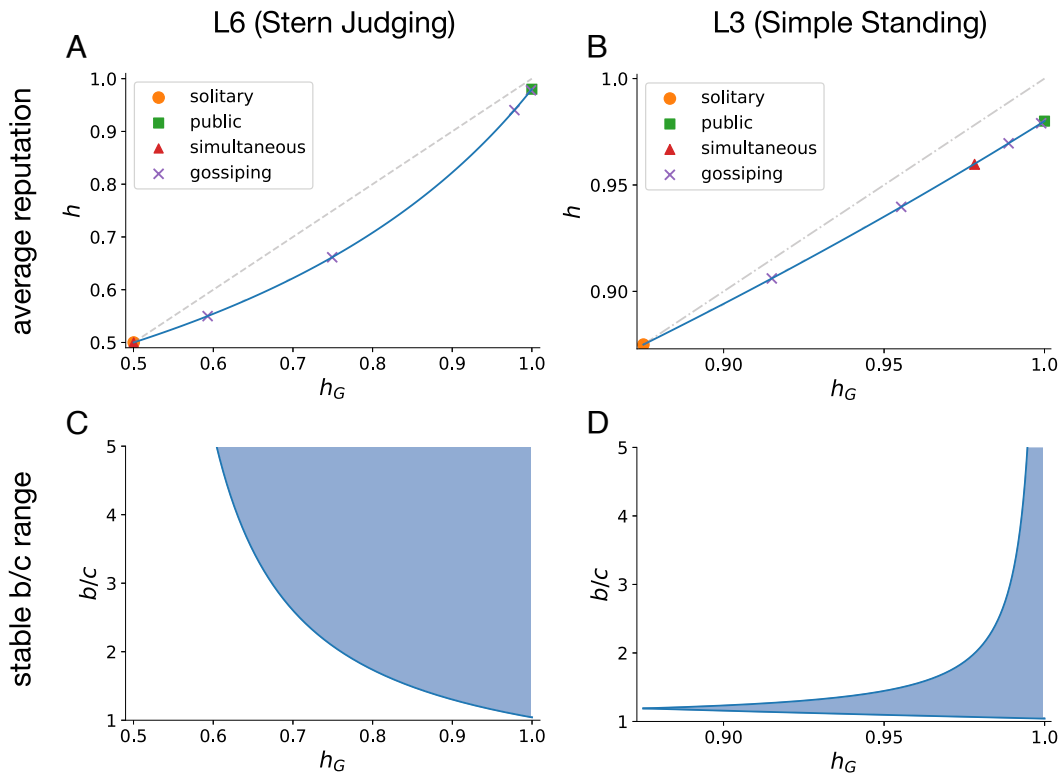
Fig. 3. (*A* and *B*) The relationship between h and h_G for the norms L6 (Stern Judging) and L3 (Simple Standing). The gray dashed lines indicate the case $h = h_G$, which is shown as a reference. For both norms, h increases as h_G increases. The leftmost h_G is the minimal value obtained for the solitary observation model; here, $h_G = h$. The rightmost h_G corresponds to the public assessment model $h_G = 1$. Results for the simultaneous observation model with $N = 100$ and $q = 1$ are shown as red triangles. Purple crosses indicate the gossiping model with $\tau = 0.1, 0.3, 1,$ and 3, from left to right. (*C* and *D*) The blue area indicates the benefit-to-cost ratios b/c for which a conditionally cooperative action rule is stable. As h_G increases, the stable range of b/c expands. We use an assessment error rate of $\mu_a = 0.02$.

more synchronized (such a positive relationship does not need to hold for other norms). In addition, we depict the realized values of h and h_G for each of the different model types considered. The most extreme models are the solitary observation model (to the *Left*) and the public assessment model (to the *Right*), respectively. The *Bottom* panels, Fig. 3 *C* and *D*, show the ranges of the benefit-to-cost ratio b/c for which conditional cooperation is stable, calculated from Eq. **15**. The stable ranges of b/c expand with h_G, indicating that cooperation is easier to sustain when opinions are more synchronized. By superimposing the *Upper* and the *Lower* panels, we can also infer for which social structure each of the two norms is stable. For example, for Stern Judging, the *Upper* panel suggests that in the simultaneous observation model, we obtain $h_G = 0.5$. For this value of h_G, the *Lower* panel suggests that Stern Judging is unstable, for any benefit-to-cost ratio, in line with previous research (26). In contrast, for Simple Standing, simultaneous observations lead to $h_G \approx 0.98$, which permits stable cooperation if b/c is sufficiently small.

Application to Specific Models

In the following, we apply this general formalism to three special cases: the public assessment model, the simultaneous observation model, and the gossiping model. In this way, we show that we are able to systematically reproduce a large set of previous results within a single framework.

Public Assessment Model. First, we consider public assessment. Here, opinions are perfectly synchronized, $h_G = 1$ and $h_B = 0$. For this model, the set of stable norms that sustain cooperation

has been fully characterized in ref. 24. According to Eq. **15**, conditional cooperation with $P(G) = 1$ and $P(B) = 0$ is optimal when

$$\begin{cases} \alpha_G > 0 & \iff b\Delta_G > c \\ \alpha_B < 0 & \iff b\Delta_B < c \end{cases}. \qquad [16]$$

A conditionally cooperative response is the unique best action rule when

$$\begin{cases} b\left[R(G, C) - R(G, D)\right] > c \\ b\left[R(B, C) - R(B, D)\right] < c \end{cases}, \qquad [17]$$

These results perfectly align with the previous characterization of stable second-order norms in ref. 24.

Simultaneous Observation Model. Next, we consider a model where multiple observers may assess the donor simultaneously. To this end, we build on the work of Fujimoto and Ohtsuki (32–34). They approximate the distribution of the goodness in the stationary state for $q = 1$. For example, when errors are rare and the population is large, $\mu_a \ll 1$ and $N \to \infty$, the distribution of goodness under the L3 norm can be written as the sum of delta functions (see figure 3D in ref. 33):

$$p(g) = (1 - 2\mu_a)\delta\left(g - (1 - \mu_a)\right) \\ + \mu_a\delta\left(g - (1 - 3\mu_a)\right) + \mu_a\delta\left(g - 2\mu_a\right) + O(\mu_a^2). \qquad [18]$$

From Eq. **18**, h and h_G are derived as

$$h = 1 - 2\mu_a + O(\mu_a^2), \quad h_G = 1 - \mu_a + O(\mu_a^2). \qquad [19]$$

It then follows from Eqs. **15** and **19** that conditional cooperation is stable if and only if both of the following conditions are met,

$$\begin{cases} \alpha_G > 0 & \Longleftrightarrow \quad \frac{b}{c} > 1 + O(\mu_a) \\ \alpha_B < 0 & \Longleftrightarrow \quad \frac{b}{c} < 2 + O(\mu_a). \end{cases} \quad \text{[20]}$$

Again, these conditions reproduce the results by Fujimoto and Ohtsuki (33, 34). In addition, we consider other general cases in *SI Appendix*. In particular, we consider strategies that interpolate between L3 and L6, and results for $q < 1$. For these cases, we rely on numerical simulations to calculate h and h_G. Once h and h_G are obtained, the stable range of b/c is derived using Eq. **15**. The resulting theoretical predictions again agree with the results of numerical simulations.

Gossiping Model. Another application of the above theory is the gossiping model by Kawakatsu et al. (39). In their model, the gossip duration τ quantifies the amount of peer-to-peer gossip between private observation periods. Kawakatsu et al. derive an analytic relationship of h, h_G and τ as follows:

$$1 - h_G = (1 - h)e^{-\tau}. \quad \text{[21]}$$

When $\tau \to 0$, the gossiping process is equivalent to the solitary observation model. When $\tau \to \infty$, it reproduces the public assessment model.

For an illustration, we consider the assessment rule L6 (for more details, see *SI Appendix*). Kawakatsu et al. (39) consider the replicator dynamics when L6 competes with both ALLC and ALLD. They show that a pure L6 population can only be stable when $b/c > (b/c)^* \equiv 1/(1 - 2\mu_a)$. If that condition is satisfied, they obtain the critical gossip duration τ^* above which a pure L6 population is stable,

$$\tau^* = \log\left[\left(2 - \frac{\frac{b}{c}}{\frac{b}{c} - \frac{1}{2(1-\mu_a)}}\right)\left(\frac{\frac{b}{c}}{\frac{b}{c} - \frac{1}{1-2\mu_a}}\right)\right]. \quad \text{[22]}$$

The same conclusion can be derived from our framework. From Eqs. **9** and **21**, h and h_G are uniquely determined. The conditionally cooperative action rule is stable when $\alpha_G > 0$ and $\alpha_B < 0$, as defined in Eq. **14**. By analyzing these conditions, we obtain the same critical benefit-to-cost ratio, and the same critical gossip duration. While the previous study considers only ALLC and ALLD as possible evolutionary competitors, we conclude that the conditionally cooperative strategy is stable against mutants with any action rule, including stochastic ones.

Discussion

In this paper, we propose a general framework to analyze the evolutionary stability of indirect reciprocity. The literature on indirect reciprocity is vast, and researchers have established several distinct types of models (12–43). Unfortunately, the different model types are often studied in isolation, and they sometimes lead to conflicting results. Here, we show that all this previous work can be organized by considering a single key quantity: the degree to which individual opinions are correlated. This correlation in turn depends on the social norm in place, on the observability of interactions, and on the degree to which individuals share their views. As a rule of thumb, we find that the more opinions are correlated, the easier it becomes to sustain cooperation. Conversely, if opinions turn out to be completely uncorrelated, cooperative norms become evolutionarily unstable. Some previous work has already hinted at the negative effects of disagreements on cooperation (e.g., refs. 25–28). Yet there has been little work to quantify these effects. Our study highlights the role of opinion synchronization in a mathematically explicit manner.

Within our framework, one extreme case is the public assessment model, where opinions are perfectly synchronized. The other extreme is the solitary observation model, where opinions are statistically independent. Although these two extreme cases may be strong idealizations, they serve as useful benchmarks due to their analytical tractability. Between these extremes, there are several models in which opinions are correlated, but incompletely so. These incomplete correlations can arise, for example, when several individuals tend to witness the same event, as in the simultaneous observation model (28–35). Alternatively, they can arise when individuals use gossip to partly synchronize their views, or at least the information they have (39). These intermediate models are more realistic, but they render analytical solutions more difficult to obtain. Our results indicate that we do not need to understand the respective image matrices in full detail. Instead, we only need to know the two quantities h and h_G (the first two moments of the goodness distribution). They provide all the information needed to characterize whether a given social norm can sustain cooperation. Future theoretical studies on opinion synchronization, such as those in refs. 39 and 49, can provide a more detailed understanding of how the values of h and h_G depend on the exact social setup in which interactions take place.

Herein, we show that some degree of opinion synchronization is crucial to maintain cooperative relationships. This may have implications for our understanding of typical group sizes in human societies, such as those indicated by Dunbar's number (51, 52). In smaller populations, opinions tend to synchronize more readily, thereby facilitating cooperation. In contrast, as a group increases, opinion synchronization becomes progressively more challenging. This difficulty in aligning opinions in larger groups ultimately sets a limit on the group size within which cooperation can be maintained. Our research implies the importance of group size in stabilizing cooperative interactions via indirect reciprocity.

Our findings on the importance of opinion synchronization also resonate with previous work on effective punishment after norm violations (53, 54). This literature studies under which conditions group members are able to sanction certain offenses in the presence of uncertainty. The model of Dalkiran et al (53) suggests that they can only do so if it becomes common knowledge (or more precisely, "common p-belief") that a norm violation occurred in the first place. Their result follows a similar logic as in our model. Once individuals disagree on whether or not a norm violation occurred, it becomes too costly for any single individual to take action.

Without a doubt, our study has several limitations. While we have shown that opinion synchronization is crucial for evolutionarily stable indirect reciprocity, this does not mean that opinion synchronization is sufficient. In particular, we have only considered mutants who deviate from a population's norm by choosing different actions. Instead, mutants may also differ in how they assess other people's actions. For instance, a mutant with an assessment rule that leads to a higher level of synchronization might be able to invade. Furthermore, while

opinion synchronization does help stabilize cooperation, the lack of evolutionarily stable strategies in the solitary observation model does not rule out cooperation entirely. Even if opinions are uncorrelated, indirect reciprocity might evolve if additional mechanisms for cooperation are in place, such as group structure (55). Another valuable direction is to explore models with continuous degrees of cooperation (35, 56) and models with an explicit punishment option (57–59).

Finally, an important open question is how mechanisms for opinion synchronization coevolve with social norms. We studied models with simultaneous observation and gossiping, but other possibilities exist. These mechanisms could have coexisted or evolved in a specific sequence. Furthermore, these actions to promote synchronization may themselves be costly, which could again affect the stability of cooperation. Investigating the evolution of these mechanisms is another promising direction for future research on indirect reciprocity.

Materials and Methods

Numerical Simulations. We conducted Monte Carlo simulations to validate the theoretical predictions. We consider a population of N players. The reputation state is represented by an image matrix of size $N \times N$. At each time step, a randomly chosen player i is selected as the donor, and a randomly chosen player j is selected as the recipient. The donor i decides its action based on its action rule. Then, the reputation of the donor (i'th column of the image matrix) is updated. How the reputation is updated depends on the assessment rule and the model. More details and pseudocodes for each of these models are described in *SI Appendix*.

We first conducted t_{init} steps to equilibrate the image matrix and then ran t_{measure} steps to measure the quantities. The values of h and h_G are calculated by measuring the average and the variance of the goodness of the players (32, 33).

Here, we define the goodness of player-i, g_i, as the fraction of the good image in the i-th column of the image matrix excluding the diagonal element:

$$g_i = \sum_{j \neq i} \delta(m_{ji}, G)/(N-1), \qquad \textbf{[23]}$$

where $\delta(x, y)$ is the Kronecker delta function. The average goodness taken over i equals h: $h = \langle g_i \rangle$. The product hh_G is the expected probability that two randomly chosen players agree that another randomly chosen player is G. Thus, for a finite N,

$$hh_G = \left\langle g_i \frac{(N-1)\, g_i - 1}{N-2} \right\rangle = \frac{N-1}{N-2} \left\langle g_i^2 \right\rangle - \frac{\langle g_i \rangle}{N-2}$$

$$h_G = \frac{1}{N-2} \left[(N-1) \frac{\langle g_i^2 \rangle}{\langle g_i \rangle} - 1 \right]. \qquad \textbf{[24]}$$

When $N \to \infty$, $h_G \to \langle g_i^2 \rangle / \langle g_i \rangle$.

The parameters used in this study are $N = 100$, $\mu_a = 0.02$, $t_{\text{init}} = 10^5$, and $t_{\text{measure}} = 10^6$. The source code used in this study is available in Github (60).

Data, Materials, and Software Availability. Source code data have been deposited in GitHub (https://github.com/yohm/sim_indirect_opinion_sync) (60). All other data are included in the manuscript and/or *SI Appendix*.

1. A. P. Melis, D. Semmann, How is human cooperation different? *Philos. Trans. R. Soc. B* **365**, 2663–2674 (2010).
2. D. G. Rand, M. A. Nowak, Human cooperation. *Trends Cogn. Sci.* **117**, 413–425 (2012).
3. M. A. Nowak, Five rules for the evolution of cooperation. *Science* **314**, 1560–1563 (2006).
4. D. Semmann, H. J. Krambeck, M. Milinski, Strategic investment in reputation. *Behav. Ecol. Sociobiol.* **56**, 5 (2004).
5. D. Engelmann, U. Fischbacher, Indirect reciprocity and strategic reputation building in an experimental helping game. *Games Econ. Behav.* **67**, 399–407 (2009).
6. 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).
7. V. Swakman, L. Molleman, A. Ule, M. Egas, Reputation-based cooperation: Empirical evidence for behavioral strategies. *Evol. Hum. Behav.* **37**, 230–235 (2016).
8. L. Molleman, E. van den Broek, M. Egas, Personal experience and reputation interact in human decisions to help reciprocally. *Proc. R. Soc. B* **28**, 20123044 (2013).
9. C. Wedekind, M. Milinski, Cooperation through image scoring in humans. *Science* **288**, 850–852 (2000).
10. C. Wedekind, V. A. Braithwaite, The long-term benefits of human generosity in indirect reciprocity. *Curr. Biol.* **12**, 1012–1015 (2002).
11. M. Milinski, D. Semmann, H. J. Krambeck, Reputation helps solve the "tragedy of the commons". *Nature* **415**, 424–426 (2002).
12. H. Ohtsuki, Y. Iwasa, How should we define goodness? Reputation dynamics in indirect reciprocity. *J. Theor. Biol.* **231**, 107–120 (2004).
13. H. Ohtsuki, Y. Iwasa, The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.* **239**, 435–444 (2006).
14. 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).
15. F. P. Santos, F. C. Santos, J. M. Pacheco, Social norms of cooperation in small-scale societies. *PLoS Comput. Biol.* **12**, e1004709 (2016).
16. 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).
17. 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).
18. N. Masuda, Ingroup favoritism and intergroup cooperation under indirect reciprocity based on group reputation. *J. Theor. Biol.* **311**, 8–18 (2012).
19. U. Berger, Learning to cooperate via indirect reciprocity. *Games Econ. Behav.* **72**, 30–37 (2011).
20. M. Nakamura, H. Ohtsuki, Indirect reciprocity in three types of social dilemmas. *J. Theor. Biol.* **355**, 117–127 (2014).
21. H. Ohtsuki, Y. Iwasa, M. A. Nowak, Reputation effects in public and private interactions. *PLoS Comput. Biol.* **11**, e1004527 (2015).
22. D. Clark, D. Fudenberg, A. Wolitzky, Indirect reciprocity with simple records. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 11344–11349 (2020).
23. Y. Murase, M. Kim, S. K. Baek, Social norms in indirect reciprocity with ternary reputations. *Sci. Rep.* **12**, 455 (2022).
24. Y. Murase, C. Hilbe, Indirect reciprocity with stochastic and dual reputation updates. *PLoS Comput. Biol.* **19**, e1011271 (2023).
25. S. Uchida, Effect of private information on indirect reciprocity. *Phys. Rev. E* **82**, 036111 (2010).
26. S. Uchida, T. Sasaki, Effect of assessment error and private information on stern-judging in indirect reciprocity. *Chaos Solitons Fractals* **56**, 175–180 (2013).
27. C. Perret, M. Krellner, T. A. Han, The evolution of moral rules in a model of indirect reciprocity with private assessment. *Sci. Rep.* **11**, 23581 (2021).
28. 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).
29. L. Schmid, P. Shati, C. Hilbe, K. Chatterjee, The evolution of indirect reciprocity under action and assessment generosity. *Sci. Rep.* **11**, 17443 (2021).
30. L. Schmid, K. Chatterjee, C. Hilbe, M. A. Nowak, A unified framework of direct and indirect reciprocity. *Nat. Hum. Behav.* **5**, 1292 (2021).
31. L. Schmid, F. Ekbatani, C. Hilbe, K. Chatterjee, Quantitative assessment can stabilize indirect reciprocity under imperfect information. *Nat. Commun.* **14**, 2086 (2023).
32. Y. Fujimoto, H. Ohtsuki, Reputation structure in indirect reciprocity under noisy and private assessment. *Sci. Rep.* **12**, 10500 (2022).
33. 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).
34. Y. Fujimoto, H. Ohtsuki, Who is a leader in the leading eight? Indirect reciprocity under private assessment. *PRX Life* **2**, 023009 (2024).
35. S. Lee, Y. Murase, S. K. Baek, Local stability of cooperation in a continuous model of indirect reciprocity. *Sci. Rep.* **11**, 14225 (2021).
36. 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).
37. I. Okada, T. Sasaki, Y. Nakai, A solution for private assessment in indirect reciprocity using solitary observation. *J. Theor. Biol.* **455**, 7–15 (2018).
38. I. Okada, Two ways to overcome the three social dilemmas of indirect reciprocity. *Sci. Rep.* **10**, 1–9 (2020).
39. 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).
40. 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).
41. A. L. Radzvilavicius, A. J. Stewart, J. B. Plotkin, Evolution of empathetic moral evaluation. *eLife* **8**, e44269 (2019).
42. A. L. Radzvilavicius, T. A. Kessinger, J. B. Plotkin, Adherence to public institutions that foster cooperation. *Nat. Commun.* **12**, 3567 (2021).

43. M. Krellner, T. A. Han, Pleasing enhances indirect reciprocity-based cooperation under private assessment. *Artif. Life* **27**, 246–276 (2022).

44. R. B. Cialdini, N. J. Goldstein, Social influence: Compliance and conformity. *Annu. Rev. Psychol.* **55**, 591–621 (2004).

45. K. K. Denton, Y. Ram, U. Liberman, M. W. Feldman, Cultural evolution of conformity and anticonformity. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 13603-13614 (2020).

46. K. K. Denton, U. Liberman, M. W. Feldman, Polychotomous traits and evolution under conformity. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2205914119 (2022).

47. J. Henrich, R. Boyd, Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *J. Theor. Biol.* **208**, 79–89 (2001).

48. A. Romano, D. Balliet, Reciprocity outperforms conformity to promote cooperation. *Psychol. Sci.* **28**, 1490-1502 (2017).

49. M. Krellner, T. A. Han, We both think you did wrong–how agreement shapes and is shaped by indirect reciprocity. arXiv [Preprint] (2023). https://arxiv.org/abs/2304.14826 (Accessed 16 September 2024).

50. X. Pan, V. Hsiao, D. S. Nau, M. Gelfand, Explaining the evolution of gossip. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2214160121 (2024).

51. R. I. Dunbar, Neocortex size as a constraint on group size in primates. *J. Hum. Evol.* **22**, 469–493 (1992).

52. R. I. Dunbar, The social brain hypothesis and its implications for social evolution. *Ann. Hum. Biol.* **36**, 562–572 (2009).

53. N. A. Dalkiran, M. Hoffman, R. Paturi, D. Ricketts, A. Vattani, "Common knowledge and state-dependent equilibria" in *Lecture Notes in Computer Science*, M. Serna, Ed. (Springer, Berlin/Heidelberg, Germany, 2012), vol. 7615.

54. M. Hoffman, E. Yoeli, *Hidden Games* (Basic Books, 2022).

55. 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).

56. 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).

57. H. Ohtsuki, Y. Iwasa, M. A. Nowak, Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature* **457**, 79 (2009).

58. A. Schlaepfer, The emergence and stability of reputation systems that drive cooperative behaviour. *Proc. R. Soc. B* **285**, 20181508 (2018).

59. Y. Murase, Costly punishment sustains indirect reciprocity under low defection detectability. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2409.09701 (Accessed 16 September 2024).

60. Y. Murase, Source code (2024). Github. https://github.com/yohm/sim_indirect_opinion_sync. Deposited 03 September 2024.