Supporting Information for

A mechanistic model of gossip, reputations, and cooperation

Mari Kawakatsu*, Taylor A. Kessinger*, Joshua B. Plotkin

Correspondence to: marikawa@sas.upenn.edu (M.K.), tkess@sas.upenn.edu (T.A.K.), jplotkin@sas.upenn.edu (J.B.P.)

This PDF file includes:

Supporting Information Text
Figs. S1 to S8
Tables S1 to S5
SI References

Supporting Information Text

Contents

1. Relationship between single-source gossip and empathetic perspective taking

Here we show the mathematical relationship between the model of gossip with a single source (Fig. 1C; see also *A Model of Gossip, Reputations, and Social Behavior* in the main text) and the model of empathetic moral evaluation (1).

The equilibrium of the reputation ODEs, $\frac{dr_s}{dt} = p_s(t) - r_s(t)$ (Eq. 2 in the main text), satisfies $r_s = p_s$. In particular, the equilibrium reputation of discriminators is

$$r_{\text{DISC}} = p_{\text{DISC}} = \tilde{g}_2 P_{GC} + \tilde{d}_2 \left(P_{BC} + P_{GD}\right) + \tilde{b}_2 P_{BD} \ . \tag{S1}$$

Recall from Eq. 5 that the rates of agreement and disagreement after a period of gossip are given by

$$\begin{aligned}
\tilde{g}_2 &= \left(1 - Q^2\right) \cdot g_2 + Q^2 \cdot r \ , \\
\tilde{b}_2 &= \left(1 - Q^2\right) \cdot b_2 + Q^2 \cdot (1 - r) \ , \\
\tilde{d}_2 &= \left(1 - Q^2\right) \cdot d_2 \ .
\end{aligned}$$

Substituting these expressions into Eq. S1, we obtain

$$\begin{aligned}
r_{\text{DISC}} &= \left[Q^2 r + \left(1 - Q^2\right) g_2\right] P_{GC} + \left[\left(1 - Q^2\right) d_2\right] \left(P_{BC} + P_{GD}\right) + \left[Q^2 (1 - r) + \left(1 - Q^2\right) b_2\right] P_{BD} \\
&= Q^2 \left[r P_{GC} + (1 - r) P_{BD}\right] + \left(1 - Q^2\right) \left[g_2 P_{GC} + d_2 \left(P_{BC} + P_{GD}\right) + b_2 P_{BD}\right] \ .
\end{aligned}$$

This expression is equivalent in form to Eq. 5 in Radzvilavicius et al. (1), but with $Q^2 = E$, where E is the degree of empathy, i.e., the probability that an observer uses the donor's view of the recipient's reputation when updating the donor's reputation. This relationship means that consulting a single gossip source is like a slower form of empathetic perspective-taking: in the former, two individuals will agree with full certainty only if they have both consulted the shared information source, while in the latter, a single individual can guarantee agreement with another by unilaterally adopting her view.

2. Stability of cooperation under unbiased peer-to-peer gossip

2.1. Impact of assessment and execution errors on the critical gossip duration under the Stern Judging norm. To determine how errors in assessment or execution impact the amount of gossip needed to stabilize cooperation under Stern Judging (condition (ii) in main text), we evaluate the partial derivatives of the critical gossip duration τ^* with respect to the assessment error rate u_a and the execution error rate u_e:

$$\begin{aligned}
\frac{\partial \tau^*}{\partial u_a} &= \frac{2}{(1 - 2u_a)\left(\frac{(b/c)}{(b/c)^*} - 1\right)} + \frac{\frac{(b/c)}{(b/c)^*}}{\left(\frac{1}{(b/c)^*} - (1 - u_a)\frac{(b/c)}{(b/c)^*}\right)\left(2(1 - u_a)\frac{(b/c)}{(b/c)^*} - \frac{1}{(b/c)^*}\right)} \ , \\
\frac{\partial \tau^*}{\partial u_e} &= \frac{1}{(1 - u_e)\left(\frac{(b/c)}{(b/c)^*} - 1\right)} \ ,
\end{aligned}$$

where $(b/c)^* = \frac{1}{(1-2u_a)(1-u_e)}$ as in condition (i) in main text. Both derivatives are positive whenever condition (i) is satisfied, i.e., $b/c > (b/c)^*$. Hence, under Stern Judging, the critical gossip duration τ^* increases monotonically with u_a and with u_e. We confirm this numerically in Fig. S3.

2.2. Equilibrium reputations at the all-DISC equilibrium under any norm. To derive conditions for the stability of cooperation, we begin by computing the reputation equilibrium in a population of discriminators. To do so, we set the right-hand sides of the reputation ODEs (Eq. 2) to zero and solve for r_{ALLC}, r_{ALLD}, and r_{DISC} at $f_{\text{DISC}} = 1$. More explicitly, the reputation equilibrium at the all-DISC equilibrium satisfies

$$\begin{aligned}
r_{\text{ALLC}} &= r_{\text{DISC}} P_{GC} + (1 - r_{\text{DISC}}) P_{BC} \ , \\
r_{\text{ALLD}} &= r_{\text{DISC}} P_{GD} + (1 - r_{\text{DISC}}) P_{BD} \ , \\
r_{\text{DISC}} &= \left(r_{\text{DISC}}^2 + r_{\text{DISC}} (1 - r_{\text{DISC}}) \cdot \left(1 - e^{-\tau}\right)\right) P_{GC} + \left(r_{\text{DISC}} (1 - r_{\text{DISC}}) \cdot e^{-\tau}\right) \left(P_{BC} + P_{GD}\right) \\
&\qquad + \left((1 - r_{\text{DISC}})^2 + r_{\text{DISC}} (1 - r_{\text{DISC}}) \cdot \left(1 - e^{-\tau}\right)\right) P_{BD} \\
&= \left[r_{\text{DISC}} \cdot P_{GC} + (1 - r_{\text{DISC}}) \cdot P_{BD}\right] - e^{-\tau} \left[r_{\text{DISC}} (1 - r_{\text{DISC}}) \cdot (P_{GC} - P_{BC} - P_{GD} + P_{BD})\right] \ .
\end{aligned} \tag{S2}$$

These expressions are obtained from Eq. 3 by setting $f_{\text{DISC}} = 1$, substituting in the agreement and disagreement rates evaluated at $f_{\text{DISC}} = 1$ (Eq. 4), and letting $p_s = r_s$ (Eq. 2).

Solving for r_{DISC} (with $0 \leq r_{\text{DISC}} \leq 1$), we obtain the equilibrium reputation of discriminators at the discriminator-only equilibrium, given by

$$r_{\text{DISC}} = \frac{1}{2}\left(1 + \frac{e^{\tau}(1 - P_{GC} + P_{BD})}{P_{GC} - P_{BC} - P_{GD} + P_{BD}} - \sqrt{\left(1 + \frac{e^{\tau}(1 - P_{GC} + P_{BD})}{P_{GC} - P_{BC} - P_{GD} + P_{BD}}\right)^2 - \frac{e^{\tau} \cdot 4P_{BD}}{P_{GC} - P_{BC} - P_{GD} + P_{BD}}}\right) \ .$$

We then obtain the equilibrium r_{ALLC} and r_{ALLD} by substituting the expression for r_{DISC} into the first two equations in Eq. S2. The explicit expression for the equilibrium value of r_{DISC} for a norm parametrized by (p, q) (*Materials and Methods*) can be obtained using Eq. 12.

2.2.1. Equilibrium reputations at the all-DISC equilibrium under the Stern Judging norm. The Stern Judging norm is given by $(p, q) = (0, 1)$, so we have $1 - P_{GC} + P_{BD} = 1 + (1 - 2u_a)\, u_e$ and $P_{GC} - P_{BC} - P_{GD} + P_{BD} = 2\,(1 - 2u_a)\,(1 - u_e)$. The equilibrium reputation of DISC at the all-DISC equilibrium is then

$$r_{\text{DISC}} = \frac{1}{2}\left(1 + \frac{e^{\tau}(1 + (1 - 2u_a)\, u_e)}{2\,(1 - 2u_a)\,(1 - u_e)} - \sqrt{\left(1 + \frac{e^{\tau}(1 + (1 - 2u_a)\, u_e)}{2\,(1 - 2u_a)\,(1 - u_e)}\right)^2 - \frac{e^{\tau} \cdot 4\,(1 - u_a)}{2\,(1 - 2u_a)\,(1 - u_e)}}\right) .$$

2.2.2. Equilibrium reputations at the all-DISC equilibrium under the Simple Standing norm. The Simple Standing norm is given by $(p, q) = (1, 1)$, so we have $1 - P_{GC} + P_{BD} = 1 + (1 - 2u_a)\, u_e$ and $P_{GC} - P_{BC} - P_{GD} + P_{BD} = (1 - 2u_a)\,(1 - u_e)$. The equilibrium reputation of DISC at the all-DISC equilibrium is then

$$r_{\text{DISC}} = \frac{1}{2}\left(1 + \frac{e^{\tau}(1 + (1 - 2u_a)\, u_e)}{(1 - 2u_a)\,(1 - u_e)} - \sqrt{\left(1 + \frac{e^{\tau}(1 + (1 - 2u_a)\, u_e)}{(1 - 2u_a)\,(1 - u_e)}\right)^2 - \frac{e^{\tau} \cdot 4\,(1 - u_a)}{(1 - 2u_a)\,(1 - u_e)}}\right) .$$

2.2.3. Equilibrium reputations at the all-DISC equilibrium under the Shunning norm. The Shunning norm is given by $(p, q) = (0, 0)$, so we have $1 - P_{GC} + P_{BD} = (1 - 2u_a)\, u_e + 2u_a$ and $P_{GC} - P_{BC} - P_{GD} + P_{BD} = (1 - 2u_a)\,(1 - u_e)$. The equilibrium reputation of DISC at the all-DISC equilibrium is then

$$r_{\text{DISC}} = \frac{1}{2}\left(1 + \frac{e^{\tau}((1 - 2u_a)\, u_e + 2u_a)}{(1 - 2u_a)\,(1 - u_e)} - \sqrt{\left(1 + \frac{e^{\tau}((1 - 2u_a)\, u_e + 2u_a)}{(1 - 2u_a)\,(1 - u_e)}\right)^2 - \frac{e^{\tau} \cdot 4u_a}{(1 - 2u_a)\,(1 - u_e)}}\right) .$$

2.3. Stability of the all-DISC equilibrium against ALLD under any norm. We focus on the stability of the DISC equilibrium in the main text because it is the only equilibrium under the Stern Judging norm that stably supports cooperation (this is also the case under the Shunning norm). However, the Simple Standing norm admits a stable mixed equilibrium along the ALLC–DISC axis, such that cooperation can be sustained even when an all-discriminator population can be invaded by defectors. To facilitate meaningful comparisons across norms, we analyze the stability of the all-discriminator equilibrium against only ALLD (vs against both ALLD and ALLC as in the analysis under Stern Judging reported in main text).

The replicator dynamic ODE for the competition between ALLD and DISC is given by

$$\frac{df_{\text{DISC}}}{d\eta} = f_{\text{DISC}}\,(1 - f_{\text{DISC}})\,(\pi_{\text{DISC}} - \pi_{\text{ALLD}}) .$$

Here, $r_{\text{ALLD}}, r_{\text{DISC}} \in [0, 1]$ are evaluated at the reputation equilibrium, as before.

The Jacobian of this ODE at the all-discriminator equilibrium ($f_{\text{DISC}} = 1$) is given by

$$J = \left[(1 - u_e)\,(b r_{\text{ALLD}} - (b - c)\, r_{\text{DISC}})\right]\Big|_{f_{\text{DISC}}=1} .$$

The all-discriminator equilibrium is locally stable if and only if $(1 - u_e)\,(b r_{\text{ALLD}} - (b - c)\, r_{\text{DISC}}))\,|_{f_{\text{DISC}}=1} < 0$. For a general norm parametrized by (p, q) (see *Social Norms* in *Materials and Methods*), this condition simplifies to

(i′) $\quad \dfrac{b}{c} > \left(\dfrac{b}{c}\right)^{*} = \dfrac{1}{(1 - 2u_a)\,(1 - u_e)}$ $\quad$ and

(ii′)
$$\begin{cases} \tau > \tau^{*}_{\text{ALLD}} = \log\left[(1 - p + q)\left(1 - \dfrac{u_a + q(1 - 2u_a)}{1 + q(1 - 2u_a)} \cdot \dfrac{\left(\dfrac{b}{c}\right)}{\left(\dfrac{b}{c}\right) - \dfrac{1}{1 + q(1 - 2u_a)}}\right)\left(\dfrac{\dfrac{b}{c}}{\dfrac{b}{c} - \left(\dfrac{b}{c}\right)^{*}}\right)\right] & \text{if } \tau^{*}_{\text{ALLD}} \geq 0 , \\ \tau \geq 0 & \text{if } \tau^{*}_{\text{ALLD}} < 0 . \end{cases}$$

Note that τ^{*}_{ALLD} is undefined for the Scoring norm ($(p, q) = (1, 0)$). This is consistent with the fact that, under a first-order norm, gossip will not impact the stability of DISC because equilibrium reputations do not depend on the level of agreement about social reputations. For any norm other than Scoring (i.e., $(p, q) \in [0, 1]^2 \setminus (1, 0)$), τ^{*}_{ALLD} is a decreasing function of b/c for all $(b/c) > (b/c)^{*}$ (i.e., when (i′) is satisfied), meaning that less gossip is required to stabilize cooperation when the benefit-to-cost ratio is larger.

2.3.1. The effect of the social norm on the critical gossip duration. We evaluate conditions (i′) and (ii′) for the three second-order norms of interest: Stern Judging ($(p,q)=(0,1)$), Simple Standing ($(p,q)=(1,1)$), and Shunning ($(p,q)=(0,0)$) (Fig. S2A). Consistent with our intuition, the duration of gossip τ^*_{ALLD} needed to stabilize DISC against ALLD is the highest for the Shunning norm, the lowest for the Simple Standing norm, and intermediate for the Stern Judging norm.

The conditions above also allow us to study the impact of a general social norm (p,q) on the critical gossip duration. We find analytically that the critical gossip duration τ^*_{ALLD} is decreasing in p (i.e., $\partial\tau^*_{\text{ALLD}}/\partial p<0$ for any $b>c>0$ and $0<u_a,u_e<1/2$; see a numerical example in Fig. S2B). This is consistent with the intuition that increasing the parameter p makes the norm more 'lenient', incentivizes cooperating with bad individuals, and therefore reduces the amount of gossip needed to stabilize cooperation.

In contrast, we find that τ^*_{ALLD} is increasing or decreasing in q depending on parameter conditions: if condition (i′) is satisfied ($(b/c)>(b/c)^*$), then

$$\frac{\partial\tau^*_{\text{ALLD}}}{\partial q}>0 \iff u_a\geq\frac{1-u_e}{2(2-u_e)} \quad \text{or} \quad \frac{b}{c}>\frac{1}{2u_a} \quad \text{or} \quad p>\frac{1-2u_a(b/c)}{(b/c)(1-2u_a)}\,.$$

Numerical examples in Fig. S2C and D are consistent with these analytical results. Increasing the parameter q generally makes the norm more 'strict' and incentivizes defecting against bad individuals. This can in turn promote cooperation and thus lower the critical gossip duration, at least when assessments are relatively accurate (low u_a) and cooperating with bad individuals is disincentivized (low p) (Fig. S2C). When p is high(er), however, this effect is reversed for some combinations of the benefit-to-cost ratio b/c and the assessment error rate u_a (Fig. S2D).

2.3.2. The effect of assessment and execution errors on the critical gossip duration. To determine how errors in assessment or execution impact the amount of gossip needed to stabilize a population of DISC against ALLD, we evaluate the derivatives of the critical gossip duration τ^*_{ALLD} with respect to the assessment error rate u_a and the execution error rate u_e.

The critical gossip duration τ^*_{ALLD} is increasing in u_e (see numerical examples in Fig. S7A–C): if condition (i′) is satisfied ($(b/c)>(b/c)^*$), then we have

$$\frac{\partial\tau^*_{\text{ALLD}}}{\partial u_e}=\frac{(b/c)^*}{(1-u_e)\left((b/c)-(b/c)^*\right)}>0$$

for any $b>c>0$ and $0<u_a,u_e<1/2$.

In contrast, τ^*_{ALLD} is increasing or decreasing in u_a depending on the social norm and parameter values. Under Stern Judging and Simple Standing, τ^*_{ALLD} is increasing in u_a (i.e., $\partial\tau^*_{\text{ALLD}}/\partial u_a>0$ for any $b>c>0$ and $0<u_a,u_e<1/2$ (numerical examples in Fig. S7D and E). However, under Shunning, τ^*_{ALLD} can be monotonic in u_a (numerical example in Fig. S7F): we have

$$\frac{\partial\tau^*_{\text{ALLD}}}{\partial u_a}>0 \quad\iff\quad \frac{b}{c}<\frac{4}{3-4u_a-\sqrt{1+8u_e+8u_a(1-2u_a-4(1-u_a)u_e)}}$$

assuming condition (i′) is satisfied ($(b/c)>(b/c)^*$). For $u_a=u_e=0.02$, the condition on the right-hand side evaluates to $b/c<2.248$.

2.4. Impact of paradoxical cooperators on the stability of cooperation under the Stern Judging norm. While our analysis has focused on the competition among cooperators (ALLC), defectors (ALLD), and discriminators (DISC), prior work has shown that, under a public reputation scheme, the existence of paradoxical discriminators (pDISC)—who cooperate with bad individuals and defect against good individuals—can be beneficial for cooperation under Stern Judging due to the symmetry of this norm (2, 3).

To study whether pDISC has a similar effect on gossip-based cooperation, we expand our model to include four strategies (ALLC, ALLD, DISC, pDISC), and we analyze the stability of the all-DISC and all-pDISC equilibria under the Stern Judging norm.

2.4.1. Strategy updates in the presence of paradoxical discriminators. As in the main text (Eq. 6), the dynamics of strategy frequencies are governed by the replicator ODEs,

$$\frac{df_s}{d\eta}=f_s(\eta)\left(\pi_s(\eta)-\bar{\pi}(\eta)\right), \quad s\in S=\{\text{ALLC},\text{ALLD},\text{DISC},\text{pDISC}\}\,.$$

The presence of paradoxical discriminators alters the payoffs for ALLC, ALLD, and DISC because a pDISC strategist provides a benefit b when interacting with a recipient with a bad reputation (barring errors). Let f_{pDISC} and π_{pDISC} be the frequency and payoff of paradoxical discriminators, respectively. The new payoffs for ALLC, ALLD, and DISC are given by

$$\begin{aligned}
\pi_{\text{ALLC}}&=(1-u_e)\left[b\left(f_{\text{ALLC}}+f_{\text{DISC}}\cdot r_{\text{ALLC}}+f_{\text{pDISC}}\cdot(1-r_{\text{ALLC}})\right)-c\right],\\
\pi_{\text{ALLD}}&=(1-u_e)\left[b\left(f_{\text{ALLC}}+f_{\text{DISC}}\cdot r_{\text{ALLD}}+f_{\text{pDISC}}\cdot(1-r_{\text{ALLD}})\right)\right],\\
\pi_{\text{DISC}}&=(1-u_e)\left[b\left(f_{\text{ALLC}}+f_{\text{DISC}}\cdot r_{\text{DISC}}+f_{\text{pDISC}}\cdot(1-r_{\text{DISC}})\right)-c\cdot r\right],
\end{aligned}$$

where, as before, $r\triangleq\sum_{s\in S}f_s\cdot r_s$, $S=\{\text{ALLC},\text{ALLD},\text{DISC},\text{pDISC}\}$ is the average reputation of the population. Similarly, the payoff for pDISC is given by

$$\pi_{\text{pDISC}}=(1-u_e)\left[b\left(f_{\text{ALLC}}+f_{\text{DISC}}\cdot r_{\text{pDISC}}+f_{\text{pDISC}}\cdot(1-r_{\text{pDISC}})\right)-c\cdot(1-r)\right],$$

2.4.2. Reputation updates in the presence of paradoxical discriminators. As before (Eq. 2 in the main text), the reputation dynamics are governed by the system of ODEs,

$$\frac{dr_s}{dt} = p_s(t) - r_s(t), \quad s \in S = \{\text{ALLC}, \text{ALLD}, \text{DISC}, \text{pDISC}\} \ ,$$

where $p_s(t)$ is the probability that an individual of strategic type s will be assigned a good reputation by an observer, which depends on the current reputations in the population. The quantities p_{ALLC}, p_{ALLD}, and p_{DISC} are as before (Eq. 3); below we derive p_{pDISC}.

A paradoxical discriminator (pDISC) gains a good reputation by

(1) interacting with someone who has a *good* reputation in the eyes of both the donor and the observer (which occurs with probability $\tilde{g}_2$), intending to *defect*, and being assigned a good reputation (with probability P_{GD});

(2) interacting with someone who has a good reputation in the eyes of the donor but a *bad* reputation in the eyes of the observer (with probability $\tilde{d}_2$), intending to *defect*, and being assigned a good reputation (with probability P_{BD});

(3) interacting with someone who has a bad reputation in the eyes of the donor but a *good* reputation in the eyes of the observer (with probability $\tilde{d}_2$), intending to *cooperate*, and being assigned a good reputation (with probability P_{GC}); or

(4) interacting with someone who has a *bad* reputation in the eyes of both the donor and the observer (with probability $\tilde{b}_2$), intending to *cooperate*, and being assigned a good reputation (with probability P_{BC}).

Thus, the probability that a paradoxical discriminator earns a good reputation is given by

$$p_{\text{pDISC}} = \tilde{g}_2 P_{GD} + \tilde{d}_2 \left(P_{BD} + P_{GC}\right) + \tilde{b}_2 P_{BC} \ .$$

2.4.3. Stability of the all-DISC and all-pDISC equilibria. To determine the amount of gossip required to sustain cooperation in the expanded strategy space, we analyzed the local stability of the discriminator equilibrium (all-DISC, $f_{\text{DISC}} = 1$) and the paradoxical-discriminator equilibrium (all-pDISC, $f_{\text{pDISC}} = 1$).

We find that the stability conditions for the all-DISC and all-pDISC equilibria are identical: both equilibria are locally stable under the Stern Judging norm if and only if conditions (i) and (ii) in the main text are satisfied. In other words, the presence of pDISC has no impact on the minimum duration of gossip τ^* needed to stabilize the all-DISC equilibrium. Notably, the equilibrium reputation of discriminators at the all-DISC equilibrium ($r_{\text{DISC}}|_{f_{\text{DISC}}=1}$) and the equilibrium reputation of paradoxical discriminators at the all-pDISC equilibrium ($r_{\text{pDISC}}|_{f_{\text{pDISC}}=1}$) sum to 1:

$$r_{\text{DISC}}|_{f_{\text{DISC}}=1} = 1 - r_{\text{pDISC}}|_{f_{\text{pDISC}}=1} \ .$$

This implies that the rate of cooperation achieved at the all-DISC and all-pDISC equilibria, which are proportional to the two quantities in the equation above, are identical:

$$(1-u_e)\left(r_{\text{DISC}}|_{f_{\text{DISC}}=1}\right) = (1-u_e)\left(1 - r_{\text{pDISC}}|_{f_{\text{pDISC}}=1}\right) \ .$$

These results are consistent with the previous finding that Stern Judging is a symmetric norm that benefits from both DISC and pDISC strategies (2, 3).

3. Agreement and disagreement after biased peer-to-peer gossip

In this section, we derive the expressions for the agreement and disagreement terms, $\tilde{g}_2$, $\tilde{b}_2$, and $\tilde{d}_2$, after biased gossip (Eq. 9).

3.1. Gossip process for a focal individual. We consider a population of N individuals engaged in gossip. Suppose that, at time T, there are ℓ individuals who believe a focal individual i is good and $N-\ell$ who believe i is bad. We assume that, with probability u (v), an individual who considered i as good (bad) "mutates" to the opposite opinion between time T and $T+1$. Thus, the dynamics of biased gossip for a focal individual follow a Wright-Fisher process in a haploid population with two alleles, which keeps track of how many individuals view the focal individual as good (allele one) or bad (allele two) over discrete generations (rounds) of gossip.

Let $R_{i,T} \in \{0, 1/N, \ldots, (N-1)/N, 1\}$ be a random variable that tracks the frequency of allele one at time T (after T rounds of gossip). The probability that there are m individuals who believe i is good at time $T+1$ is

$$p_{m\ell} = \mathbb{P}\left(R_{i,T+1} = \frac{m}{N} \,\middle|\, R_{i,T} = \frac{\ell}{N} \right) = \binom{N}{m} \left(g\left(\frac{\ell}{N}\right) \right)^m \left(1 - g\left(\frac{\ell}{N}\right) \right)^{N-m}$$

for $0 \leq m \leq N$, where the function

$$g(r_{i,T}) = r_{i,T}\,(1-u) + (1-r_{i,T})\,v = (1-u-v)\,r_{i,T} + v$$

gives the proportion of gossip transmitted between T and $T+1$ that is positive (i.e., views i as good), provided that a fraction $r_{i,T}$ view i as good at time T. In the absence of mutation ($u = v = 0$), we recover the model of gossip as pure drift ($g(r_{i,T}) = r_{i,T}$).

The mean and variance of the distribution of $R_{i,T}$ are given by (see Tataru et al. (4, 5) for derivations)

$$\begin{aligned}
\mathbb{E}\left[R_{i,T} \,\middle|\, R_{i,0} = r_{i,0} \right] &= \left(r_{i,0} - \frac{v}{u+v} \right) (1-u-v)^T + \frac{v}{u+v}\,, \\
\mathrm{Var}\left(R_{i,T} \,\middle|\, R_{i,0} = r_{i,0} \right) &= \frac{v}{u+v}\left(1 - \frac{v}{u+v} \right) \left[\frac{1 - \left(1 - \frac{1}{N}\right)^T (1-(u+v))^{2T}}{N - (N-1)\,(1-(u+v))^2} \right] \\
&\quad + \left(1 - 2 \cdot \frac{v}{u+v} \right) \left(r_{i,0} - \frac{v}{u+v} \right) (1-(u+v))^T \left[\frac{1 - \left(1 - \frac{1}{N}\right)^T (1-(u+v))^T}{N - (N-1)\,(1-(u+v))} \right] \\
&\quad - \left(r_{i,0} - \frac{v}{u+v} \right)^2 (1-(u+v))^{2T} \left[1 - \left(1 - \frac{1}{N} \right)^T \right]\,.
\end{aligned}$$

Assuming (1) N is large and (2) u and v are small, we can approximate these quantities as (see 4, 5)

$$\begin{aligned}
\mathbb{E}\left[R_{i,\tau} \,\middle|\, R_{i,0} = r_{i,0} \right] &= \left(r_{i,0} - \frac{\nu}{\mu+\nu} \right) e^{-(\mu+\nu)\tau} + \frac{\nu}{\mu+\nu}\,, \\
\mathrm{Var}\left(R_{i,\tau} \,\middle|\, R_{i,0} = r_{i,0} \right) &= \frac{\nu}{\mu+\nu}\left(1 - \frac{\nu}{\mu+\nu} \right) \cdot \frac{1}{1+2\,(\mu+\nu)} \left(1 - e^{-(2(\mu+\nu)+1)\tau} \right) \\
&\quad + \left(1 - \frac{2\nu}{\mu+\nu} \right) \left(r_{i,0} - \frac{\nu}{\mu+\nu} \right) \frac{1}{1+(\mu+\nu)} \cdot e^{-(\mu+\nu)\tau} \left(1 - e^{-((\mu+\nu)+1)\tau} \right) \\
&\quad - \left(r_{i,0} - \frac{\nu}{\mu+\nu} \right)^2 e^{-2(\mu+\nu)\tau} \left(1 - e^{-\tau} \right)\,,
\end{aligned} \tag{S3}$$

as reported in Eq. 8, where $\mu = Nu$ and $\nu = Nv$ are the scaled mutation rates and $\tau = T/N$ is the scaled gossip duration. We recover the case of noiseless gossip by letting $\mu = \nu = 0$ and setting $0/0 := 1$ (5).

3.2. Population-level agreement and disagreement. We derive the agreement and disagreement terms, $\tilde{g}_2$, $\tilde{b}_2$, and $\tilde{d}_2$, by first computing the following quantities for a focal individual i:

$$\begin{aligned}
\mathbb{E}\left[R_{i,\tau}^2 \,\middle|\, R_{i,0} = r_{i,0} \right] &= \mathrm{Var}\left(R_{i,\tau} \,\middle|\, R_{i,0} = r_{i,0} \right) + \mathbb{E}\left[R_{i,\tau} \,\middle|\, R_{i,0} = r_{i,0} \right]^2\,, \\
\mathbb{E}\left[(1-R_{i,\tau})^2 \,\middle|\, R_{i,0} = r_{i,0} \right] &= 1 - 2\,\mathbb{E}\left[R_{i,\tau} \,\middle|\, R_{i,0} = r_{i,0} \right] + \mathbb{E}\left[R_{i,\tau}^2 \,\middle|\, R_{i,0} = r_{i,0} \right]\,, \\
\mathbb{E}\left[R_{i,\tau}(1-R_{i,\tau}) \,\middle|\, R_{i,0} = r_{i,0} \right] &= \mathbb{E}\left[R_{i,\tau} \,\middle|\, R_{i,0} = r_{i,0} \right] - \mathbb{E}\left[R_{i,\tau}^2 \,\middle|\, R_{i,0} = r_{i,0} \right]\,.
\end{aligned}$$

As discussed in the main text (*Noisy Gossip Is Less Beneficial for Cooperation*), the gossip process for a focal individual i is initialized as follows: at the start of each gossip period ($T = 0$), we assume that the fraction r_i of those engaged in gossip who view a given focal individual i of type s as good is identical to the fraction r_s of the population who view type s as good in the context of the reputation ODEs (Eq. 2). In

other words, the initial allele one frequency in the gossip process about individual i, $r_{i,0}$, is r_s. Therefore, the quantities $\tilde{g}_2$, $\tilde{b}_2$, and $\tilde{d}_2$ can be computed as

$$\tilde{g}_2 = \sum_s f_s \cdot \mathbb{E}\left[R_{i,\tau}^2 \,\middle|\, R_{i,0} = r_s\right] \ ,$$
$$\tilde{b}_2 = \sum_s f_s \cdot \mathbb{E}\left[(1 - R_{i,\tau})^2 \,\middle|\, R_{i,0} = r_s\right] \ ,$$
$$\tilde{d}_2 = \sum_s f_s \cdot \mathbb{E}\left[R_{i,\tau}(1 - R_{i,\tau}) \,\middle|\, R_{i,0} = r_s\right] \ ,$$

as reported in Eq. 9.

3.3. Impact of noisy gossip on cooperation under the Scoring norm. Under the Scoring norm, noiseless gossip will not impact cooperation; this is because noiseless gossip changes the degree of agreement about reputations but not the mean of reputations, and only the latter impacts the stability of cooperation under a first-order norm. Indeed, in the absence of noise ($\mu = \nu = 0$), the all-discriminator equilibrium is stable against defectors under the Scoring norm as long as $(b/c) > (b/c)^*$ (condition (i′) in Section 2.3; identical to condition (i) in the main text), regardless of the duration of gossip τ.

In contrast, noisy gossip does affect cooperation under Scoring because it alters the mean reputation of each strategic type (Eq. S3 and Eq. 8 in the main text). For example, for the standard error rates $u_e = u_a = 0.02$, infinitely long gossip with a maximally positive bias ($\tau \to \infty$ and $\beta = +1$, i.e., $\mu = 0$, $\nu > 0$) can stabilize DISC against ALLD under Scoring provided $(b/c) >= 1.021$, which is less strict than condition (i') (which evaluates to $(b/c) > (b/c)^* = 1.063$).

In fact, we can show analytically that the minimum b/c required to stabilize DISC against ALLD is lower for noisy gossip than for noiseless gossip, provided the bias is sufficiently positive. More precisely, we find that the critical b/c threshold for given scaled mutation rates (μ and ν) and gossip duration (τ) is less than $(b/c)^*$ for noiseless gossip (of any duration) if

$$\beta > \beta^* = \frac{-(1 - 2u_a)u_e}{1 - (1 - 2u_a)(1 - u_e)} \ .$$

Note that β^* is constrained to $-1 < \beta^* < 0$ for $0 < u_a, u_e < 1/2$ (as a numerical example, for $u_e = u_a = 0.02$, the condition evaluates to $\beta > \beta^* = -0.3243$). This implies that gossip with any positive bias ($\beta > 0$) will outperform noiseless gossip, reducing the critical b/c threshold and making it easier to sustain cooperation under Scoring. It also implies that gossip with maximally negative bias ($\beta = -1$) will never outperform noiseless gossip.

4. Continuous-time model of gossip and interactions

In the main text, we develop a model of reputation change as a result of two discrete processes: pairwise interactions between all players in the population, and "epochs" of gossip. In this section, we develop an intercalated model in which interactions and gossip occur as Poisson processes on overlapping time scales, at rate 1 and rate γ, respectively. We apply this analysis to both peer-to-peer gossip and single-source gossip. We show that this model, in the large-N, mean-field approximation, yields results identical to the models in the main text, subject to a change of variables.

Our approach is as follows:

1. We define an "image matrix" r_{ij}; a value of 1 means that player i has a good view of j and a value of 0 means that i has a bad view of j.
2. We consider the expected change in r_{ij} as a result of private observation of interactions and as a result of gossip, treated as discrete events. Observations occur at rate 1, and gossip occurs at rate γ.
3. We repeat this analysis for the "good reputation agreement" term $g_{ijk} = r_{ik}r_{jk}$, which is 1 if i and j agree that k is good and 0 otherwise. This is a discrete analogue of $\tilde{g}_2$ in the main text. Similarly, the analogue for $\tilde{d}_2$ is $r_{ik} - g_{ijk}$ or $r_{jk} - g_{ijk}$ (the i and j labels are arbitrary and may be interchanged), and the analogue for $\tilde{b}_2$ is $1 - r_{ik} - r_{jk} + g_{ijk}$.
4. We sum these terms over the population to recast the discrete random variables r_{ij} and g_{ijk} in terms of the average reputation r_s and agreement g_s associated with behavioral strategy s; these are continuous variables whose time evolution can be described by ordinary differential equations.
5. We show that the equilibria of these ODEs behave exactly as in the main text, subject to a change of variables.

The principal result is that, under peer-to-peer gossip, the time evolution of reputations and agreement in a population can be described by the ordinary differential equations

$$\begin{aligned} \frac{dr_s}{d\phi} &= \frac{1}{N^2}\left(p_s - r_s\right), \\ \frac{dg_s}{d\phi} &= \frac{2}{N^2}\left(r_s p_s - g_s + \Gamma(r_s - g_s)\right) = \frac{2}{N^2}\left(r_s\left(p_s + \Gamma\right) - g_s\left(1 + \Gamma\right)\right), \end{aligned} \qquad \text{[S4]}$$

where $\Gamma = \gamma/N$ is the normalized gossip rate and p_s is the equilibrium reputation of strategy s, which is equivalent to the probability that an individual of strategy s is assigned a good reputation at equilibrium (Table S1). The steady-state conditions are

$$\begin{aligned} r_s &= p_s, \\ g_s &= \frac{r_s(p_s + \Gamma)}{1 + \Gamma} = \frac{r_s(r_s + \Gamma)}{1 + \Gamma}. \end{aligned}$$

By rescaling time in the first equation in Eq. S4, we obtain the reputation ODE used in the main text (Eq. 2),

$$\frac{dr_s}{dt} = p_s(t) - r_s(t)\ .$$

Finally, we obtain similar relations for gossip with a single source.

Variable	Definition
i, j, k	focal players whose reputations we track
I, J, K	behavioral strategies of i, j, k (respectively)
$\sigma_I, \sigma_J, \sigma_K$	the sets of all players using strategies I, J, K (respectively)
r_{ij}	player i's view of player $j \in \{0, 1\}$
g_{ijk}	$r_{ik}r_{jk}$, whether players i and j agree that k is good
r_{iJ}	i's average view of J-players
$r_{\bullet J}$	the whole population's average view of J-players
p_{ijk}	i's expected new view of j, after i sees j interact with k
ω, θ, ρ	arbitrary players, used primarily as indexing variables
γ	rate at which pairwise gossip occurs
N	population size
Γ	normalized gossip rate γ/N

Table S1. Notation used in Section 4.

4.1. Notation. Throughout this section, we use lowercase letters i, j, and k to denote individual players and capital letters I, J, and K to denote their strategies. We further denote by σ_I the set of all individuals with behavioral strategy I. Each behavioral strategy s is described by two numbers a^G and a^B, which indicate whether they cooperate (1) or defect (0) with someone with a good (G) or bad (B) reputation; we recover the strategies ALLC, ALLD, DISC, and pDISC by setting $(a^G, a^B) = (1,1), (0,0), (1,0), (0,1)$, respectively. In principle, the values of a^G, a^B need not be integers; they may take on values between 0 and 1 and be considered either as probabilities or as "fractional" cooperation events (6–8).

For simplicity, we introduce the notation

$$\langle r_{ij} \rangle_i = \frac{1}{N} \sum_{i=1}^{N} r_{ij} \triangleq r_{\bullet j},$$
$$\langle r_{ij} \rangle_{\sigma_I} = \frac{1}{f_I} \frac{1}{N} \sum_{i \in \sigma_I} r_{ij} \triangleq r_{Ij}.$$

That is, a bullet $\bullet$ subscript indicates that the sum has been carried out over the entire population, and a behavioral strategy I (or J) subscript indicates that the sum has been carried out only over σ_I, so that $r_{\bullet j}$ is the population's average view of player j, and likewise r_{Ij} is the average I-player's view of individual j. We will later demonstrate that the first index does not matter and drop it, so that

$$r_J = r_{\bullet J},$$
$$r = r_{\bullet\bullet} = \sum_{J \in S} f_J r_J,$$

where (as in the main text) S ranges over all strategies. At that point, we will revert to using s to denote behavioral strategies, consistent with the main text.

When the average has more than one subscript, it is to be understood that both indices are summed over; for example,

$$\langle r_{ij} \rangle_{ij} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} r_{ij} \triangleq r_{\bullet\bullet},$$
$$\langle r_{ij} \rangle_{\sigma_I} = \frac{1}{f_I f_J} \frac{1}{N^2} \sum_{i \in \sigma_I} \sum_{i \in \sigma_J} r_{ij} \triangleq r_{IJ}.$$

The notation is similar for the agreement term:

$$\langle g_{ijk} \rangle_{ijk} = \frac{1}{N^3} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} g_{ijk} \triangleq g_{\bullet\bullet\bullet},$$
$$\langle g_{ijk} \rangle_{\sigma_I \sigma_J \sigma_K} = \frac{1}{f_I f_J f_K} \frac{1}{N^3} \sum_{\sigma_I} \sum_{\sigma_J} \sum_{\sigma_K} g_{ijk} \triangleq g_{IJK},$$
$$\langle g_{IJK} \rangle_{ijk} = \sum_I \sum_J \sum_K f_I f_J f_K g_{IJK} = g_{\bullet\bullet\bullet}.$$

As with r_s, we will demonstrate that the first two indices do not matter and use

$$g_s = g_{\bullet\bullet K},$$
$$g = \sum_{K \in S} f_K g_K.$$

Note that

$$g_{\bullet\bullet k} = \langle r_{ik} r_{jk} \rangle_{ij} = r_{\bullet k}^2.$$

If all i and j agree that a given k is good, we have $r_{\bullet k} = 1$, and if all i and j agree that a given k is bad, we have $r_{\bullet k} = 0$; if either of these conditions obtains for all k, then we will have $g_{\bullet\bullet K} = r_{\bullet K}$. If, on the other hand, each i and j has an independent view of k, then we will instead have $g_{\bullet\bullet K} = (r_{\bullet K})^2$. As we will see, privately observed interactions cause $g_{\bullet\bullet K}$ to decay toward $(r_{\bullet K})^2$, and gossip causes it to decay toward $r_{\bullet K}$.

4.2. Incorporating errors. The expressions we provide here for the equilibrium reputation p_s of strategy s assume no errors. Errors may be incorporated as follows:

4.2.1. Assessment error. To incorporate assessment error, it is sufficient to transform

$$p_s \to p_s(1-2u_a) + u_a.$$

This mapping effectively sends 1 to $1-u_a$ and 0 to u_a, thus "perturbing" the reputation equilibria toward $1/2$; this is often sufficient to guarantee that the reputation dynamics have a single stable equilibrium. It may also be interpreted as follows. With probability $1-2u_a$, the observer i applies the social norm correctly, i.e., i determines j's reputation correctly based on j's behavior toward k and i's view of k. With probability $2u_a$, i instead flips a coin.

4.2.2. Execution error. Execution error is implemented in the main text, as is common in the literature, as follows; with probability $1-u_e$, an individual who intends to cooperate accidentally defects. An individual who intends to defect never accidentally cooperates. This is tantamount to saying that, with probability $1-u_e$, they play their intended strategy; with probability u_e, they instead play ALLD. Thus, execution error can be incorporated via the transformation

$$p_s \to (1-u_e)p_s + u_e p_{\text{ALLD}}.$$

Note that incorporating both assessment error and execution error yields

$$\left((1-u_e)p_s + u_e p_{\text{ALLD}}\right)(1-2u_a) + u_a = \left(p_s(\varepsilon - u_a) + p_{\text{ALLD}}(1-u_a-\varepsilon)\right) + u_a.$$

4.2.3. Perception error. A third type of error that appears in the literature is "perception error" (9), wherein an individual is occasionally misperceived as though they had performed the opposite action. That is, if j cooperated with k, i may accidentally judge j as if j had defected with k; likewise, if j defected with k, i may accidentally judge j as if j had cooperated with k. We denote the probability of such an error as u_p. Although we did not discuss this type of error in the main text, it can be incorporated via the transformation

$$p_s \to (1-u_p)p_s + u_p p_{s^*},$$

where s^* is a strategy such that $a^G_{s^*} = 1 - a^G_s$ and $a^B_{s^*} = 1 - a^B_s$; it is effectively the "opposite" strategy of s. For ALLC, this is ALLD; for DISC, it is pDISC; and vice versa.

Note that assessment error and perception error may behave differently depending on the norm. For example, under Stern Judging, we have (in the absence of errors) $\{P_{GC}, P_{GD}, P_{BC}, P_{BD}\} = \{1,0,0,1\}$. Under public assessment, this yields

$$p_{\text{DISC}} = \begin{cases} 1-u_a, & \text{assessment error only,} \\ 1-u_p, & \text{perception error only;} \end{cases}$$

these types of error behave the same. In contrast, under the Shunning norm, we have (in the absence of errors) $\{P_{GC}, P_{GD}, P_{BC}, P_{BD}\} = \{1,0,0,0\}$. Under public assessment, this yields

$$p_{\text{DISC}} = \begin{cases} r(1-u_a) + (1-r)u_a, & \text{assessment error only,} \\ r(1-u_p), & \text{perception error only.} \end{cases}$$

4.3. Change in image and agreement. We will analyze the change in the reputation terms r_{IJ} by considering the expected change in r_{ij} in a single time step $\Delta\phi$; we will repeat this analysis for g_{IJK} and the other "moments" above. We consider three processes, in which a randomly chosen individual ω updates their view of θ, $r_{\omega\theta}$.

1. *Interaction*: During a given time step, a randomly chosen observer ω updates their view of a random donor θ by observing θ interact with a recipient ρ. ω's new opinion of θ is then given by $p_{\omega\theta\rho}$, the exact value of which we enumerate in the next section.
2. *Peer-to-peer gossip*: During a given time step, a randomly chosen observer ω updates their view of a random player θ by gossiping with a random partner ρ, thereby adopting ρ's view of θ. ω's new opinion of θ becomes $r_{\rho\theta}$.
3. *Gossip with a single source*: During a given time step, a randomly chosen observer ω consults a designated player known as a *gossip source*, labeled z, for their opinion on θ. In doing so, ω adopts z's opinion of θ.

To obtain the change in the image matrix r_{ij}, we consider an arbitrary process that replaces $r_{\omega\theta}$ with $\xi_{\omega\theta\rho}$, where ρ is some randomly chosen third party. We will later "plug in" the correct values of $\xi_{\omega\theta\rho}$. Let r^+_{ij} be the updated value of r_{ij} in the next time step. We have

$$\begin{aligned} \mathbb{E}[\Delta r_{ij}] &= \mathbb{E}[r^+_{ij}] - r_{ij} \\ &= \left\langle r_{ij} + \delta_{\omega i}\delta_{\theta j}\left(\xi_{\omega\theta\rho} - r_{\omega\theta}\right)\right\rangle_{\omega\theta\rho} - r_{ij} \\ &= \frac{1}{N^2}\left(\xi_{ij\bullet} - r_{ij}\right). \end{aligned} \qquad \text{[S5]}$$

We repeat this calculation for g_{ijk};

$$\begin{aligned}
\mathbb{E}[\Delta g_{ijk}] &= \mathbb{E}[g^+_{ijk}] - g_{ijk} \\
&= \langle (r_{ik} + \delta_{\omega i}\delta_{\theta k}\left(\xi_{\omega\theta\rho} - r_{\omega\theta}\right))\left(r_{jk} + \delta_{\omega j}\delta_{\theta k}\left(\xi_{\omega\theta\rho} - r_{\omega\theta}\right)\right)\rangle_{\omega\theta\rho} - g_{ijk} \\
&= \frac{1}{N^2}\left\langle r_{ik}(\xi_{jk\rho} - r_{jk}) + r_{jk}(\xi_{ik\rho} - r_{ik}) + \delta_{ij}\left(\xi_{ik\rho} - r_{ik}\right)^2\right\rangle_\rho \\
&= \frac{1}{N^2}\left\langle r_{ik}\xi_{jk\rho} + r_{jk}\xi_{ik\rho} - 2g_{ijk} + \delta_{ij}(\xi_{ik\rho} - r_{ik})^2\right\rangle_\rho, \\
&= \frac{1}{N^2}\left(r_{ik}\xi_{jk\bullet} + r_{jk}\xi_{ik\bullet} - 2g_{ijk} + \delta_{ij}(\left\langle\xi^2_{ik\rho}\right\rangle_\rho - 2\xi_{ik\bullet}r_{ik} + r_{ik})\right),
\end{aligned} \tag{S6}$$

where δ is the Kronecker delta. In general, the Kronecker delta terms are corrections for double-counting, which are of order $1/N$. Depending on the specific process $\xi_{\omega\theta\rho}$, these corrections may not contribute to the dominant balance, and thus it may be possible to neglect them.

4.4. Effect of interactions. Suppose that a randomly chosen observer ω updates their view of a random θ by observing θ's interaction with a random recipient ρ. The following outcomes are possible:

1. If ω and θ agree that ρ is *good*, then ω assigns θ a good reputation with probability $a^G_\theta P_{GC} + (1 - a^G_\theta)P_{GD}$.
2. If ω believes ρ is *good* but θ believes ρ is *bad*, then ω assigns θ a good reputation with probability $a^B_\theta P_{GC} + (1 - a^B_\theta)P_{GD}$.
3. If ω believes ρ is *bad* but θ believes ρ is *good*, then ω assigns θ a good reputation with probability $a^G_\theta P_{BC} + (1 - a^G_\theta)P_{BD}$.
4. If ω and θ agree that ρ is *bad*, then ω assigns θ a good reputation with probability $a^B_\theta P_{BC} + (1 - a^B_\theta)P_{BD}$.

Thus, the total probability that ω assigns θ a good reputation is

$$\begin{aligned}
p_{\omega\theta\rho} \triangleq\ & r_{\omega\rho}r_{\theta\rho}\left(a^G_\theta P_{GC} + (1 - a^G_\theta)P_{GD}\right) \\
&+ r_{\omega\rho}(1 - r_{\theta\rho})\left(a^B_\theta P_{GC} + (1 - a^B_\theta)P_{GD}\right) \\
&+ (1 - r_{\omega\rho})r_{\theta\rho}\left(a^G_\theta P_{BC} + (1 - a^G_\theta)P_{BD}\right) \\
&+ (1 - r_{\omega\rho})(1 - r_{\theta\rho})\left(a^B_\theta P_{BC} + (1 - a^B_\theta)P_{BD}\right) \\
=\ & r_{\omega\rho}r_{\theta\rho}(a^G_\theta - a^B_\theta)(P_{GC} - P_{GD} - P_{BC} + P_{BD}) \\
&+ r_{\omega\rho}\left(a^B_\theta(P_{GC} - P_{GD} - P_{BC} + P_{BD}) + P_{GD} - P_{BD}\right) \\
&+ r_{\theta\rho}(a^G_\theta - a^B_\theta)(P_{BC} - P_{BD}) \\
&+ a^B_\theta(P_{BC} - P_{BD}) + P_{BD}.
\end{aligned} \tag{S7}$$

Defining

$$\begin{aligned}
\nu^A_\theta &= (a^G_\theta - a^B_\theta)(P_{GC} - P_{GD} - P_{BC} + P_{BD}), \\
\nu^B_\theta &= a^B_\theta(P_{GC} - P_{GD} - P_{BC} + P_{BD}) + P_{GD} - P_{BD}, \\
\nu^C_\theta &= (a^G_\theta - a^B_\theta)(P_{BC} - P_{BD}), \\
\nu^D_\theta &= a^B_\theta(P_{BC} - P_{BD}) + P_{BD}
\end{aligned}$$

allows us to rewrite Eq. S7 as

$$p_{\omega\theta\rho} = g_{\omega\theta\rho}\nu^A_\theta + r_{\omega\rho}\nu^B_\theta + r_{\theta\rho}\nu^C_\theta + \nu^D_\theta.$$

The general forms of p for the social norms considered in the main text, as well as the behavioral strategies we consider, are summarized in Table S2. As we have previously noted, these expressions are valid for zero error rates.

norm	$p^{\mathbf{ALLC}}_{ijk}$	$p^{\mathbf{ALLD}}_{ijk}$	$p^{\mathbf{DISC}}_{ijk}$	$p^{\mathbf{pDISC}}_{ijk}$
Stern Judging	r_{ik}	$-r_{ik} + 1$	$2g_{ijk} - r_{ik} - r_{jk} + 1$	$-2g_{ijk} + r_{ik} + r_{jk}$
Simple Standing	1	$-r_{ik} + 1$	$g_{ijk} - r_{ik} + 1$	$-g_{ijk} + 1$
Scoring	1	0	r_{jk}	$-r_{jk} + 1$
Shunning	r_{ik}	0	g_{ijk}	$-g_{ijk} + r_{ik}$

Table S2. Error-free values of p_{ijk} for the behavioral strategies and social norms discussed in the main text. Note that the expression for $p^{\mathbf{DISC}}_{ijk}$ under Stern Judging can be simplified to $r_{ik}r_{jk} + (1 - r_{ik})(1 - r_{jk})$.

Then, setting $\xi_{\omega\theta\rho} = p_{\omega\theta\rho}$ in Eq. S5 yields

$$\mathbb{E}[\Delta r_{ij}] = \frac{1}{N^2}\left(p_{ij\bullet} - r_{ij}\right),$$
$$\therefore \mathbb{E}[\Delta r_{IJ}] = \frac{1}{N^2}\left(p_{IJ\bullet} - r_{IJ}\right), \quad \text{with} \quad p_{IJ\bullet} = g_{IJ\bullet}\nu_J^A + r_{I\bullet}\nu_J^B + r_{J\bullet}\nu_J^C + \nu_J^D.$$

Note that this has no dependence on I other than via $r_{I\bullet}$ and $g_{IJ\bullet}$. We repeat this procedure to obtain the expected change in g_{ijk} via Eq. S6:

$$\mathbb{E}[\Delta g_{ijk}] = \frac{1}{N^2}\left(r_{ik}p_{jk\bullet} + r_{jk}p_{ik\bullet} - 2g_{ijk} + \delta_{ij}(\langle p_{ik\rho}\rangle_\rho - 2p_{ik\bullet}r_{ik} + r_{ik})\right).$$

The $r_{ik}p_{jk\bullet}$ and $r_{jk}p_{ik\bullet}$ terms can be summed as follows:

$$\begin{aligned}
\langle r_{ik}p_{jk\bullet}\rangle_{\sigma_I\sigma_J\sigma_K} &= \left\langle r_{ik}g_{jk\bullet}\nu_k^A + r_{ik}r_{j\bullet}\nu_k^B + r_{ik}r_{k\bullet}\nu_k^C + r_{ik}\nu_k^D\right\rangle_{\sigma_I\sigma_J\sigma_K} \\
&= \left\langle r_{ik}g_{jk\bullet}\nu_k^A + r_{ik}r_{j\bullet}\nu_k^B + r_{ik}r_{k\bullet}\nu_k^C + r_{ik}\nu_k^D\right\rangle_{\sigma_I\sigma_J\sigma_K} \\
&= r_{IK}g_{JK\bullet}\nu_K^A + r_{IK}r_{J\bullet}\nu_K^B + r_{IK}r_{K\bullet}\nu_K^C + r_{IK}\nu_K^D, \\
\langle r_{jk}p_{ik\bullet}\rangle_{\sigma_I\sigma_J\sigma_K} &= r_{JK}g_{IK\bullet}\nu_K^A + r_{JK}r_{I\bullet}\nu_K^B + r_{JK}r_{K\bullet}\nu_K^C + r_{JK}\nu_K^D,
\end{aligned}$$

and so

$$\mathbb{E}[\Delta g_{IJK}] = \frac{1}{N^2}\left(r_{IK}p_{JK\bullet} + r_{JK}p_{IK\bullet} - 2g_{IJK} + \delta_{IJ}\frac{f_I}{N}\left\langle\left\langle p_{ik\rho}^2\right\rangle_\rho - 2p_{ik\bullet}r_{ik} + r_{ik}\right\rangle_{\sigma_I\sigma_K}\right).$$

Note that this sum technically requires that we assume

$$\begin{aligned}
\langle r_{ij}r_{ik}\rangle_{\sigma_I\sigma_J\sigma_K} &= r_{IJ}r_{IK}, \\
\langle r_{ij}r_{jk}\rangle_{\sigma_I\sigma_J\sigma_K} &= r_{IJ}r_{JK}.
\end{aligned}$$

These are safe assumptions, however; the values of r_{ij} and r_{ik} are the result of independent Bernoulli trials and thus are uncorrelated, and while interactions do induce short-term correlations between r_{ij} and r_{jk}, these correlations decay as soon as either i updates their view of j or j updates their view of k. This is in contrast to $g_{ijk} = r_{ik}r_{jk}$, where we will later explicitly consider a process that creates such correlations (namely gossip). In the limit of large N, the δ_{IJ} term in g_{IJK} can be neglected, and we have

$$\mathbb{E}[\Delta g_{IJK}] = \frac{1}{N^2}\left(r_{IK}p_{JK\bullet} + r_{JK}p_{IK\bullet} - 2g_{IJK}\right).$$

Armed with these terms, we obtain an ordinary differential equation for the change of the image r_{IJ} and g_{IJK} due to interactions by sending the length of time steps $\Delta\phi \to 0$:

$$\begin{aligned}
\frac{dr_{IJ}}{d\phi} &= \frac{1}{N^2}\left(p_{IJ\bullet} - r_{IJ}\right), \\
\frac{dg_{IJK}}{d\phi} &= \frac{1}{N^2}\left(r_{IK}p_{JK\bullet} + r_{JK}p_{IK\bullet} - 2g_{IJK}\right).
\end{aligned}$$

Since $p_{IJ\bullet}$ does not depend on I except transiently via r_{IJ} and g_{IJK}, we may, without loss of generality, average over the point of view of all observers. (This assumption is valid only because the population is well-mixed and because the ν terms depend solely on the behavior of the focal individual. Relaxing this condition, for example by allowing different individuals to apply different assessment rules, would invalidate this assumption.)

We use the subscript s for consistency with the main text and write down

$$\begin{aligned}
r_s &= r_{\bullet s}, \\
r &= r_\bullet = \sum_{s\in S} f_s r_s, \\
g_s &= g_{\bullet\bullet s}, \\
g &= g_\bullet = \sum_{s\in S} f_s g_s, \\
p_s &= p_{\bullet s\bullet} = g\nu_s^A + r(\nu_s^B + \nu_s^C) + \nu_s^D,
\end{aligned} \tag{S8}$$

which yields

$$\begin{aligned}
\frac{dr_s}{d\phi} &= \frac{1}{N^2}\left(p_s - r_s\right), \\
\frac{dg_s}{d\phi} &= \frac{2}{N^2}\left(r_s p_s - g_s\right).
\end{aligned} \tag{S9}$$

The steady-state behavior is

$$\begin{aligned}
r_s &= p_s, \\
g_s &= r_s p_s = r_s^2.
\end{aligned}$$

4.5. Effect of peer-to-peer gossip. Suppose that a randomly chosen observer ω updates their view of a random θ by gossiping with a random ρ about θ, thereby adopting ρ's view of θ. Then Eq. S5 becomes

$$\mathbb{E}[\Delta r_{ij}] = \frac{1}{N^2}\left(r_{\bullet j} - r_{ij}\right),$$

and so

$$\mathbb{E}[\Delta r_{IJ}] = \frac{1}{N^2}\left(r_{\bullet J} - r_{IJ}\right).$$

The change in g_{ijk} may be obtained via

$$\begin{aligned}\mathbb{E}[\Delta g_{ijk}] &= \frac{1}{N^2}\left(r_{ik}r_{\bullet k} + r_{jk}r_{\bullet k} - 2g_{ijk} + \delta_{ij}\left(\left\langle r_{\rho k}^2\right\rangle_\rho - 2r_{\bullet k}r_{ik} + r_{ik}\right)\right)\\ &= \frac{1}{N^2}\left(g_{i\bullet k} + g_{j\bullet k} - 2g_{ijk} + \delta_{ij}\left(r_{\bullet k} - 2g_{i\bullet k} + r_{ik}\right)\right),\end{aligned}$$

and so

$$\mathbb{E}[\Delta g_{IJK}] = \frac{1}{N^2}\left(g_{I\bullet K} + g_{J\bullet K} - 2g_{IJK} + \delta_{IJ}\frac{f_I}{N}\left(r_{\bullet K} - 2g_{I\bullet K} + r_{IK}\right)\right).$$

Note that the equations for r_{IJ} and g_{IJK} depend only transiently on I and J. We then recast these difference equations as ordinary differential equations to obtain

$$\begin{aligned}\frac{dr_{IJ}}{d\phi} &= \frac{1}{N^2}\left(r_{\bullet J} - r_{IJ}\right),\\ \frac{dg_{IJK}}{d\phi} &= \frac{1}{N^2}\left(g_{I\bullet K} + g_{J\bullet K} - 2g_{IJK} + \delta_{IJ}\frac{f_I}{N}\left(r_{\bullet K} - 2g_{I\bullet K} + r_{IK}\right)\right).\end{aligned}$$

This shows that r_{IJ} approaches $r_{\bullet J}$, and g_{IJK} approaches the average of $g_{I\bullet K}$ and $g_{J\bullet K}$, plus the $1/N$ correction factor. Similarly,

$$\begin{aligned}\frac{dr_{\bullet J}}{d\phi} &= 0,\\ \frac{dg_{I\bullet K}}{d\phi} &= \frac{1}{N^2}\left(g_{\bullet\bullet K} - g_{I\bullet K} + \frac{f_I}{N}\left(r_{\bullet K} - 2g_{\bullet\bullet K} + r_{IK}\right)\right),\\ \frac{dg_{J\bullet K}}{d\phi} &= \frac{1}{N^2}\left(g_{\bullet\bullet K} - g_{J\bullet K} + \frac{f_J}{N}\left(r_{\bullet K} - 2g_{\bullet\bullet K} + r_{JK}\right)\right).\end{aligned}$$

This, too, shows that $g_{I\bullet K}$ and $g_{J\bullet K}$ decay toward $g_{\bullet\bullet K}$, the behavior of which is described by

$$\frac{dg_{\bullet\bullet K}}{d\phi} = \frac{2}{N^3}\left(r_{\bullet K} - g_{\bullet\bullet K}\right).$$

We then invoke Eq. S8 to obtain

$$\begin{aligned}\frac{dr_s}{d\phi} &= 0,\\ \frac{dg_s}{d\phi} &= \frac{2}{N^3}(r_s - g_s).\end{aligned} \qquad \text{[S10]}$$

Thus, gossip has no effect on the average reputation r_s of a strategic type s, but it *does* cause the total agreement g_s to decay toward r_s. In effect, it does so by causing the individual g_k terms to "drift" towards either 0 or 1.

4.6. Effect of interactions and peer-to-peer gossip. We now allow both interactions and gossip to interleave; that is, we specify that interactions occur at (normalized) rate 1 and gossip occurs at rate γ. This is mathematically equivalent to stating that, in a given time step, an event occurs: with probability $1/(\gamma+1)$ it is an interaction between players, and with probability $\gamma/(\gamma+1)$, it is gossip. Equivalently, if i updates their view of j by observing an interaction, γ is the average number of times i gossips about j before the next observation. This diverges slightly from the main text model, in which periods of interaction are separated by "epochs" of gossip of length T, at which point r_s and g_s are "refreshed" to their equilibrium values; in this version of the model, r_{ij} and g_{ijk} are independently updated by coterminous processes, which yields different behavior but similar long-term dynamics, once variables are transformed correctly.

We will soon see that γ needs to be of order N or greater in order for gossip to be significant, so for notational convenience, we define $\Gamma = \gamma/N$. Combining Eqs. S9 and S10 yields the system

$$\begin{aligned}\frac{dr_s}{d\phi} &= \frac{1}{N^2}\left(p_s - r_s\right),\\ \frac{dg_s}{d\phi} &= \frac{2}{N^2}\left(r_s p_s - g_s + \Gamma(r_s - g_s)\right) = \frac{2}{N^2}\left(r_s\left(p_s + \Gamma\right) - g_s\left(1+\Gamma\right)\right), \quad \text{with} \quad p_s = g\nu_s^A + r(\nu_s^B + \nu_s^C) + \nu_s^D.\end{aligned}$$

The steady-state conditions are

$$r_s = p_s,$$
$$g_s = \frac{r_s(p_s+\Gamma)}{1+\Gamma} = \frac{r_s(r_s+\Gamma)}{1+\Gamma}.$$

When $\gamma = 0$, this is simply r_s^2 (the private assessment limit), and when $\gamma \to \infty$, this is r_s (the public assessment limit), as expected. Tuning Γ allows us to smoothly interpolate between these extremes. The general form for g_s thus becomes

$$g_s = \frac{g_s^{\text{private}} + \Gamma g_s^{\text{public}}}{1+\Gamma},$$

which may be compared to the main text model, in which $g_s = g_s^{\text{private}} e^{-\tau} + g_s^{\text{public}}(1-e^{-\tau})$.

All of the expressions in these models are equivalent under the transformation

$$e^{-\tau} = \frac{1}{1+\Gamma}, \quad \text{so that} \quad \Gamma = e^{\tau} - 1 \quad \text{or} \quad \tau = \log(1+\Gamma). \tag{S11}$$

The slight difference between these versions of the model is that in the main text, interactions and gossip occur as discrete events: $T = \tau N$ is the total number of peer-to-peer gossip events between interactions, and the disagreement rate is "refreshed" by observation only after T gossip events have occurred. In this version of the model, $\gamma = \Gamma N$ is the average number of gossip events between interactions *for a given pair of players*. That is, for every time i observes j act, i gossips about j roughly γ times. γ is thus comparable to T, but the fact that the rest of the population is observing and gossiping at the same time prevents the strategy-wise average disagreement g_s from decaying as quickly as an exponential; it instead decays with γ as $1/(1+\gamma)$, which is slower.

4.7. Effect of interactions and consultation with a single gossip source. We now consider an alternative gossip model, namely consultation with a *gossip source* (rather than peer-to-peer gossip). The source, z, is randomly selected from the population and behaves like any other individual, but they are occasionally solicited for their opinion: a player i may seek z's opinion about j and adopt it. We thus need to pay special attention to the dynamics of the source's opinions r_{zj} and agreement g_{zjk}.

As usual,

$$\begin{aligned}
\mathbb{E}[\Delta r_{zj}] &= \frac{1}{N^2}\left(p_{zj\bullet} - r_{zj}\right) \\
\therefore \mathbb{E}[\Delta r_{zJ}] &= \frac{1}{N^2}\left(p_{zJ\bullet} - r_{zJ}\right), \text{with} \\
p_{zJ\bullet} &= g_{zJ\bullet}\nu_J^A + r_{z\bullet}\nu_J^B + r_{J\bullet}\nu_J^C + \nu_J^D, \text{and} \\
\mathbb{E}[\Delta g_{zjk}] &= \frac{1}{N^2}\left(r_{zk}p_{jk\bullet} + r_{jk}p_{zk\bullet} - 2g_{zjk} + \delta_{zj}\left(\left\langle p_{jk\rho}^2\right\rangle_\rho - 2p_{jk\bullet}r_{jk} + r_{jk}\right)\right) \\
\therefore \mathbb{E}[\Delta g_{zJK}] &\approx \frac{1}{N^2}\left(r_{zK}p_{JK\bullet} + r_{JK}p_{zK\bullet} - 2g_{zJK}\right).
\end{aligned}$$

In the expression for $\mathbb{E}[\Delta g_{zJK}]$, we dropped the δ term because it introduces corrections of order $1/N$; the dominant balance of the dynamics is given by the preceding terms. When gossip with the source occurs, we have $\xi_{\omega\theta\rho} = r_{z\theta}$ in Eq. S6, and so

$$\begin{aligned}
\mathbb{E}[\Delta r_{ij}] &= \frac{1}{N^2}\left(r_{zj} - r_{ij}\right) \\
\therefore \mathbb{E}[\Delta r_{IJ}] &= \frac{1}{N^2}\left(r_{zJ} - r_{IJ}\right) \\
\mathbb{E}[\Delta g_{zjk}] &= \frac{1}{N^2}\left(r_{zk}^2 + r_{jk}r_{zk} - 2g_{zjk} + \delta_{zj}\left(r_{zk} - 2r_{zk}r_{jk} + r_{jk}\right)\right) \\
&= \frac{1}{N^2}\left(r_{zk} - g_{zjk} + \delta_{zj}\left(r_{zk} - 2r_{zk}r_{jk} + r_{jk}\right)\right) \\
\therefore \mathbb{E}[\Delta g_{zJK}] &\approx \frac{1}{N^2}\left(r_{zK} - g_{zJK}\right), \\
\mathbb{E}[\Delta g_{ijk}] &= \frac{1}{N^2}\left(r_{ik}r_{zk} + r_{jk}r_{zk} - 2g_{ijk} + \delta_{ij}\left(r_{zk} - 2r_{zk}r_{ik} + r_{ik}\right)\right) \\
&= \frac{1}{N^2}\left(g_{zik} + g_{zjk} - 2g_{ijk} + \delta_{ij}\left(r_{zk} - 2g_{zik} + r_{ik}\right)\right) \\
\therefore \mathbb{E}[\Delta g_{IJK}] &= \frac{1}{N^2}\left(g_{zIK} + g_{zJK} - 2g_{IJK} + \frac{f_I}{N}\delta_{IJ}\left(r_{zK} - 2g_{zIK} + r_{IK}\right)\right) \\
&\approx \frac{1}{N^2}\left(g_{zIK} + g_{zJK} - 2g_{IJK}\right).
\end{aligned}$$

Again, we drop the δ terms because they introduce small corrections that are swamped by the dominant balance of the preceding terms.

If gossip with the source occurs at rate γ and interactions occur at (normalized) rate 1, the dynamics of reputations and agreement are described by the ODEs

$$\begin{aligned}
\frac{dr_{zJ}}{d\phi} &= \frac{1}{N^2}\left(p_{zJ\bullet} - r_{zJ}\right), \\
\frac{dr_{IJ}}{d\phi} &= \frac{1}{N^2}\left(p_{IJ\bullet} - r_{IJ}\right) + \frac{\gamma}{N^2}\left(r_{zJ} - r_{IJ}\right), \\
\frac{dg_{zJK}}{d\phi} &= \frac{1}{N^2}\left(r_{zK}p_{JK\bullet} + r_{JK}p_{zK\bullet} - 2g_{zJK}\right) + \gamma\frac{1}{N^2}\left(r_{zK} - g_{zJK}\right), \\
\frac{dg_{IJK}}{d\phi} &= \frac{1}{N^2}\left(r_{IK}p_{JK\bullet} + r_{JK}p_{IK\bullet} - 2g_{IJK}\right) + \gamma\frac{1}{N^2}\left(g_{zIK} + g_{zJK} - 2g_{IJK}\right).
\end{aligned}$$

We introduce the notation $\zeta_j = r_{zj}$, $Z_j = p_{zj\bullet}$, and $\chi_{ij} = g_{zij}$; they are, respectively, the source's opinion of z, the expected value of the source's opinion of j after an observation, and an indicator variable for whether or not the source and i agree that j is good. These, and the other terms, can be rewritten in terms of s by observing that there is no dependence on the strategic identity of the observer, except transiently. The form of Z_s is

$$\begin{aligned}
Z_s &= \chi\nu_s^A + \zeta\nu_s^B + r\nu_s^C + \nu_s^D \\
&= p_s + (\chi - g)\nu_s^A + (\zeta - r)\nu_s^B, \text{with} \\
\chi &= \chi_\bullet, \\
\zeta &= \zeta_\bullet,
\end{aligned}$$

and the steady-state conditions are given by setting

$$\begin{aligned}
\frac{d\zeta_s}{d\phi} &= \frac{1}{N^2}\left(Z_s - \zeta_s\right), \\
\frac{dr_s}{d\phi} &= \frac{1}{N^2}\left(p_s - r_s + \gamma\left(\zeta_s - r_s\right)\right), \\
\frac{d\chi_s}{d\phi} &= \frac{1}{N^2}\left(\zeta_s p_s + r_s Z_s - 2\chi_s + \gamma\left(\zeta_s - \chi_s\right)\right), \\
\frac{dg_s}{d\phi} &= \frac{2}{N^2}\left(r_s p_s - g_s + \gamma\left(\chi_s - g_s\right)\right).
\end{aligned}$$

to zero. These yield

$$\begin{aligned}
\zeta_s &= Z_s, \\
r_s &= \frac{p_s + \gamma Z_s}{1+\gamma}, \\
\chi_s &= \frac{Z_s\left(p_s(2+\gamma) + \gamma(1 + Z_s + \gamma)\right)}{(1+\gamma)(2+\gamma)}, \\
g_s &= \frac{p_s^2(2+\gamma) + 2p_s Z_s\gamma(2+\gamma) + Z_s\gamma^2(1+Z_s+\gamma)}{(1+\gamma)^2(2+\gamma)}.
\end{aligned}$$

At this point we change variables slightly. We define $\iota = 1 - 1/(1+\gamma)$, so that $\gamma = 1/(1-\iota) - 1$; sending $\gamma \to 0$ yields $\iota = 0$, and sending $\gamma \to \infty$ yields $\iota = 1$. The preceding expressions then become

$$\begin{aligned}
\zeta_s &= Z_s, \\
r_s &= p_s(1-\iota) + Z_s\iota, \\
\chi_s &= Z_s\left(p_s(1-\iota) - \iota\frac{Z_s(1-\iota)+1}{2-\iota}\right), \\
g_s &= p_s^2(1-\iota)^2 + 2p_s Z_s\iota(1-\iota) + Z_s\iota^2\left(\frac{Z_s(1-\iota)+1}{2-\iota}\right).
\end{aligned}$$

Finally, we obtain

$$g_s - r_s^2 = \left(\frac{\iota^2}{2-\iota}\right) Z_s(1 - Z_s). \tag{S12}$$

This is the difference between full agreement about strategy s (so that $g_s = r_s$) and fully private assessment (so that $g_s = r_s^2$), which correspond to $\iota = 1$ and $\iota = 0$, respectively.

We now return to the expression $\tilde{g}_2$ from the main text (Eq. 19):

$$\begin{aligned}\tilde{g}_2 &= \left(1-Q^2\right)\cdot g_2 + Q^2\cdot r' \\ &= \left(1-Q^2\right)\cdot \sum_{s\in S} f_s (r'_s)^2 + Q^2\cdot \sum_{s\in S} f_s r'_s \\ &= \sum_{s\in S} f_s \left((r'_s)^2 + Q^2\left(r'_s - (r'_s)^2\right)\right).\end{aligned}$$

(We have used r'_s to refer to the average reputation of strategy s denoted r_s in Eq. 19, to avoid confusion with the values of r_s used here.) Denote by $\tilde{g}_{2,s}$ the component of $\tilde{g}_2$ that is due to strategy s, divided by f_s (so that, when multiplied by f_s and summed, we recover $\tilde{g}_2$). Then we have (after pulling out a power of r'_s from the second term)

$$\begin{aligned}\tilde{g}_{2,s} &= (r'_s)^2 + Q^2 r'_s\left(1-r'_s\right) \\ \therefore \tilde{g}_{2,s} - (r'_s)^2 &= Q^2 r'_s\left(1-r'_s\right).\end{aligned}$$

This is the same as Eq. S12 under the transformation

$$\begin{aligned}Q^2 r'_s\left(1-r'_s\right) &= \left(\frac{\iota^2}{2-\iota}\right) Z_s(1-Z_s) \\ \therefore Q^2 &= \frac{\iota^2}{2-\iota} \\ &= \frac{\gamma^2}{(1+\gamma)(2+\gamma)}, \\ r'_s &= Z_s.\end{aligned}$$

Thus, the continuous-time process described here, where interactions *and* consultation of the gossip source occur on overlapping time scales, can easily be rescaled into the discrete-time process outlined in the main text, where interactions and gossip are considered as discrete processes. The reputations used as the "input" for the gossip phase of the discrete-time process are precisely the steady-state equilibrium reputations Z_s in the eyes of the gossip source z. This procedure may be iterated to obtain comparable expressions for r_s, χ_s, and g_s.

5. The "leading eight" norms

In this section, we extend our analysis of gossip and indirect reciprocity to the "leading eight" social norms (10), which are labelled $L1$–$L8$. Two of the "leading eight" norms have already been analyzed—$L3$ is Simple Standing and $L6$ is Stern Judging—whereas the remaining leading eight are third-order norms, meaning that their judgment of a donor depends upon the donor's action, the recipient's reputation, and the donor's current reputation. The resulting equations for reputation dynamics are more complicated than for third-order norms but still tractable. Our analysis below leverages the fact that a third-order norm can often be written as a mixture of two second-order norms, with a mixture rate that depends upon the current reputation of the donor. Based on this idea, we will show that the dynamics of gossip, reputations, and behavior are qualitatively similar for specific pairs of the third-order leading-eight norms: namely, the pairs $(L1, L2)$, $(L4, L5)$, and $(L7, L8)$.

This classification can be contrasted with the dynamical analysis by Fujimoto and Ohtsuki (14), which binned third-order norms into three categories according to how the rest of the population views an arbitrary player under private assessment; type 1 norms ($L1$, $L3$, $L4$, $L7$) in which most players end up with good reputations, type 2 norms ($L2$, $L5$) in which some players end up with good reputations and others end up with intermediate or bad reputations, and type 3 norms ($L6$, $L8$) in which almost everyone has either an intermediate or bad reputation. After sending the scaled gossip rate $\Gamma \to 0$, our mean-field model generally agrees with their private assessment results.

For this portion of our analysis, equations are written in terms of the scaled gossip rate Γ in the intercalated model of Section 4; however, equivalent expressions for τ in the main-text model can be derived using Eq. S11. We are aided by the fact that both models behave identically as τ or Γ go to 0 or infinity.

Under third-order norms, the rule an observer i uses to update the reputation of a donor j depends not only on j's action toward k and on r_{ik} but *also* on r_{ij}; that is, i considers j's current reputation when updating j's reputation. Thus, if i's view of j is $R_{ij} \in \{G, B\}$, i's view of k is $R_{ik} \in \{G, B\}$, and j performed action $A_{jk} \in \{C, D\}$ toward k, then the probability that i assigns j a good reputation may be denoted $p_{R_{ij}R_{ik}A_{jk}}$. Likewise, j's action toward k may depend not only on r_{jk} but on r_{jj}, i.e., j's view of themselves; their behavior is described by four values $a_j^{R_{jj}R_{jk}}$, so that the first superscript index corresponds to j's view of j's own reputation, and the second index corresponds to j's view of k. The expected value of r_{ij} after an observed interaction is given by

$$
\begin{aligned}
p_{ijk} = {} & r_{jj}r_{jk}a_j^{GG}r_{ij}r_{ik}p_{GGC} + r_{jj}r_{jk}a_j^{GG}r_{ij}(1-r_{ik})p_{GBC} \\
& + r_{jj}r_{jk}a_j^{GG}(1-r_{ij})r_{ik}p_{BGC} + r_{jj}r_{jk}a_j^{GG}(1-r_{ij})(1-r_{ik})p_{BBC} \\
& + r_{jj}r_{jk}(1-a_j^{GG})r_{ij}r_{ik}p_{GGD} + r_{jj}r_{jk}(1-a_j^{GG})r_{ij}(1-r_{ik})p_{GBD} \\
& + r_{jj}r_{jk}(1-a_j^{GG})(1-r_{ij})r_{ik}p_{BGD} + r_{jj}r_{jk}(1-a_j^{GG})(1-r_{ij})(1-r_{ik})p_{BBD} \\
& + r_{jj}(1-r_{jk})a_j^{GB}r_{ij}r_{ik}p_{GGC} + r_{jj}(1-r_{jk})a_j^{GB}r_{ij}(1-r_{ik})p_{GBC} \\
& + r_{jj}(1-r_{jk})a_j^{GB}(1-r_{ij})r_{ik}p_{BGC} + r_{jj}(1-r_{jk})a_j^{GB}(1-r_{ij})(1-r_{ik})p_{BBC} \\
& + r_{jj}(1-r_{jk})(1-a_j^{GB})r_{ij}r_{ik}p_{GGD} + r_{jj}(1-r_{jk})(1-a_j^{GB})r_{ij}(1-r_{ik})p_{GBD} \\
& + r_{jj}(1-r_{jk})(1-a_j^{GB})(1-r_{ij})r_{ik}p_{BGD} + r_{jj}(1-r_{jk})(1-a_j^{GB})(1-r_{ij})(1-r_{ik})p_{BBD} \\
& + (1-r_{jj})r_{jk}a_j^{BG}r_{ij}r_{ik}p_{GGC} + (1-r_{jj})r_{jk}a_j^{BG}r_{ij}(1-r_{ik})p_{GBC} \\
& + (1-r_{jj})r_{jk}a_j^{BG}(1-r_{ij})r_{ik}p_{BGC} + (1-r_{jj})r_{jk}a_j^{BG}(1-r_{ij})(1-r_{ik})p_{BBC} \\
& + (1-r_{jj})r_{jk}(1-a_j^{BG})r_{ij}r_{ik}p_{GGD} + r_{jj}(1-r_{jk})(1-a_j^{BG})r_{ij}(1-r_{ik})p_{GBD} \\
& + (1-r_{jj})r_{jk}(1-a_j^{BG})(1-r_{ij})r_{ik}p_{BGD} + r_{jj}(1-r_{jk})(1-a_j^{BG})(1-r_{ij})(1-r_{ik})p_{BBD} \\
& + (1-r_{jj})(1-r_{jk})a_j^{BB}r_{ij}r_{ik}p_{GGC} + (1-r_{jj})(1-r_{jk})a_j^{BB}r_{ij}(1-r_{ik})p_{GBC} \\
& + (1-r_{jj})(1-r_{jk})a_j^{BB}(1-r_{ij})r_{ik}p_{BGC} + (1-r_{jj})(1-r_{jk})a_j^{BB}(1-r_{ij})(1-r_{ik})p_{BBC} \\
& + (1-r_{jj})(1-r_{jk})(1-a_j^{BB})r_{ij}r_{ik}p_{GGD} + (1-r_{jj})(1-r_{jk})(1-a_j^{BB})r_{ij}(1-r_{ik})p_{GBD} \\
& + (1-r_{jj})(1-r_{jk})(1-a_j^{BB})(1-r_{ij})r_{ik}p_{BGD} + (1-r_{jj})(1-r_{jk})(1-a_j^{BB})(1-r_{ij})(1-r_{ik})p_{BBD}.
\end{aligned}
$$

We will shortly see that most of these terms vanish.

5.1. Analysis of the self-image. One complication is that two of the norms, $L1$ and $L2$, demand that we know the donor's "self-image" $h_j = r_{jj}$, since they feature a "repentant discriminator" strategy such that $a_j^{BB} = 1$ rather than 0; that is, if i has a bad view of j, they will defect with j, *unless* i believes themselves to be bad, as well. We abbreviate this strategy as rDISC. In principle, we would also need to track the "agreement" term $\psi_{ij} = h_j r_{ij}$, which corresponds to a donor j who views themselves as good *and* is viewed as good by an observer i. Under public reputations, it is commonly assumed that r_{jj} is the same as any other player's view of j or, equivalently, that j conditions their behavior on the (known) consensus view of j, not their own personal feelings. But when reputations are private, a decision must be made regarding how to treat the self-image. We consider two possible approaches:

1. *Consistent self-image*: Treating the self-image as good, i.e., assuming that a player always views themselves as good. This emerges naturally in models of private assessment (e.g., 12, 13); it corresponds to setting $h_j = 1$ and $\psi_{ij} = r_{ij}$.
2. *Variable self-image*: Treating the self-image like any other element of the image matrix (e.g., 14). That is, j updates their self-image any time they either "observe" their own action or any time they gossip with a third party about themselves.

We analyze the dynamics of the self-image and "self-agreement" terms in a manner similar to Eqs. S5 and S6:

$$\begin{aligned}
\mathbb{E}[\Delta h_j] &= \mathbb{E}[h_j^+] - h_j \\
&= \langle h_j + \delta_{\omega j}\delta_{\theta j}\left(\xi_{\omega\theta\rho} - h_\omega\right)\rangle_{\omega\theta\rho} - h_j \\
&= \frac{1}{N^2}\left(\xi_{jj\bullet} - h_j\right), \\
\mathbb{E}[\Delta \psi_{ij}] &= \mathbb{E}[\psi_{ij}^+] - \psi_{ij} \\
&= \langle \left(r_{ij} + \delta_{\omega i}\delta_{\theta j}\left(\xi_{\omega\theta\rho} - r_{\omega\theta}\right)\right)\left(h_j + \delta_{\omega j}\delta_{\theta j}\left(\xi_{\omega\theta\rho} - r_{\omega\theta}\right)\right)\rangle_{\omega\theta\rho} - \psi_{ij} \\
&= \frac{1}{N^2}\left\langle r_{ij}(\xi_{jj\rho} - h_j) + h_j(\xi_{ij\rho} - r_{ij}) + \delta_{ij}\left(\xi_{ij\rho} - r_{ij}\right)^2\right\rangle_\rho \\
&= \frac{1}{N^2}\left\langle r_{ij}\xi_{jj\rho} + h_j\xi_{ij\rho} - 2\psi_{ij} + \delta_{ij}(\xi_{ij\rho}^2 - 2\xi_{ij\rho}r_{ij} + r_{ij})\right\rangle_\rho \\
&= \frac{1}{N^2}\left(r_{ij}\xi_{jj\bullet} + h_j\xi_{ij\bullet} - 2\psi_{ij} + \delta_{ij}(\left\langle\xi_{ij\rho}^2\right\rangle_\rho - 2\xi_{ij\bullet}r_{ij} + r_{ij})\right) \\
&= \frac{1}{N^2}\left(r_{ij}\xi_{jj\bullet} + h_j\xi_{ij\bullet} - 2\psi_{ij} + \delta_{ij}(\left\langle\xi_{jj\rho}^2\right\rangle_\rho - 2\xi_{jj\bullet}h_j + h_j)\right).
\end{aligned}$$

In the last line, we took advantage of the δ_{ij} prefactor in the last term to send $i \to j$ ahead of time.

5.1.1. Effect of interactions with variable self-image. Just as with second-order norms, we set $\xi_{\omega\theta\rho} = p_{\omega\theta\rho}$ and obtain

$$\begin{aligned}
\mathbb{E}[\Delta h_j] &= \frac{1}{N^2}\left(p_{jj\bullet} - h_j\right), \\
\mathbb{E}[\Delta \psi_{ij}] &= \frac{1}{N^2}\left(r_{ij}p_{jj\bullet} + h_j p_{ij\bullet} - 2\psi_{ij} + \delta_{ij}(\left\langle p_{jj\rho}^2\right\rangle_\rho - 2p_{jj\bullet}h_j + h_j)\right).
\end{aligned}$$

Define

$$\begin{aligned}
p_{J\bullet}^{\S} &= \langle p_{jj\rho}\rangle_{\rho\sigma_J}, \\
p_{J\bullet}^{\S\S\S} &= \left\langle p_{jj\rho}^2\right\rangle_{\rho\sigma_J}.
\end{aligned}$$

Summing yields

$$\begin{aligned}
\mathbb{E}[\Delta h_J] &= \frac{1}{N^2}\left(p_{JJ\bullet} - h_J\right), \\
\mathbb{E}[\Delta \psi_{IJ}] &= \frac{1}{N^2}\left(r_{IJ}p_{J\bullet}^{\S} + h_J p_{IJ\bullet} - 2\psi_{IJ} + \delta_{IJ}\frac{f_I}{N}(p_{J\bullet}^{\S\S\S} - 2p_{J\bullet}^{\S}h_J + h_J)\right).
\end{aligned} \qquad \text{[S13]}$$

The δ_{IJ} term may be neglected; it yields a $\mathcal{O}(1/N)$ correction, and the dominant balance will be given by the preceding terms. Rewriting this as a differential equation for a strategic type S yields

$$\begin{aligned}
\frac{dh_s}{d\phi} &= \frac{1}{N^2}\left(p_s^{\S} - h_s\right), \\
\frac{d\psi_s}{d\phi} &= \frac{1}{N^2}\left(r_s p_s^{\S} + h_s p_s - 2\psi_s\right).
\end{aligned}$$

Thus, interactions cause the self-image h_s to tend toward $p_s^{\S}$, and they cause the "self-agreement" ψ_s to tend toward $(r_s p_s^{\S} + h_s p_s)/2$; in steady state (with no gossip), this becomes $p_s^{\S} p_s$.

norm	P_{GGC}	P_{GGD}	P_{GBC}	P_{GBD}	P_{BGC}	P_{BGD}	P_{BBC}	P_{BBD}	s^{GG}	s^{GB}	s^{BG}	s^{BB}
$L1$	1	0	1	1	1	0	1	0	1	0	1	1
$L2$	1	0	0	1	1	0	1	0	1	0	1	1
$L3$	1	0	1	1	1	0	1	1	1	0	1	0
$L4$	1	0	1	1	1	0	0	1	1	0	1	0
$L5$	1	0	0	1	1	0	1	1	1	0	1	0
$L6$	1	0	0	1	1	0	0	1	1	0	1	0
$L7$	1	0	1	1	1	0	0	0	1	0	1	0
$L8$	1	0	0	1	1	0	0	0	1	0	1	0

Table S3. The "leading eight" social norms, including both assessment rules and action rules, modeled after Table 2 of Podder et al. (11). Norm $L3$ is *Simple Standing*, and norm $L6$ is *Stern Judging*; both are symmetric with respect to the reputation of the donor and, thus, are second-order norms.

norm	p_{ijk}^{DISC} or p_{ijk}^{rDISC}	p_{ijk}^{ALLD}	p_{ijk}^{ALLC}
$L1$	$\psi_{ij}(g_{ijk} - r_{ik} - r_{jk} + 1) + h_j(r_{jk} - 1) + 1$	$r_{ij}(1 - r_{ik})$	1
$L2$	$2\psi_{ij}(g_{ijk} - r_{ik} - r_{jk} + 1) + r_{ij}(r_{ik} - 1) + h_j(r_{jk} - 1) + 1$	$r_{ij}(1 - r_{ik})$	$r_{ij}(r_{ik} - 1) + 1$
$L3$	$g_{ijk} - r_{ik} + 1$	$1 - r_{ik}$	1
$L4$	$r_{ij}(r_{jk} - g_{ijk}) + 2g_{ijk} - r_{ik} - r_{jk} + 1$	$1 - r_{ik}$	$r_{ij}(1 - r_{ik}) + r_{ik}$
$L5$	$r_{ij}(g_{ijk} - r_{jk}) + g_{ijk} - r_{ik} + 1$	$1 - r_{ik}$	$r_{ij}(r_{ik} - 1) + 1$
$L6$	$2g_{ijk} - r_{ik} - r_{jk} + 1$	$1 - r_{ik}$	r_{ik}
$L7$	$r_{ij}(1 - r_{ik}) + g_{ijk}$	$r_{ij}(1 - r_{ik})$	$r_{ij}(1 - r_{ik}) + r_{ik}$
$L8$	$r_{ij}(g_{ijk} - r_{ik} - r_{jk} + 1) + g_{ijk}$	$r_{ij}(1 - r_{ik})$	r_{ik}

Table S4. Error-free values of p_{ijk} for the leading eight social norms, when j follows the strategy listed in the superscript. In the middle column, the expressions provided are for p_{ijk}^{DISC} except for norms $L1$ and $L2$, which use the "repentant discriminator" strategy rDISC; this differs from DISC insofar as $a^{BB} = 1$ rather than 0, so that players who perceive themselves as bad cooperate even with recipients they consider bad.

norm	p_{DISC} or p_{rDISC}	p_{ALLD}	p_{ALLC}
$L1$	$\psi(g - 2r + 1) + h(r - 1) + 1$	$r_{\text{ALLD}}(1 - r)$	1
$L2$	$2\psi(g - 2r + 1) + (h + r)(r - 1) + 1$	$r_{\text{ALLD}}(1 - r)$	$1 - r_{\text{ALLC}}(1 - r)$
$L3$	$g - r + 1$	$(1 - r)$	1
$L4$	$(r - g)(r - 2) + 1$	$(1 - r)$	$r_{\text{ALLC}}(1 - r) + r$
$L5$	$(g - r)(r + 1) + 1$	$(1 - r)$	$1 - r_{\text{ALLC}}(1 - r)$
$L6$	$2g - 2r + 1$	$(1 - r)$	r
$L7$	$r(1 - r) + g$	$r_{\text{ALLD}}(1 - r)$	$r_{\text{ALLC}}(1 - r) + r$
$L8$	$r(g - 2r + 1) + g$	$r_{\text{ALLD}}(1 - r)$	r

Table S5. Population-wide generalizations of p_{ijk} in a monomorphic population of discriminators (or, for norms $L1$ and $L2$, repentant discriminators). As in our analysis of second-order norms, assessment errors may be incorporated by multiplying p by $(1 - 2u_a)$ and adding u_a. Execution errors may be incorporated for discriminators by multiplying p_{DISC} or p_{rDISC} by $(1 - u_e)$ and adding $u_e p_{\text{ALLD}}$.

5.1.2. Effect of gossip with variable self-image. As usual for gossip, we set $\xi_{\omega\theta\rho} = r_{\rho\theta}$:

$$\begin{aligned}
\mathbb{E}[\Delta h_j] &= \frac{1}{N^2}\left(r_{\bullet j} - h_j\right), \\
\mathbb{E}[\Delta \psi_{ij}] &= \frac{1}{N^2}\left(r_{ij}r_{\bullet j} + h_j r_{\bullet j} - 2\psi_{ij} + \delta_{ij}(r_{\bullet j} - 2r_{\bullet j}h_j + h_j)\right) \\
&= \frac{1}{N^2}\left(g_{i\bullet j} + \psi_{\bullet j} - 2\psi_{ij} + \delta_{ij}(r_{\bullet j} + h_j - 2\psi_{\bullet j})\right).
\end{aligned}$$

Summing yields

$$\begin{aligned}
\mathbb{E}[\Delta h_J] &= \frac{1}{N^2}\left(r_{\bullet J} - h_J\right), \\
\mathbb{E}[\Delta \psi_{IJ}] &= \frac{1}{N^2}\left(g_{I\bullet J} + \psi_{\bullet J} - 2\psi_{IJ} + \delta_{IJ}\frac{f_I}{N}(r_{\bullet J} + h_J - 2\psi_{\bullet J})\right).
\end{aligned}$$

We ignore the δ_{IJ} term, since it yields a $\mathcal{O}(1/N)$ correction and is negligible compared to the dominant balance of the preceding terms. Doing so, and rewriting this as a differential equation for a strategic type s, yields

$$\begin{aligned}
\frac{dh_s}{d\phi} &= \frac{1}{N^2}\left(r_s - h_s\right), \\
\frac{d\psi_s}{d\phi} &= \frac{1}{N^2}\left(g_s - \psi_s\right).
\end{aligned}$$

Thus, gossip causes the "self-image" h_s to tend toward r_s, the population average view of s, and it causes ψ_s to tend toward g_s.

Combining this with Eq. S13 yields

$$\begin{aligned}
\frac{dh_s}{d\phi} &= \frac{1}{N^2}\left(p_s^{\S} - h_s + \gamma\left(r_s - h_s\right)\right) \\
&= \frac{1}{N^2}\left(p_s^{\S} + \gamma r_s - (1 + \gamma)h_s\right) \\
\frac{d\psi_s}{d\phi} &= \frac{1}{N^2}\left(r_s p_s^{\S} + h_s p_s - 2\psi_s + \gamma\left(g_s - \psi_s\right)\right) \\
&= \frac{1}{N^2}\left(r_s p_s^{\S} + h_s p_s + \gamma g_s - (2 + \gamma)\psi_s\right)
\end{aligned}$$

The steady-state conditions are given by

$$h_s = \frac{p_s^{\S} + \gamma r_s}{1+\gamma}$$
$$\psi_s = \frac{r_s p_s^{\S} + h_s p_s + \gamma g_s}{2+\gamma}.$$

From our analysis of second-order norms, we know that, in order to significantly affect the reputation dynamics, $\Gamma = \gamma/N$ must be of order 1, which means γ must be of order N. Since we are working in the large population size limit, this is tantamount to requiring $\gamma \gg 1$. Thus, any appreciable amount of gossip quickly causes the self-image and self-agreement to "decay" towards r_s and g_s, respectively. As a result, for our "variable self-image" analysis, we set $h_s = r_s$ and $\psi_s = g_s$.

It is intriguing to observe that $p_s^{\S}$ is not guaranteed to be 1, even in the absence of errors. For example, under both norms $L7$ and $L8$, we have $p_{jjk} = h_j(1 - r_{jk}) + r_{jk}$. In each case, a player who regards themselves as good ($h_i = 1$) but interacts with a bad player ($r_{jk} = 0$) will continue to view themselves as good, but a player who regards themselves as bad ($h_i = 0$) and interacts with a bad player ($r_{jk} = 0$) will regard themselves as bad. This curious feature was also noted by ref. (14).

A natural extension would be to posit that the self-image h_s updates more slowly in response to gossip than the general image r_s, for example via a "stubbornness" parameter that controls how resistant individuals are to updating their own self-image. Setting the "stubbornness" parameter to 0 would yield variable self-image, wherein h_s updates just as quickly as r_s and thus decays to r_s almost immediately in the presence of even modest gossip; setting it to 1 would yield consistent self-image, i.e., $h_s = 1$. We restrict ourselves to these two regimes, since they are the limiting cases.

5.2. Invasibility by defectors. Most of the leading eight norms—$L3$ through $L8$—employ DISC as an action rule. In a monomorphic population of DISC, we have $r = r_{\text{DISC}}$ and

$$\pi_{\text{DISC}} = (b-c)r,$$
$$\pi_{\text{ALLD}} = br_{\text{ALLD}},$$

so that the critical b/c needed to resist invasion is

$$\left(\frac{b}{c}\right)^{\text{crit}} = \frac{r}{r - r_{\text{ALLD}}}.$$

The exceptions are norms $L1$ and $L2$, which employ rDISC as an action rule. Under these norms, we have $r = r_{\text{rDISC}}$, and the self-image $h = h_{\text{DISC}}$ becomes relevant. For an arbitrary recipient following strategy s, rDISC cooperates any time they interact with a recipient they view as good (probability r_s) *and* any time they interact with a recipient they view as bad, provided they view themselves as bad, too (probability $(1-h)(1-r_s)$). Thus

$$\pi_{\text{rDISC}} = (b-c)\left(r + (1-h)(1-r)\right),$$
$$\pi_{\text{ALLD}} = b\left(r_{\text{ALLD}} + (1-h)(1-r_{\text{ALLD}})\right);$$

the critical b/c needed to resist invasion is

$$\left(\frac{b}{c}\right)^{\text{crit}} = \frac{1-h(1-r)}{h(r-r_{\text{ALLD}})} = \begin{cases} \dfrac{r}{r-r_{\text{ALLD}}} & \text{consistent self-image,} \\ \dfrac{1-r(1-r)}{r(r-r_{\text{ALLD}})} & \text{variable self-image.} \end{cases} \qquad \text{[S14]}$$

The expression for consistent self-image is the same as for the other leading eight norms, which employ DISC. This is because, with consistent self-image, DISC and rDISC behave identically. Solutions for r_{ALLD} can be obtained via the expressions in Table S5.

For the remainder of this analysis, we neglect execution error, but it can be implemented as in Section 4.2. In Fig. S8, we present the critical gossip length τ^*_{ALLD} needed to stabilize cooperation for a fixed value of b/c for each of the third-order leading eight norms.

5.3. $L1$ and $L2$ norms. The $L1$ and $L2$ norms feature a "repentant discriminator" strategy, which we abbreviate rDISC; it has $a^{BB} = 1$ rather than 0. It may be decomposed as follows:

1. If j views themselves as good ($h_j = 1$), they play DISC.
2. If j views themselves as bad ($h_j = 0$), they play ALLC.

The assessment rule of both of these norms may likewise be summarized in terms of second-order norms as follows:

1. If i views j as *bad*, then i judges j according to the Scoring norm (under both $L1$ and $L2$).
2. If i views j as *good*, then i judges j according to Simple Standing (under $L1$) or Stern Judging (under $L2$).

Putting these together yields

$$\begin{aligned}
p_{ijk}^{\text{rDISC},L1} &= r_{ij}\left(h_j p_{\text{DISC}}^{SS} + (1-h_j)p_{\text{ALLC}}^{SS}\right) + (1-r_{ij})\left(h_j p_{\text{DISC}}^{SC} + (1-h_j)p_{\text{ALLC}}^{SC}\right) \\
&= \psi_{ij}p_{\text{DISC}}^{SS} + (r_{ij}-\psi_{ij})p_{\text{ALLC}}^{SS} + (h_j-\psi_{ij})p_{\text{DISC}}^{SC} + (1-r_{ij}-h_j+\psi_{ij})p_{\text{ALLC}}^{SC} \\
&= \psi_{ij}(p_{\text{DISC}}^{SS} - p_{\text{ALLC}}^{SS} - p_{\text{DISC}}^{SC} + p_{\text{ALLC}}^{SC}) + r_{ij}(p_{\text{ALLC}}^{SS} - p_{\text{ALLC}}^{SC}) + h_j(p_{\text{DISC}}^{SC} - p_{\text{ALLC}}^{SC}) + p_{\text{ALLC}}^{SC} \\
&= \psi_{ij}(g_{ijk} - r_{ik} - r_{jk} + 1) + h_j(r_{jk} - 1) + 1, \\
p_{ijk}^{\text{rDISC},L2} &= r_{ij}\left(h_j p_{\text{DISC}}^{SJ} + (1-h_j)p_{\text{ALLC}}^{SJ}\right) + (1-r_{ij})\left(h_j p_{\text{DISC}}^{SC} + (1-h_j)p_{\text{ALLC}}^{SC}\right) \\
&= \psi_{ij}p_{\text{DISC}}^{SJ} + (r_{ij}-\psi_{ij})p_{\text{ALLC}}^{SJ} + (h_j-\psi_{ij})p_{\text{DISC}}^{SC} + (1-r_{ij}-h_j+\psi_{ij})p_{\text{ALLC}}^{SC} \\
&= \psi_{ij}(p_{\text{DISC}}^{SJ} - p_{\text{ALLC}}^{SJ} - p_{\text{DISC}}^{SC} + p_{\text{ALLC}}^{SC}) + r_{ij}(p_{\text{ALLC}}^{SJ} - p_{\text{ALLC}}^{SC}) + h_j(p_{\text{DISC}}^{SC} - p_{\text{ALLC}}^{SC}) + p_{\text{ALLC}}^{SC} \\
&= 2\psi_{ij}(g_{ijk} - r_{ik} - r_{jk} + 1) + r_{ij}(r_{ik} - 1) + h_j(r_{jk} - 1) + 1.
\end{aligned}$$

These expressions correspond to population-level averages

$$\begin{aligned}
p_{\text{rDISC}}^{L1} &= \psi(g - 2r + 1) + h(r-1) + 1, \\
p_{\text{rDISC}}^{L2} &= 2\psi(g - 2r + 1) + (h+r)(r-1) + 1
\end{aligned} \tag{S15}$$

in a monomorphic population of rDISC.

5.3.1. Consistent self-image. When the "self-image" is always considered good, the equilibrium reputations in Eq. S15 become

$$\begin{aligned}
p_{\text{rDISC}}^{L1} &= r(g - 2r + 2), \\
p_{\text{rDISC}}^{L2} &= r(2g - 3r + 2).
\end{aligned}$$

We introduce assessment error by setting $r = p(1-2u_a) + u_a$; see Section 4.2. Combining this with $g = r(r+\Gamma)/(1+\Gamma)$ (Eq. S7) allows us to show that the general solutions for r are the real-valued solutions, between 0 and 1, of the following cubic equations:

$$r = \begin{cases} \left(\dfrac{1-2u_a}{1+\Gamma}\right)r^3 - \left(\dfrac{(2+\Gamma)(1-2u_a)}{1+\Gamma}\right)r^2 + 2(1-2u_a)r + u_a & L1, \\ 2\left(\dfrac{1-2u_a}{1+\Gamma}\right)r^3 - \left(\dfrac{(3+\Gamma)(1-2u_a)}{1+\Gamma}\right)r^2 + 2(1-2u_a)r + u_a & L2. \end{cases} \tag{S16}$$

In general, the solutions to these equations cannot be expressed in terms of radicals. With no errors, they simplify to

$$r = \begin{cases} \left(\dfrac{1}{1+\Gamma}\right)r^3 - \left(\dfrac{2+\Gamma}{1+\Gamma}\right)r^2 + 2r & L1, \\ 2\left(\dfrac{1}{1+\Gamma}\right)r^3 - \left(\dfrac{3+\Gamma}{1+\Gamma}\right)r^2 + 2r & L2. \end{cases}$$

which do have some rational roots; for $L1$, the roots are $r = \{0, 1, 1+\Gamma\}$, and for $L2$, they are $r = \{0, 1, (1+\Gamma)/2\}$. In general, the 1 corresponds to the "physical" solution of these equations, i.e., the stable equilibrium of the reputation dynamics. When the gossip rate Γ is zero (i.e., purely private assessment), we have

$$r = \begin{cases} (1-2u_a)r^3 - 2(1-2u_a)r^2 + 2(1-2u_a)r + u_a & L1, \\ 2(1-2u_a)r^3 - 3(1-2u_a)r^2 + 2(1-2u_a)r + u_a & L2. \end{cases}$$

The equation for $L2$ has an analytical solution on the interval $[0, 1]$, namely

$$r = 1/2$$

independent of u_a, just like Stern Judging ($L6$). This is sensible; good players are judged according to Stern Judging, which has $r = 1/2$ for any strategy s, and bad players are judged according to Scoring, so they gain a good reputation only by cooperating–which they will do, on average, half the time. Finally, sending the gossip rate $\Gamma \to \infty$ yields

$$r = -(1-2u_a)r^2 + 2(1-2u_a)r + u_a$$

for *both* $L1$ and $L2$, which has the solution

$$r = \frac{1 - 4u_a + \sqrt{1 - 4u_a + 8u_a^2}}{2(1-2u_a)} = \frac{1}{2} + \frac{\sqrt{1 - 4u_a + 8u_a^2} - 2u_a}{2(1-2u_a)}.$$

We now obtain the condition for stable cooperation. With assessment error, we have

$$\begin{aligned} r_{\text{ALLD}} &= r_{\text{ALLD}}(1-r)(1-2u_a) + u_a \\ &= \frac{u_a}{r(1-2u_a)+2u_a}. \end{aligned}$$

This may be used in conjunction with Eq. S14 to obtain the general condition $(b/c)^{\text{crit}}$ for stable cooperation—or, equivalently, the critical gossip rate Γ needed to stabilize cooperation for a given value of b/c. We present this result in terms of $\tau = \log(1+\Gamma)$, consistent with the main text, in figure Fig. S8. Surprisingly, both $L1$ and $L2$ easily admit stable cooperation even under *private* assessment (Γ or $\tau \to 0$), a result consistent with ref. (14). With fast gossip ($\Gamma \to \infty$), the condition for stable cooperation under both $L1$ and $L2$ becomes

$$\left(\frac{b}{c}\right)^{\text{crit}} = \frac{1+\sqrt{1-4u_a+8u_a^2}}{2(1-2u_a)}.$$

Under private assessment ($\Gamma \to 0$), it is intriguing to note that, while we have $r_{\text{DISC}} = 1/2$ under $L2$, cooperation is nonetheless possible. This is because r_{ALLD} is much less than $1/2$. This, in turn, is because bad individuals are judged according to the Scoring norm, under which defection is always bad; thus, defectors who end up with a bad reputation continue to have a bad reputation indefinitely (barring errors). In contrast, discriminators who end up with a bad reputation have the chance to ameliorate their moral standing by interacting with someone they view as good; they cooperate and thus are assigned a good reputation.

5.3.2. Variable self-image. When the "self-image" is treated like any other element of the image matrix, the expressions for p_{rDISC} become

$$\begin{aligned} p_{\text{rDISC}}^{L1} &= g(g-2r+1) - r(1-r) + 1, \\ p_{\text{rDISC}}^{L2} &= 2g(g-2r+1) - 2r(1-r) + 1. \end{aligned}$$

After introducing assessment error, the solutions are the real-valued roots (between 0 and 1) of the following quartic equations:

$$r = \begin{cases} \left(\frac{1-2u_a}{(1+\Gamma)^2}\right) r^4 - 2\left(\frac{1-2u_a}{(1+\Gamma)^2}\right) r^3 + \left(\frac{(1-2u_a)(2+\Gamma)}{(1+\Gamma)^2}\right) r^2 - \left(\frac{1-2u_a}{1+\Gamma}\right) r + 1 - u_a & L1, \\ 2\left(\frac{1-2u_a}{(1+\Gamma)^2}\right) r^4 - 4\left(\frac{1-2u_a}{(1+\Gamma)^2}\right) r^3 + 2\left(\frac{(1-2u_a)(2+\Gamma)}{(1+\Gamma)^2}\right) r^2 - 2\left(\frac{1-2u_a}{1+\Gamma}\right) r + 1 - u_a & L2. \end{cases}$$

When $u_a = 0$, both of these equations generally have $r = 1$ as a root, independent of Γ. However, the $r = 1$ solution is unstable against errors; it becomes non-physical for positive error rates. The $\Gamma \to 0$ limit, like the general case, does not appear to admit analytical solutions in terms of radicals. For $\Gamma \to \infty$, we have

$$r = 1 - u_a$$

for *both* $L1$ and $L2$; this can easily be seen from the fact that all of the non-constant terms have higher powers of Γ in the denominator than in the numerator. The condition for stability against defectors can be found via Eq. S14. A minor difference is that, under both $L1$ and $L2$, it is more difficult to sustain cooperation under variable self-image, especially for low gossip rates Γ. (For this portion of our analysis, we present the $\Gamma \to 0$ solutions as a heuristic, but we caution the reader that our variable self-image model requires that the self-image decay quickly—which means there must indeed be some gossip; Γ is really roughly of order $1/N$.)

The reason is subtle. On the one hand, rDISC enjoys a higher reputation under variable self-image than under consistent self-image; for example, with $u_a = u_e = 0.02$ and $\Gamma = 0$, we have $r_{\text{rDISC}} \approx 0.64$ under variable self-image, compared to $r_{\text{DISC}} = 1/2$ under consistent self-image. The reputation of ALLD is similar in both cases, approximately 0.244 and 0.260 under variable and consistent self-image, respectively. However, the cooperation rate toward ALLD is different. Under consistent self-image, the cooperation rate from DISC to ALLD is simply r_{ALLD}; under variable self-image, the cooperation rate from rDISC to ALLD is $r_{\text{ALLD}} + (1 - r_{\text{rDISC}})(1 - r_{\text{ALLD}}) \approx 0.517$, whereas for rDISC to rDISC, it is $r_{\text{rDISC}} + (1 - r_{\text{rDISC}})^2 \approx 0.769$. That is, under consistent self-image, DISC cooperates with ALLD about half as often as it cooperates with other DISC (0.260 versus $1/2$). Under variable self-image, it is more like two-thirds as often (0.517 versus 0.769).

5.4. $L3$ and $L6$ norms. $L3$ is Simple Standing and $L6$ is Stern Judging. Both are addressed in the main text and Section 2.2.

5.5. $L4$ and $L5$ norms. The assessment rule of $L4$ and $L5$ of these norms may be summarized in terms of second-order norms as follows. Both $L4$ and $L5$ judge some donors according to Stern Judging and some according to Simple Standing. For $L4$, good donors are judged according to Simple Standing and bad donors according to Stern Judging. For $L5$, it is the reverse. Thus

$$\begin{aligned} p_{ijk}^{\text{DISC},L4} &= r_{ij} p_{ijk}^{\text{DISC},SS} + (1-r_{ij}) p_{ijk}^{\text{DISC},SJ} \\ &= r_{ij}(g_{ijk} - r_{ik} + 1) + (1-r_{ij})(2g_{ijk} - r_{ik} - r_{jk} + 1) \\ &= r_{ij}(r_{jk} - g_{ijk}) + 2g_{ijk} - r_{ik} - r_{jk} + 1, \\ p_{ijk}^{\text{DISC},L5} &= r_{ij} p_{ijk}^{\text{DISC},SJ} + (1-r_{ij}) p_{ijk}^{\text{DISC},SS} \\ &= r_{ij}(2g_{ijk} - r_{ik} - r_{jk} + 1) + (1-r_{ij})(g_{ijk} - r_{ik} + 1) \\ &= r_{ij}(g_{ijk} - r_{jk}) + g_{ijk} - r_{ik} + 1. \end{aligned}$$

Accordingly, we obtain the population-level averages

$$p^{L4}_{\text{DISC}} = (r-g)(r-2)+1,$$
$$p^{L5}_{\text{DISC}} = (g-r)(r+1)+1.$$

Using $g = r(r+\Gamma)/(1+\Gamma)$ and introducing assessment error yields

$$r = \begin{cases} -\left(\frac{1-2u_a}{1+\Gamma}\right)r^3 + 3\left(\frac{1-2u_a}{1+\Gamma}\right)r^2 - 2\left(\frac{1-2u_a}{1+\Gamma}\right)r + 1 - u_a & L4, \\ \left(\frac{1-2u_a}{1+\Gamma}\right)r^3 - \left(\frac{1-2u_a}{1+\Gamma}\right)r + 1 - u_a & L5. \end{cases} \quad [S17]$$

These cubic equations do not appear to admit simple, general solutions, even with $\Gamma \to 0$ or $u_a \to 0$; one exception is that, for $L5$, sending both u_a and Γ to 0 yields

$$r = \frac{\sqrt{5}-1}{2},$$

which is 1 divided by the golden ratio. For $\Gamma \to \infty$, we find

$$r = 1 - u_a$$

for both $L4$ and $L5$. This is to be expected; both Stern Judging and Simple Standing have exactly this equilibrium reputation for DISC under public assessment, and both $L4$ and $L5$ are just combinations of these two norms. Generally, $L5$ admits less cooperation under fully private assessment ($\Gamma \to 0$) than $L4$ does and approaches the asymptotic limit more slowly as Γ increases. This is reasonable, as $L5$ judges good individuals according to Stern Judging, under which it is notoriously difficult to retain a good reputation under private assessment (equilibrium $r = 1/2$ irrespective of strategy); thus, good individuals are easily reshuffled into being labeled as bad. In contrast, maintaining a good reputation under Simple Standing (which $L4$ uses to judge good individuals) is easier.

The stability of DISC against invasion by ALLD may be probed via

$$r_{\text{ALLD}} = (1-2u_a)(1-r) + u_a,$$
$$\left(\frac{b}{c}\right)^{\text{crit}} = \frac{r}{r - r_{\text{ALLD}}} = \frac{r}{(2r-1)(1-u_a)},$$

provided of course that the denominator is positive—which requires $r > 1/2$, a condition that is generally satisfied. The critical value of b/c to prevent invasion of DISC by ALLD admits no general closed-form solution, but as $\Gamma \to \infty$, we obtain

$$\left(\frac{b}{c}\right)^{\text{crit}} = \frac{1}{1-2u_a};$$

once again, this is the same condition as under both Stern Judging and Simple Standing. $L5$ does not suffer the same difficulty as $L6$ (Stern Judging) under private assessment; even with high error rates, cooperation can be stabilized. The relationship between b/c and the length of gossip τ^*_{ALLD} needed to stabilize cooperation against invasion by defectors can be seen in Fig. S8B.

5.6. $L7$ and $L8$ norms. The assessment rule of both of these norms may be summarized in terms of second-order norms as follows. Both norms judge bad donors according to Shunning. However, $L7$ judges good donors according to Simple Standing; $L8$ judges good donors according to Stern Judging. Thus

$$\begin{aligned} p^{\text{DISC},L7}_{ijk} &= r_{ij}p^{SS}_{\text{DISC}} + (1-r_{ij})p^{SH}_{\text{DISC}} \\ &= r_{ij}(g_{ijk} - r_{ik} + 1) + (1-r_{ij})g_{ijk} \\ &= r_{ij}(1-r_{ik}) + g_{ijk}, \\ p^{\text{DISC},L8}_{ijk} &= r_{ij}p^{SJ}_{\text{DISC}} + (1-r_{ij})p^{SH}_{\text{DISC}} \\ &= r_{ij}(2g_{ijk} - r_{ik} - r_{jk} + 1) + (1-r_{ij})g_{ijk} \\ &= r_{ij}(g_{ijk} - r_{ik} - r_{jk} + 1) + g_{ijk}. \end{aligned}$$

In a monomorphic population of DISC, we have the population-level averages

$$p^{L7}_{\text{DISC}} = r(1-r) + g,$$
$$p^{L8}_{\text{DISC}} = r(g - 2r + 1) + g;$$

incorporating assessment error and using $g = r(r+\Gamma)/(1+\Gamma)$ yields the polynomial equations

$$r = \begin{cases} -\left(\frac{(1-2u_a)\Gamma}{1+\Gamma}\right) r^2 + 2\left(\frac{(1-2u_a)(1+2\Gamma)}{1+\Gamma}\right) + u_a & L7, \\ \left(\frac{1-2u_a}{1+\Gamma}\right) r^3 - (1-2u_a) r^2 + \frac{(1-2u_a)(1+2\Gamma)}{1+\Gamma} r + u_a & L8. \end{cases} \quad [S18]$$

The expression for $L7$ is quadratic in r; broadly speaking, the fact that good individuals are judged according to Simple Standing, whereas bad individuals are judged according to Shunning, cancels out a prospective factor of rg, so no terms of order r^3 appear. This equation admits a simple closed form solution:

$$r = \frac{1}{2} + \frac{\sqrt{\Gamma^2(1-4u_a+8u_a^2) + 4u_a^2(1+2\Gamma)} - 2u_a(1+\Gamma)}{2(1-2u_a)\Gamma} \to \begin{cases} \frac{1}{2} & (\Gamma \to 0), \\ \frac{1}{2} + \frac{\sqrt{1-4u_a+8u_a^2} - 2u_a}{2(1-2u_a)} & (\Gamma \to \infty). \end{cases}$$

No such closed form expression appears to be possible for $L8$, though there are some special cases:

$$r \to \begin{cases} \frac{1}{2} & (\Gamma = 1/2), \\ \frac{1}{2} + \frac{\sqrt{1-4u_a+8u_a^2} - 2u_a}{2(1-2u_a)} & (\Gamma \to \infty), \end{cases}$$

i.e., the limiting behavior as $\Gamma \to \infty$ is identical to that of $L7$, but the approach is slower. Note that the top condition is indeed $\Gamma = 1/2$, not 0. Note, further, that the $\Gamma \to \infty$ conditions are identical to those of $L1$ and $L2$.

It is instructive to observe that the solution for $L8$ for small Γ behaves similarly to the Shunning norm, under which

$$r = \frac{1 + 2u_a\Gamma - \sqrt{1-4u_a+4u_a^2(2+2\Gamma+\Gamma^2)}}{2(1-2u_a)} \to \begin{cases} \frac{1-\sqrt{1-4u_a+8u_a^2}}{2(1-2u_a)} & (\Gamma \to 0), \\ \frac{1}{2} & (\Gamma \to \infty). \end{cases}$$

That is, the cooperation rate under private Shunning *increases* with u_a; with private assessment, it is almost impossible for DISC to maintain a good reputation other than by accident. Errors perturb the equilibrium reputation away from 0, thus creating a small subpopulation of good individuals with whom one may cooperate to earn a good reputation. It is sensible that $L8$ behaves this way under private assessment. Good individuals are judged according to Stern Judging but bad individuals according to Shunning. Once players have fallen into having a bad reputation, it is difficult for them to climb out other than by accident. For $r < 1/2$, the majority of individuals are considered bad, but errors perturb this fraction toward $1/2$.

Finally, we obtain the b/c condition needed for resistance to invasion by defectors. We have

$$\begin{aligned} p_{ijk}^{\text{ALLD},L7} &= r_{ij} p_{\text{ALLD}}^{SS} + (1-r_{ij}) p_{\text{ALLD}}^{SH} \\ &= r_{ij}(1-r_{ik}), \\ p_{ijk}^{\text{ALLD},L8} &= r_{ij} p_{\text{ALLD}}^{SJ} + (1-r_{ij}) p_{\text{ALLD}}^{SH} \\ &= r_{ij}(1-r_{ik}), \end{aligned}$$

which yields

$$\begin{aligned} r_{\text{ALLD}} &= r_{\text{ALLD}}(1-r)(1-2u_a) + u_a \\ &= \frac{u_a}{r(1-2u_a)+u_a}. \end{aligned}$$

The b/c condition becomes

$$\left(\frac{b}{c}\right)^{\text{crit}} = \frac{r^2(1-2u_a) + 2ru_a}{r^2(1-2u_a) + u_a(2r-1)}.$$

Surprisingly, the condition for $L7$ saturates almost immediately; we have

$$\left(\frac{b}{c}\right)^{\text{crit}} \to \begin{cases} \frac{1+2u_a}{1-2u_a} & (\Gamma \to 0), \\ \frac{1+\sqrt{1-4u_a+8u_a^2}}{2(1-2u_a)} & (\Gamma \to \infty), \end{cases}$$

so that even with no gossip, $L7$ supports stable cooperation. $L8$ does not, but as $\Gamma \to \infty$, the b/c condition saturates at the same value as under $L7$. This is unsurprising; both $L7$ and $L8$, by effectively applying the Shunning norm to bad players, make it difficult for ALLD to gain a foothold. $L7$ is more successful at stabilizing cooperation for low Γ because it is relatively tolerant of errors and disagreement among those already considered good; $L8$, by contrast, is not.

References

1. AL Radzvilavicius, AJ Stewart, JB Plotkin, Evolution of empathetic moral evaluation. *eLife* **8**, e44269 (2019).
2. FP Santos, JM Pacheco, FC Santos, Evolution of cooperation under indirect reciprocity and arbitrary exploration rates. *Sci. Reports* **6**, 37517 (2016).
3. FP Santos, FC Santos, JM Pacheco, Social norms of cooperation in small-scale societies. *PLoS Comput. Biol*. **12**, e1004709 (2016).
4. P Tataru, T Bataillon, A Hobolth, Inference under a wright-fisher model using an accurate beta approximation. *Genetics* **201**, 1133–1141 (2015).
5. P Tataru, M Simonsen, T Bataillon, A Hobolth, Statistical inference in the Wright–Fisher model using allele frequency data. *Syst. Biol*. **66**, e30–e46 (2017).
6. S Lee, Y Murase, SK Baek, Local stability of cooperation in a continuous model of indirect reciprocity. *Sci. Reports* **11**, 14225 (2021).
7. S Lee, Y Murase, SK Baek, A second-order stability analysis for the continuous model of indirect reciprocity. *J. Theor. Biol*. **548**, 111202 (2022).
8. Y Mun, SK Baek, Second-order effects of mutation in a continuous model of indirect reciprocity. *Eur. Phys. J. Special Top*. (2023).
9. C Hilbe, L Schmid, J Tkadlec, K Chatterjee, MA Nowak, Indirect reciprocity with private, noisy, and incomplete information. *Proc. Natl. Acad. Sci*. **115**, 12241–12246 (2018).
10. H Ohtsuki, Y Iwasa, How should we define goodness? - reputation dynamics in indirect reciprocity. *J. Theor. Biol*. **231**, 107–120 (2004).
11. S Podder, R Simone, K Takács, Local reputation, local selection, and the leading eight norms. *Sci. Reports* **11**, 16560 (2021).
12. K Oishi, S Miyano, K. Kaski, T Shimada, Balanced-imbalanced transitions in indirect reciprocity dynamics on networks. *Phys. Rev. E* **104**, 024310 (2021).
13. L Schmid, P Shati, C Hilbe, K Chatterjee, The evolution of indirect reciprocity under action and assessment generosity. *Sci. Reports* **11**, 17443 (2021).
14. Y Fujimoto, H Ohtsuki, Evolutionary stability of cooperation by the leading eight norms in indirect reciprocity under noisy and private assessment. arXiv [Preprint]. https://arxiv.org/abs/2310.12581 (accessed 4 April 2024).

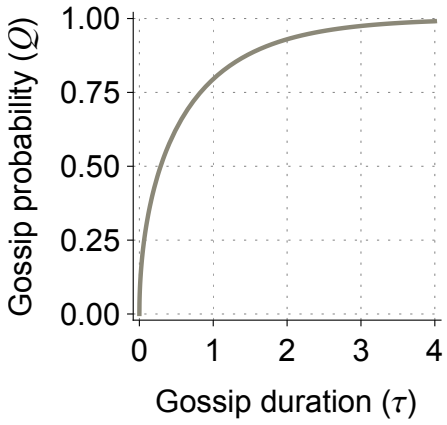

Fig. S1. Relationship between gossip with a single source versus peer-to-peer gossip. These two distinct gossip processes have the same effects on the level of agreement and equilibrium reputations in the population under a suitable transformation of parameters. We plot the transformation $\tau = -\log(1 - Q^2)$ between the duration of gossip τ in the peer-to-peer process and the probability Q of consulting the single gossip source (Eq. 7).

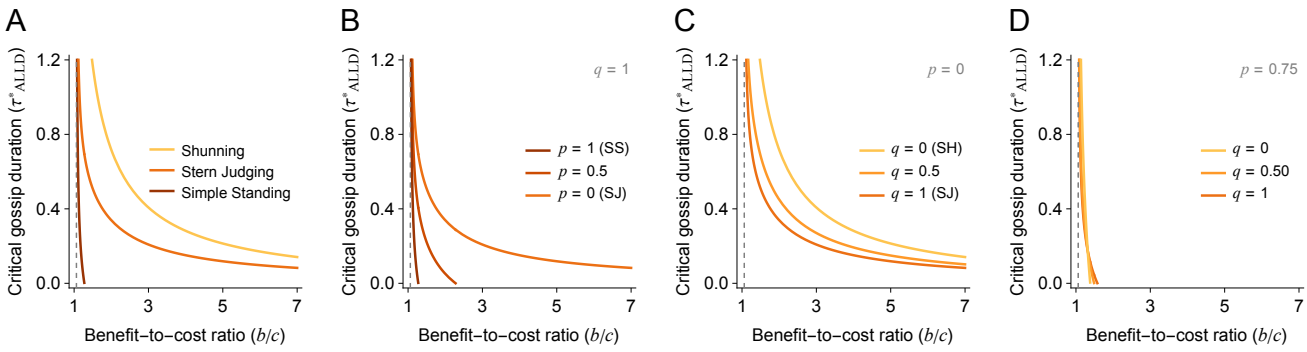

Fig. S2. Impact of social norm on the critical gossip duration for DISC to resist ALLD. Panels show the critical gossip duration τ^*_{ALLD} for a population of discriminators (DISC) to resist invasion by defectors (ALLD) as a function of the benefit-to-cost ratio. Colors denote social norms, parameterized by the probability p (q) that cooperating with (defecting against) a bad recipient yields a good reputation. **A**: For a given benefit-to-cost ratio b/c, the critical threshold τ^*_{ALLD} is the smallest for Simple Standing (SS; $(p, q) = (1, 1)$), intermediate for Stern Judging (SJ; $(p, q) = (0, 1)$), and the largest for Shunning (SH; $(p, q) = (0, 0)$). **B**: The critical gossip duration decreases with increasing p, which makes a norm more 'lenient' (i.e., incentivizes cooperating with 'bad' individuals). Parameter $q = 1$ is fixed. **C, D**: Depending on parameter values, the critical gossip duration can increase or decrease with increasing q, which makes a norm more 'strict' (i.e., incentivizes punishing 'bad' individuals). Parameter p is fixed: $p = 0$ (C) and $p = 0.75$ (D). Other parameters: $u_a = u_e = 0.02$.

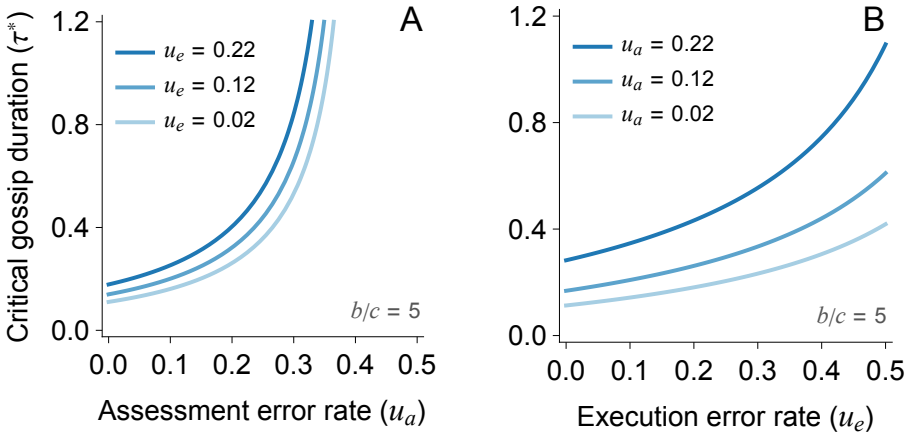

Fig. S3. Impact of assessment and execution errors on the critical gossip duration under the Stern Judging norm. Colors denote execution error rates u_e (A) or assessment error rates u_a (B). For a fixed benefit-to-cost ratio ($b/c = 5$), the critical gossip duration τ^* increases with either error rate.

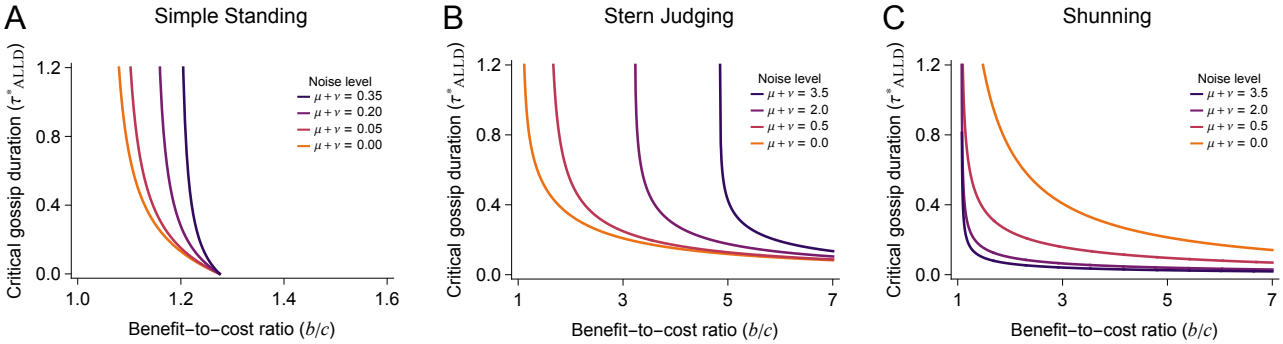

Fig. S4. Effects of noisy gossip on cooperation. Panels show the critical gossip duration τ^*_{ALLD} for a population of discriminators (DISC) to resist invasion by defectors (ALLD) as a function of the benefit-to-cost ratio, under the Simple Standing (A), Stern Judging (B), and Shunning (C) norms. Colors denote different amounts of unbiased noise in gossip ($\mu + \nu$). Each orange curve indicates the critical gossip duration for noiseless transmission ($\mu = \nu = 0$) under the corresponding norm. The critical threshold τ^*_{ALLD} increases with noise under Simple Standing (A) and Stern Judging (B; see also Fig. 2A), but the trend reverses under Shunning (C). Under the Shunning norm, reputations (before gossip) are overwhelmingly negative, and this negativity tends to be self-reinforcing because donors who cooperate with bad individuals themselves gain bad reputations; noisy gossip helps break this cycle by stochastically introducing positive gossip and, consequently, makes it easier to sustain cooperation. Other parameters: $u_a = u_e = 0.02$. Note that panel B is identical to Fig. 3A (i.e., $\tau^* = \tau^*_{\mathrm{ALLD}}$ under the Stern Judging norm) and is shown again here to facilitate comparison across norms.

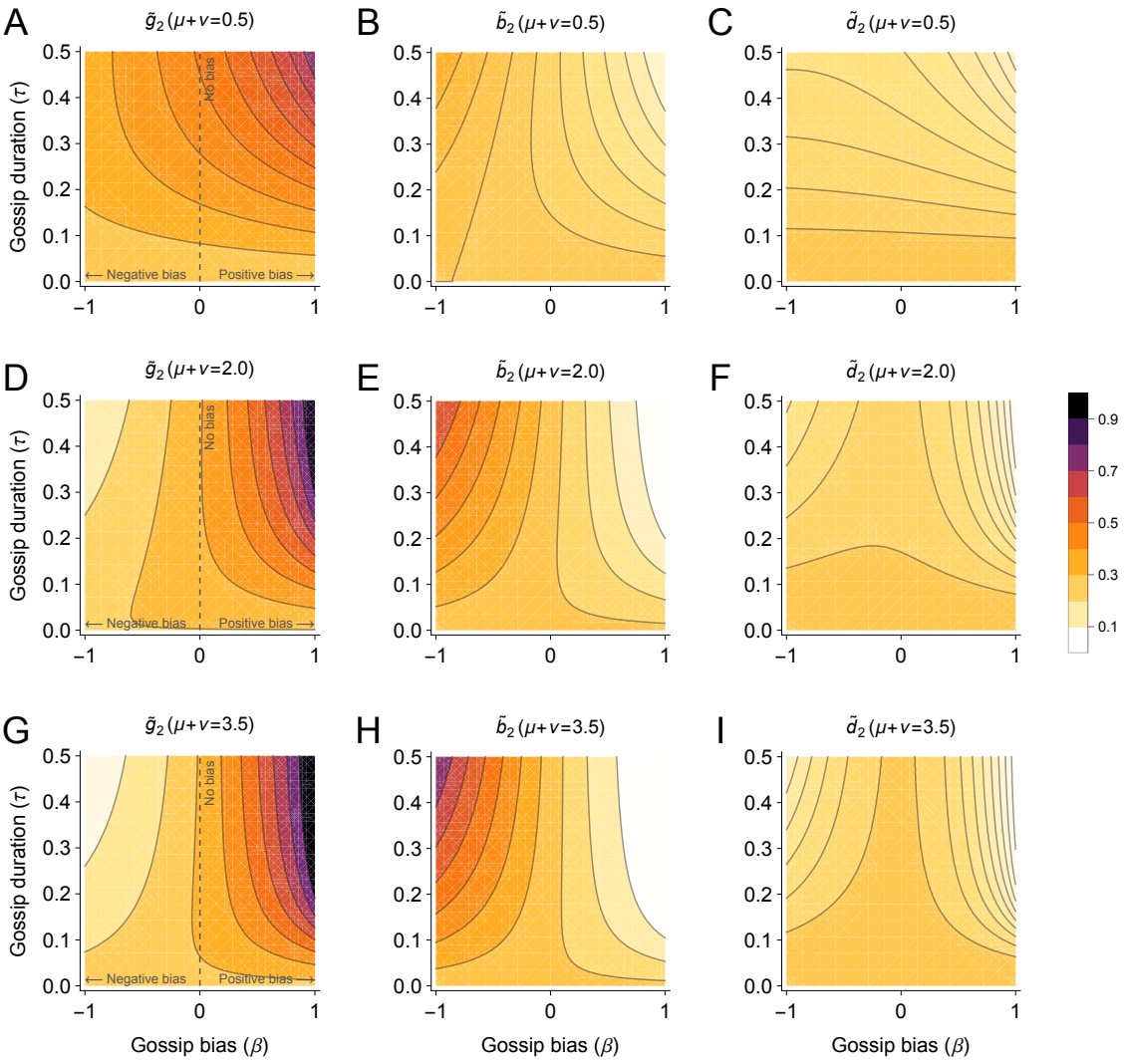

Fig. S5. Agreement and disagreement at the discriminator-only equilibrium under the Stern Judging norm. We plot the agreement ($\tilde{g}_2$, $\tilde{b}_2$) and disagreement ($\tilde{d}_2$) terms in a DISC-only population ($f_{\mathrm{DISC}} = 1$) as a function of gossip duration τ and gossip bias β in different mutation regimes ($\mu + \nu = 0.5$ in A–C, $\mu + \nu = 2.0$ in D–F, $\mu + \nu = 3.5$ in G–I). Darker colors indicate greater levels of agreement (for $\tilde{g}_2$ and $\tilde{b}_2$; A, D, G and B, E, H) or disagreement (for $\tilde{b}_2$; C, F, I). Other parameters: $u_a = u_e = 0.02$.

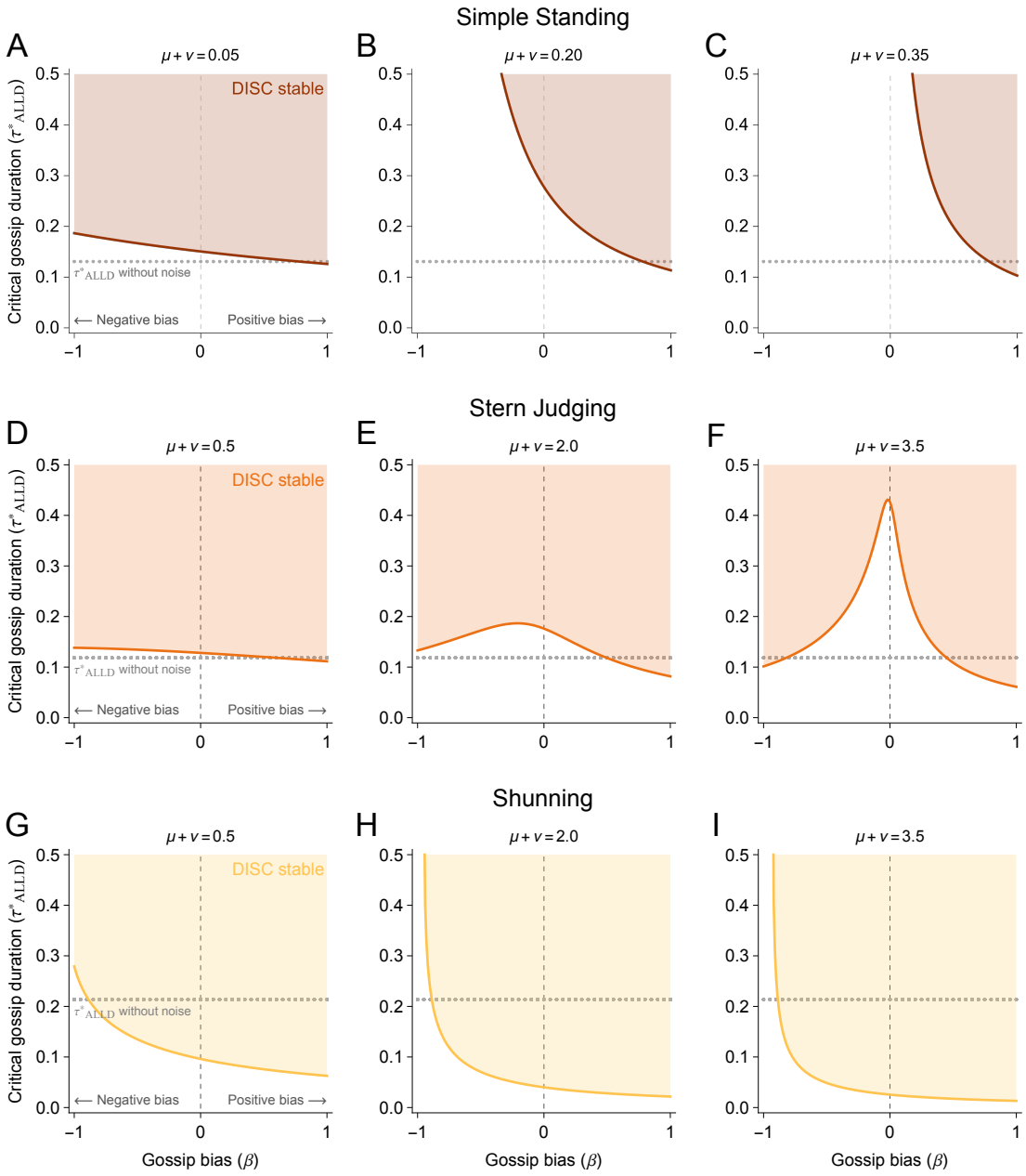

Fig. S6. Effects of biased gossip on cooperation. Panels show the critical gossip duration τ^*_{ALLD} for a population of discriminators (DISC) to resist invasion by defectors (ALLD) as a function of gossip bias (β), under the Simple Standing (A–C), Stern Judging (D–F), and Shunning (G–I) norms (denoted by colors, as in Fig. S2). Columns denote different regimes of noise as indicated. Horizontal dotted gray lines (identical across panels within each row) indicate the baseline critical gossip duration τ^*_{ALLD} in the absence of transmission noise ($\mu = \nu = 0$; orange curves in Fig. S4). We use $b/c = 1.2$ in A–C and $b/c = 5$ in D–I based on the effects of unbiased gossip identified in Fig. S4. Other parameters: $u_a = u_e = 0.02$. Note that panels D–F are identical to Fig. 4A–C (i.e., $\tau^* = \tau^*_{\mathrm{ALLD}}$ under the Stern Judging norm) and are shown again here to facilitate comparison across norms.

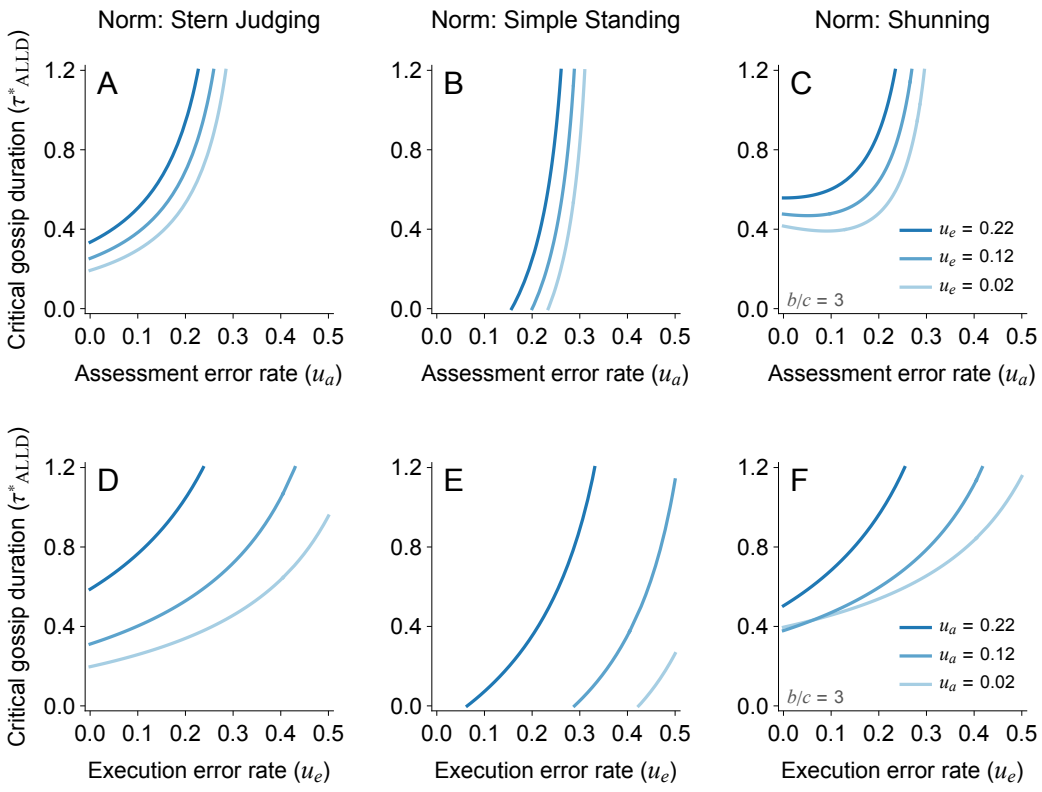

Fig. S7. Impact of errors on the critical gossip duration for DISC to resist ALLD. Colors denote execution error rates u_e (A–C) or assessment error rates u_a (D–F). **A–C**: For a fixed benefit-to-cost ratio ($b/c = 3$), the critical gossip duration τ^*_{ALLD} increases with increasing u_e. **D–F**: For a fixed benefit-to-cost ratio ($b/c = 3$), the critical gossip duration τ^*_{ALLD} increases with increasing u_a under Stern Judging and Simple Standing, but τ^*_{ALLD} is non-monotonic in u_a under Shunning.

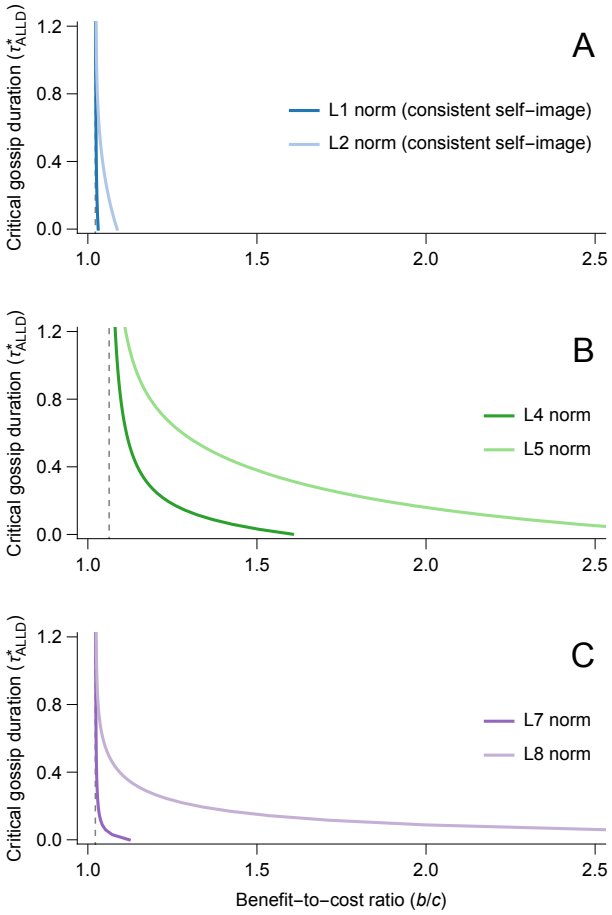

Fig. S8. Critical gossip duration for the "leading eight" norms. Panels show the critical gossip duration (τ^*_{ALLD}) as a function of the benefit-to-cost ratio (b/c) for different third-order norms, derived by solving the polynomial equations for equilibrium reputations (Eq. S16 for A; Eq. S17 for B; Eq. S18 for C) and converting Γ^* to τ^* using Eq. S11. Colors denote norms. In each panel, the two norms shown have the same vertical asymptote (dashed line), which represents the benefit-to-cost ratio b/c below which no amount of gossip can stabilize cooperation (i.e., a population of DISC can resist invasion by ALLD). **A**: $L1$ and $L2$ norms (with consistent self-image), which both judge bad donors according to Scoring. $L1$ judges good donors according to Simple Standing, while $L2$ judges good donors according to Stern Judging. Consistent self-image means that players always regard themselves as good; under this condition, the "repentant discriminator" strategy rDISC is identical to DISC. **B**: $L4$ and $L5$ norms, which both judge some donors according to Stern Judging and others according to Simple Standing. **C**: $L7$ and $L8$ norms, which both judge bad donors according to Shunning. $L7$ judges good donors according to Simple Standing, while $L8$ judges good donors according to Stern Judging. We omit $L3$ (Simple Standing) and $L6$ (Stern Judging) norms because results for these second-order norms are shown in the main text and in Fig. S2. Parameters: $u_a = u_e = 0.02$.