PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Indirect reciprocity with Bayesian reasoning and biases

Bryce Morsky[1,2]*, **Joshua B. Plotkin**[2], **Erol Akçay**[2]

1 Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **2** Department of Mathematics, Florida State University, Tallahassee, Florida, United States of America

* bmorsky@fsu.edu

OPEN ACCESS

Citation: Morsky B, Plotkin JB, Akçay E (2024) Indirect reciprocity with Bayesian reasoning and biases. PLoS Comput Biol 20(4): e1011979. https://doi.org/10.1371/journal.pcbi.1011979

Editor: Christian Hilbe, Max Planck Institute for Evolutionary Biology: Max-Planck-Institut fur Evolutionsbiologie, GERMANY

Received: September 22, 2023

Accepted: March 10, 2024

Published: April 25, 2024

Data Availability Statement: Code for analytical results, to run the numerical simulations, and to make figures is available at https://github.com/bmorsky/indirectReciprocity.

Funding: JBP acknowledges support from the John Templeton Foundation (grant #62281). EA acknowledges support from the Israel-US Binational Science Foundation (grant #2019156). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Reputations can foster cooperation by indirect reciprocity: if I am good to you then others will be good to me. But this mechanism for cooperation in one-shot interactions only works when people agree on who is good and who is bad. Errors in actions or assessments can produce disagreements about reputations, which can unravel the positive feedback loop between social standing and pro-social behaviour. Cooperators can end up punished and defectors rewarded. Public reputation systems and empathy are two possible mechanisms to promote agreement about reputations. Here we suggest an alternative: Bayesian reasoning by observers. By taking into account the probabilities of errors in action and observation and their prior beliefs about the prevalence of good people in the population, observers can use Bayesian reasoning to determine whether or not someone is good. To study this scenario, we develop an evolutionary game theoretical model in which players use Bayesian reasoning to assess reputations, either publicly or privately. We explore this model analytically and numerically for five social norms (Scoring, Shunning, Simple Standing, Staying, and Stern Judging). We systematically compare results to the case when agents do not use reasoning in determining reputations. We find that Bayesian reasoning reduces cooperation relative to non-reasoning, except in the case of the Scoring norm. Under Scoring, Bayesian reasoning can promote coexistence of three strategic types. Additionally, we study the effects of optimistic or pessimistic biases in individual beliefs about the degree of cooperation in the population. We find that optimism generally undermines cooperation whereas pessimism can, in some cases, promote cooperation.

Author summary

Cooperation is an important part of our social lives. However, selfish incentives can undermine it: I need not reciprocate with someone who has cooperated with me. One mechanism to promote cooperation is indirect reciprocity, wherein cooperation is rewarded by good reputations and defection punished by bad ones. This reward is provided indirectly, when other individuals cooperate with those of good reputation and defect against those with a bad reputation. Maintaining accurate reputations is a key part of this mechanism, since mistakes in evaluating others' behaviours and disagreements

Competing interests: The authors have declared that no competing interests exist.

about reputations can undermine the feedback loop between reputations and cooperation. Here we develop a model of individual reasoning that can correct for such mistakes and disagreements, and we explore how this impacts the level of cooperation. We find that individual reasoning to determine accurate reputations can often undermine cooperation.

Introduction

Indirect reciprocity is a well-studied mechanism that can foster cooperation among strangers, even in one-shot interactions [1–3]. Under indirect reciprocity individuals observe others' behaviours towards third parties, and assign reputations based on a social norm [3, 4]. When individuals then condition their behaviour towards each other based on reputations, for instance rewarding "good" individuals with cooperation and punishing "bad" with defection, this feedback loop can support cooperation. The dynamics of strategies under indirect reciprocity have been thoroughly explored through theoretical models [5], and this phenomenon is empirically observed in both children and adult humans [6–10] as well as non-human animals such as song sparrows [11]. Social norms, which are important in fostering cooperation generally [12, 13], govern the assignment of reputations. Several such norms have been identified as being particularly effective in maintaining cooperation [14].

Though reputations can foster cooperation through indirect reciprocity [15], this only works if reputations are accurate and individuals agree on each others' reputations. Otherwise, individuals can be wrongly punished or rewarded, leading to a cascade of assigning bad reputations and eventual defection. Importantly, errors in actions or observations can lead to such disagreements between individuals on the reputations of others. Consensus about reputations can be resolved by a public reputation system or by rapid gossip [3, 16, 17]. But when reputations of others are privately held, agreement becomes a difficult problem. Social norms that would otherwise be effective at promoting cooperation under public information, such as Stern Judging and Shunning, fail to establish cooperation under private reputation information [18]. One way to overcome the agreement problem is empathy [19], where observers can evaluate a donor through the eyes of the donor, i.e., using the recipient's reputation with the donor, which allows private reputations to better coordinate and therefore restore cooperation.

Although errors in actions and in observations play an important role in the theory of indirect reciprocity, theoretical work has assumed individuals are effectively unaware of the possibility of such errors. This means that individuals in these models always take their observations and assessments at face value. In reality, individuals may know that errors in actions or observations do sometimes occur, and they may try to account for these errors when assigning reputations (whether reputations are public or private). When observers know that errors are possible and have beliefs about the expected reputation of others, they can weigh the likelihood of a donor being good given what they have observed. To illustrate, consider an observer who observes a donor defecting against a person they deem good, under a social norm that dictates that such an action is bad. If the observer either knows or believes that there are errors in both the action (the donor intended to cooperate, but somehow failed to do so) and their own observations (the observer may have incorrectly perceived the donor's action), then they may reasonably be expected to account for this knowledge when forming an opinion of the donor's reputation. Importantly, this is true regardless of whether reputations are publicly or privately assessed.

In the example above, the observer might consider that there has been an error in action, and the donor actually meant to donate to the recipient, and so should be judged as good. Alternatively, the observer might consider the possibility that they are mistaken in believing that the recipient is good (due to an earlier error in private or public assessment of the reputation), which will change their evaluation of the donor. The observer must weigh these possible scenarios, and assess the donor as good or bad based on their perceptions of the probabilities of errors, their beliefs about prevalence of good individuals, and the social norm. For instance, if an observer believes that the vast majority of people are good and that the probability of an error is relatively high, they would likely give the donor the benefit of the doubt even if on face value they should assess the donor as bad. These considerations—based on Bayesian assignment of reputations—have the potential to change the evolutionary dynamics of cooperation.

Here, we develop a game-theoretical model of indirect reciprocity with Bayesian reasoning about reputation assignment. Players in our model engage in probabilistic reasoning [20]: they balance the probabilities of errors and their beliefs about the frequency of good individuals when forming judgements about others, using Bayes' Rule. A large literature in cognitive science has proposed that Bayesian processes explain basic aspects of human reasoning and learning [21–30]. Evidence suggests that under some circumstances human cognition may approximate a Bayesian updating process [31], or can be seen as a Bayesian sampler [32]. Bayesian updating of probabilities also can evolve under reasonable evolutionary models [33]. While other ways of reasoning about uncertain evidence exists [34, 35], Bayesian reasoning seems to be a reasonable choice to consider when modeling agents facing uncertainty. Here, we show that when observers using Bayesian reasoning to assess reputations, rather than naively accepting observations as truth, this can dramatically impact the outcome of indirect reciprocity. We consider the cases where reputations are assessed publicly or privately and the effects of biased beliefs about the reputations of others on the dynamics and equilibrium rate of cooperation. Our key finding is that Bayesian reasoning generally reduces rates of cooperation. Under the Scoring norm, however, reasoning can promote cooperation.

Methods

Indirect reciprocity

Consider a population of individuals playing a donation game where donors may, at a cost to themselves, provide a benefit to a recipient. We say that those who donate "cooperate" and those that do not donate "defect". Individuals are chosen at random to meet, one assigned to be the donor and another a recipient. A third individual, the observer, watches their interaction and assigns a reputation to the donor depending on whether or not the donor decides to cooperate and also on the reputation of the recipient (in the eyes of the observer). There is also a chance of errors that could lead to the observer evaluating the donor incorrectly. One type of error is involuntary defection [36]: with probability e_1, a donor intending to cooperate accidentally defects. The other type of error is observational: with probability e_2, an observer observes the wrong action of the donor. The probability that a donor intending to cooperate is correctly observed to be cooperating is thus $\epsilon = (1 - e_1)(1 - e_2) + e_1 e_2$, and we assume that $1 > \epsilon > \frac{1}{2} > e_2 > 0$. Note that, since there is no chance that a donor who intends to defect accidentally cooperates, the probability that a donor intending to defect is correctly observed defecting is $1 - e_2$.

We consider an infinite population of individuals engaging in such interactions. There are three strategies: always cooperate (AllC), always defect (AllD), and discriminate (Disc).

Table 1. Assessments of the donor (either *G* or *B* for good or bad) for different social norms.

	ij*: donor's action *i* and recipient's reputation *j			
Social norm	***CG***	***DG***	***CB***	***DB***
Scoring	*G*	*B*	*G*	*B*
Shunning	*G*	*B*	*B*	*B*
Simple Standing	*G*	*B*	*G*	*G*
Staying	*G*	*B*	—	—
Stern Judging	*G*	*B*	*B*	*G*

https://doi.org/10.1371/journal.pcbi.1011979.t001

As donors, AllC and AllD players will always cooperate or defect, respectively, with whomever they are matched. Discriminators, however, discriminate between "good" and "bad" recipients when deciding whether to cooperate. They will cooperate with those they deem good and defect with those they deem bad. The assignment of a reputation to a donor is determined by a set of rules called a social norm. We consider five different social norms most common in the literature: Scoring, Shunning, Simple Standing, Staying, and Stern Judging. The judgments that occur from these norm are summarized in Table 1. For example, under Simple Standing, it is considered good to cooperate with a good recipient, bad to defect against a good recipient, and good regardless of what action is taken towards a bad recipient.

The frequencies of the three strategies evolve over time, as players adopt strategies that are more successful than their own. Here we use replicator dynamics [37, 38] to model the changing strategic composition in an infinite population. Let π_i be the expected payoff to type i, and $r > 1$ be the benefit to cost ratio of cooperating. The payoffs are thus

$$\pi_x = r(x + g_x z) - 1, \quad \pi_y = r(x + g_y z), \quad \pi_z = r(x + g_z z) - g. \tag{1}$$

where x, y, and z are the frequencies of AllC, AllD, and Disc players, respectively. g_i is the frequency of i individuals with good reputations, and the total number of individuals with good reputations is $g = g_x x + g_y y + g_z z$. These reputations are assessed either publicly or privately. Under public assessment of reputations, an individual is assessed as either good or bad by all individuals and thus there will be agreement on reputations. Under private assessment, on the other hand, individuals assess reputations privately and thus players can disagree on whether an individual is good or bad. As is common in the literature, we assume that reputations equilibrate quickly before strategic frequencies change in the population. The cost to AllC players is 1 and the cost to Discriminators is g (since they will only cooperate with those they deem good). The average payoff across all individuals is $\bar{\pi} = \pi_x x + \pi_y y + \pi_z z$. Thus, the replicator equations that govern the dynamics are:

$$\dot{x} = (\pi_x - \bar{\pi})x, \quad \dot{y} = (\pi_y - \bar{\pi})y, \quad \dot{z} = (\pi_z - \bar{\pi})z. \tag{2}$$

Bayesian reasoning

Unlike previous models of indirect reciprocity, we assume observers know the error rates and consider the intentions of donors, when making reputation assessments. Observers know that donors might accidentally defect even if they intend to cooperate, and will attempt to judge them on their *intention* rather than their action. They also know that there is a chance of an assessment (observation) error; and they know the overall frequency of good individuals in the

population. The latter information is used in the Bayesian determination of whether the donor is good or not, given an observation. To make a Bayesian assessment, an observer needs to determine the probability that the donor is good given the observed action and knowledge about errors rates the frequency of good individuals in the population. Mathematically, this is:

$$\mathbb{P}(\text{donor is good}|\text{observation}) = \frac{\mathbb{P}(\text{observation}|\text{donor is good})\mathbb{P}(\text{donor is good})}{\mathbb{P}(\text{observation})}. \quad (3)$$

Each time an observer observes an interaction, they assess the donor as good with this probability.

To see how this works, consider a someone who observes a donor cooperating with a good recipient under the Simple Standing norm. Under this norm, it is considered good to contribute to a good recipient. The probability of observing a donor cooperating given that the donor is good and thus intended to give is $\mathbb{P}(\text{donor cooperates with a good recipient}|\text{donor is good}) = \epsilon$. The probability of assessing a donor as cooperating given that the donor is bad and thus intended to defect is $\mathbb{P}(\text{donor cooperates with a good recipient}|\text{donor is bad}) = e_2$. Thus, $\mathbb{P}(\text{donor cooperates with a good recipient}) = \epsilon\hat{g} + e_2(1-\hat{g})$, i.e. the chance of observing a donor cooperating is the probability of a good individual correctly giving and being assessed as having done so times the frequency of good individuals, and the probability of a bad individual mistakenly being assessed as giving times the frequency of bad individuals. Finally, we have $\mathbb{P}(\text{donor is good}) = \hat{g}$, the *perceived* frequency of good individuals in the population. As a baseline, we assume that the perceived frequency of good individuals corresponds to the true frequency, corresponding to no biased beliefs, i.e. $\hat{g} = g$. The assessment under Simple Standing for the case where an observer observes a donor cooperating with a good recipient then becomes:

$$\mathbb{P}(\text{donor is good}|\text{donor cooperates with a good recipient}) = \frac{\epsilon\hat{g}}{\epsilon\hat{g} + e_2(1-\hat{g})}. \quad (4)$$

The assessment rules for the other cases and all other norms are detailed in S1 Text.

We also analyze a generalization of our baseline model to incorporate optimism or pessimism bias in individuals' beliefs about others' reputations, as such $\hat{g} \neq g$. Under an optimism bias, $\hat{g} > g$, under a pessimism bias $\hat{g} < g$, and when there is no bias $\hat{g} = g$. Specifically, we model optimism bias by assuming $\hat{g} = (1-\lambda)g + \lambda$ where $1 > \lambda > 0$ denotes the strength of the bias. Similarly, we assume $\hat{g} = (1-\lambda)g$ under pessimism bias.

Results

Bayesian updating under public assessment of reputations

Here we summarize and discuss the results under public assessment of reputations. Our analysis proceeds as follows: first, we derive the equilibrium reputations in a population composed of a given mixture of AllD, AllC and Disc strategists for Scoring and Shunning, and we prove that reputations converge to a unique stable equilibrium for a given strategic mixture. We use these steady state reputations to calculate the evolutionary dynamics of the three strategic types of players. For the other norms, we prove that there is no strategic equilibrium in the interior of the simplex, and then show convergence of reputations along the AllD-Disc axis and analyze the strategic dynamics. The mathematical details of these results are in S1 Text. Throughout we compare these results to previous literature on indirect reciprocity under public and private assessments without Bayesian reasoning. We denote the results from prior models as the "non-reasoning" case.

Consider first the Scoring norm, where it is good to cooperate and bad to defect regardless of the recipient's reputation. In both public and private assessment without reasoning, all players playing AllD is the sole stable equilibrium [18, 39], because errors in action and observation erode the reputation of cooperators and favour AllD. However, with reasoning, we find that cooperation can be sustained. Moreover, and unlike the other social norms, there can be coexistence of all three strategies. (We note that the equations for the frequencies of good individuals for public and private assessment under Scoring are identical, and so all of the following analysis applies to both of them.) If there are sufficiently few Discriminators in the population's initial state (e.g. the lower portion of the ternary Fig 1A), then the system evolves to pure defection. Without the presence of AllC players, the average reputation g goes to zero, and thus Discriminators always defect. Therefore, a mix of AllD and Disc players can coexist at equilibrium. On the other hand, if the frequency of Discriminators is sufficiently high, paths can be attracted to an internal equilibrium supporting positive frequency of all three strategic types. The AllD-Disc boundary is still stable, but now a curve describing a polymorphic population of all three types is semi-stable, e.g. there are regions of phase space where it is attracting

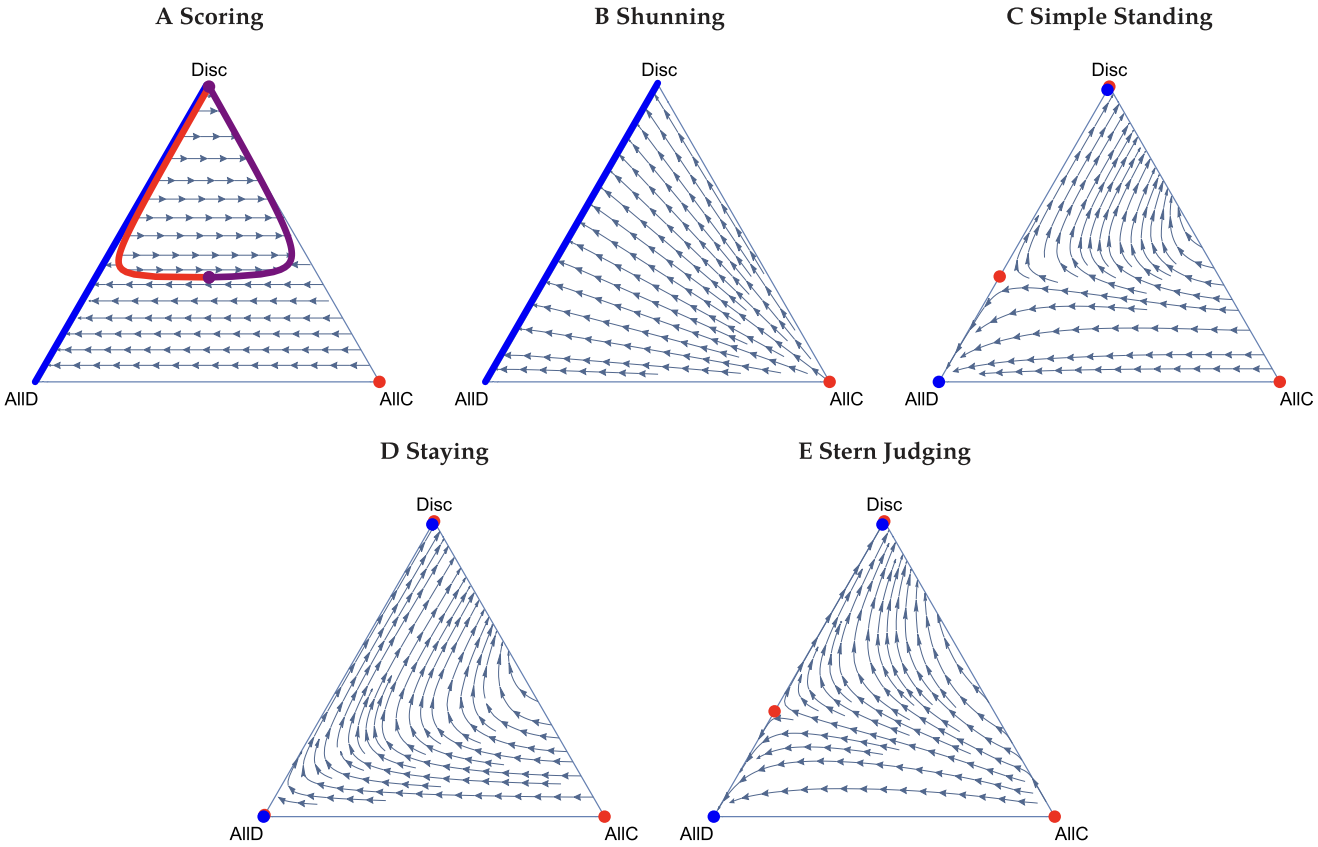

Fig 1. Ternary plots for the leading five norms under public assessment of reputations. Stable, semi-stable, and unstable equilibria are plotted in blue, purple, and red, respectively. Circles are singular equilibria, and lines sets of them. A: For Scoring, if z is sufficiently low, then the system can only evolve to the AllD-Disc boundary of the simplex where Discriminators always defect. If z is higher, then we observe bistability, which includes a set of equilibria through the interior of the simplex. Note that the red curve representing a set of unstable equilibria is interior to the simplex and only intersects with the blue curve at $z = 1$. The purple curve in turn is semistable since all points along it are attractors except at the end of the curve where Discriminators at their lowest frequency, which is unstable. B: Under Shunning, the AllD-Disc boundary is a stable set of equilibria (note that this also applies to private assessment). C-E: Simple Standing, Staying, and Stern Judging give qualitatively similar results. The system either evolves to AllD or to a position on the AllD-Disc boundary. Note that the stable equilibria on the AllD-Disc boundary are very near $z = 1$. In all figures, the benefit to cost ratio is $r = 3$ and the error rates are $e_1 = e_2 = 0.01$.

https://doi.org/10.1371/journal.pcbi.1011979.g001

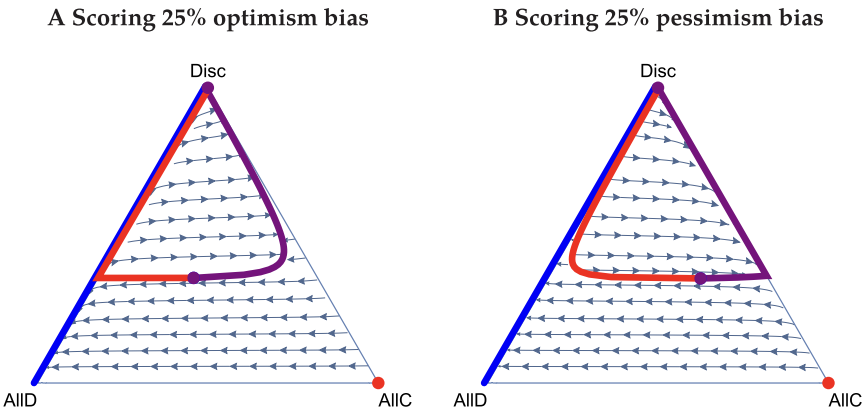

Fig 2. Ternary plots for Scoring under biases, 25% optimism (A) and 25% pessimism (B). Stable, semistable, and unstable equilibria are again plotted in blue, purple, and red, respectively. Circles are singular equilibria, and lines sets of them. These biases have shifted the internal equilibria relative to Fig 1A.

https://doi.org/10.1371/journal.pcbi.1011979.g002

and regions where it is repelling. Between these two is a set of unstable equilibria. The semi-stable polymorphic equilibria are in the interior the simplex, which is not observed in any other norms with reasoning. The points on this curve of the semi-stable polymorphic equilibria all support different, positive levels of cooperation. This curve is only semi-stable and not stable, since the end point (where Discriminators are at their nadir) is unstable to perturbations that decrease the frequency of Discriminators. The equilibria are otherwise attracting. We find that the lower the error rate, the larger the region of coexistence of strategies. However, there is a limit on how large it may be, even with no errors. On the other hand, if the error rate is sufficiently large, then the interior equilibria disappear and the only stable equilibrium is the AllD-Disc boundary. The presence of optimism or pessimism bias contorts the semi-stable interior equilibria that arise under the Scoring norm. Optimism bias draws them to the AllD-Disc boundary, and pessimism bias draws them to the AllC-Disc boundary. Fig 2 depicts this phenomena for a 25% optimism bias and a 25% pessimism bias.

Next consider the Shunning norm. Again, public and private assessments lead to qualitatively identical results. Without Bayesian reasoning, Shunning results in either AllD and Disc both being stable equilibria (and the system is thus bistable), or AllD being globally asymptotically stable [18]. With reasoning, we find a stable set of equilibria that excludes cooperation as depicted in Fig 1B. This set exists regardless of the benefit to cost ratio r, errors rates, or the presence of bias. It arises because Bayesian reasoning under the Shunning norm drives the reputations of all types to zero. Therefore, Discriminators behave exactly as AllD players and never cooperate.

Finally, we consider Simple Standing, Staying, and Stern Judging together, as they have qualitatively similar dynamics with public assessment of reputations. These three norms can all display bistable dynamics, a monomorphic population of AllD is always stable under all three, but there may be an additional stable equilibrium on the AllD-Disc boundary. Staying has a larger basin of attraction of the cooperative equilibrium (the one with Discriminators) than both Simply Standing and Stern Judging. Fig 1C–1E depict ternary figures for these norms when there is no bias.

Bias in the prior beliefs can change the qualitative nature of the equilibria by destroying the equilibrium along the AllD-Disc edge for all three of these norms, and both for optimism and pessimism bias, so long as it is sufficiently large (see S1 Fig for ternary figures with only 25%

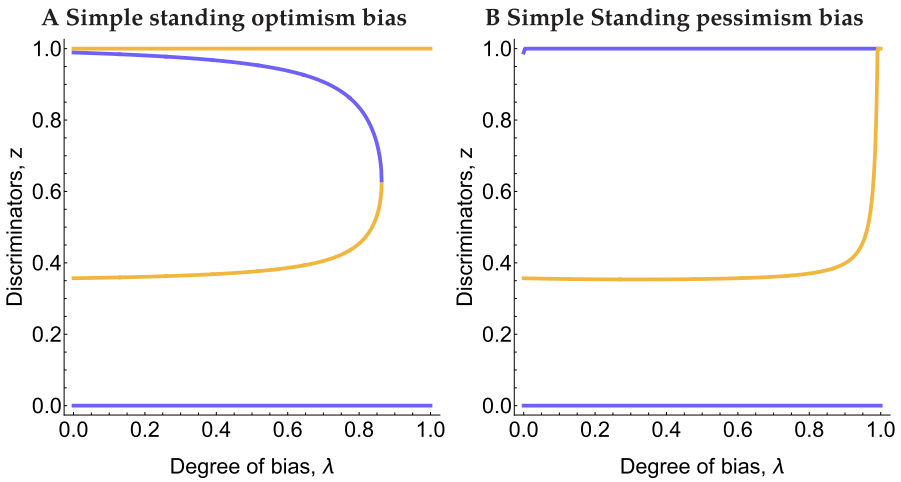

Fig 3. Bifurcation diagrams for optimism (A) and pessimism (B) biases under Simple Standing. Here AllC strategists are excluded and thus $1 - z = y$. Violet curves are stable equilibria and orange curves are unstable.

https://doi.org/10.1371/journal.pcbi.1011979.g003

bias). For optimism bias, the frequency of Discriminators decreases in the edge equilibrium (and therefore the frequency of good players and cooperation) as the degree of bias increases. At a critical bifurcation point, the internal equilibria are annihilated and the sole stable equilibrium is only composed of AllD players. An initial small degree of pessimism bias can benefit Discriminators pushing the equilibrium of only Discriminators. However, for very large negative bias, the basin of attraction to this equilibrium evaporates leaving the monomorphic equilibrium of AllD players as the only stable equilibrium. The bifurcation diagrams in Fig 3 depict these phenomena under Simple Standing. Staying and Stern Judging produce qualitatively similar diagrams (see S2 Fig).

Thus far, we have fixed the error rates in the figures produced. Here we explore the impact of various error rates on the amount of cooperation at equilibrium for the Bayesian reasoning model compared to the non-reasoning model. These results are plotted in Fig 4. The system is initialized at strategy frequencies evenly spread across the simplex and then numerically solved and averaged. The average frequency of cooperation under non-reasoning is then subtracted from the amount from Bayesian reasoning. Bayesian reasoning generally does worse than non-reasoning: cooperation is lower for Bayesian reasoning than for non-reasoning except for Scoring, which always does better than non-reasoning. Further, Bayesian reasoning generally performs relatively best when error rates are low. We also explored varying errors rates and biases (results in S3 Fig).

Private assessment of reputations

Here we consider private assessment of reputations. Since Scoring and Shunning norms result in the same behaviour for public and private assessment, we focus on Simple Standing, Staying, and Stern Judging. Under Simple Standing, in the absence of any bias, cooperation can be maintained. However, AllC strategists cannot be part of an equilibrium. Therefore, all equilibria are along the AllD-Disc boundary. We also find that the monomorphic AllD equilibrium ($y^* = 1$) is always stable and the monomorphic Disc equilibrium ($z^* = 1$) is always unstable. Between these two equilibria there may be zero, one, or two equilibria depending on the benefit to cost ratio r, and error rates e_1 and e_2 (see S1 Text for details of these conditions). For parameters where there is no equilibrium along the Disc-AllD line, AllD ($y^* = 1$) is globally

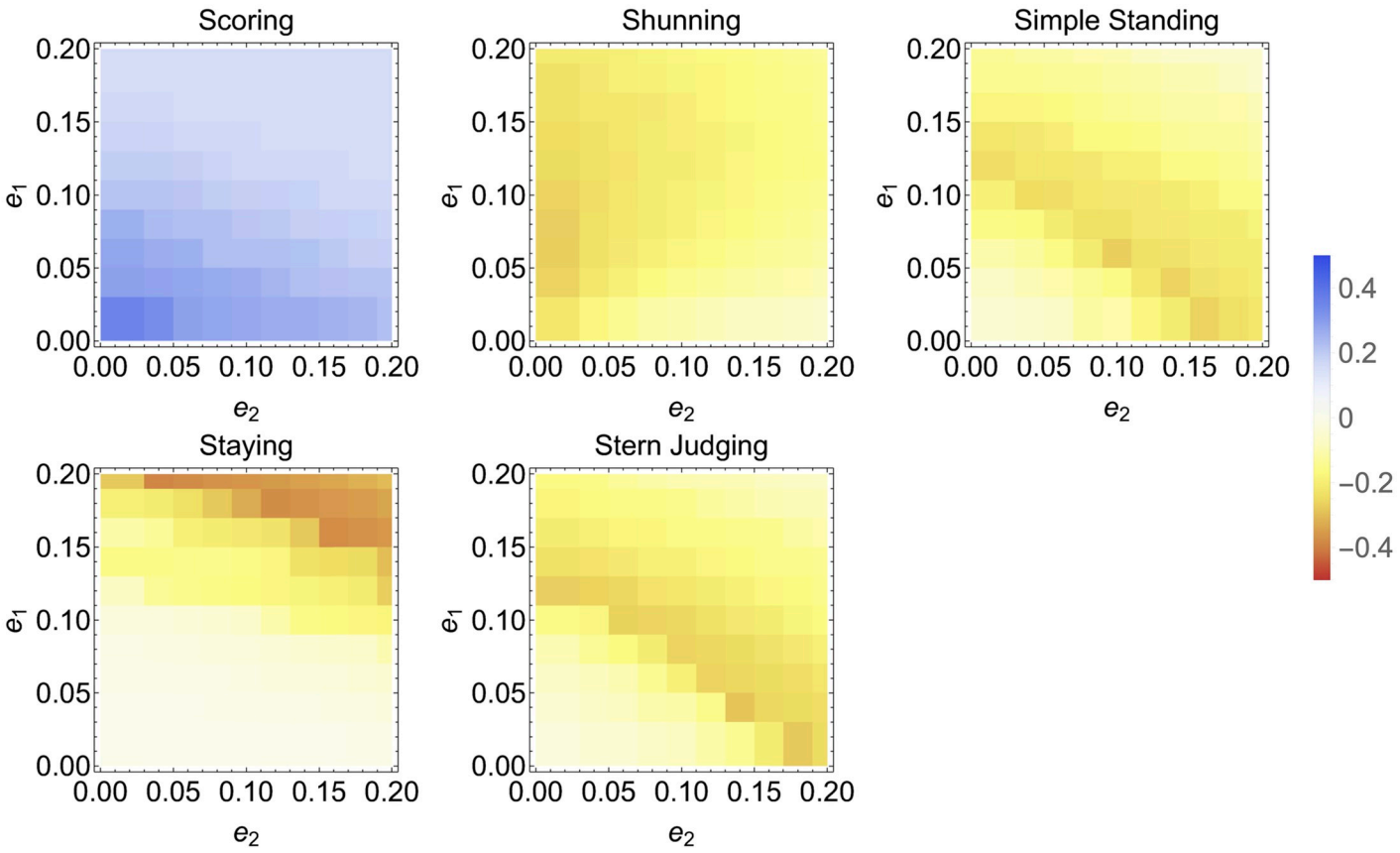

Fig 4. Heat maps of relative cooperation for Scoring, Shunning, Simple Standing, Staying, and Stern Judging, i.e. each cell represents the average amount of cooperation under Bayesian reasoning minus the average amount of cooperation under non-reasoning for different error rates. The average is over initial conditions evenly spread across the simplex and $r = 3$. Note that reasoning generally under-performs compared to non-reasoning except for Scoring.

https://doi.org/10.1371/journal.pcbi.1011979.g004

stable. If there is one equilibrium containing a mixture of AllD and Disc, it is semi-stable along the AllD-Disc boundary, i.e. it is stable coming from the direction of AllD, and unstable from the direction of Disc. In the case where there are two equilibria, then the one closest to $z^* = 1$ is stable, while the other is not, and the dynamics of the system converge to either to the monomorphic AllD or the stable mixture of AllD and Disc. The latter equilibrium can maintain a high level of cooperation as it can mostly consist of Discriminators. Fig 5A depicts this case for $r = 3$ and $e_1 = e_2 = 0.01$. Note that in the figure, the stable mixture is mostly Discriminators so that it is very close to the corner with all Discriminators (which itself is an unstable equilibrium).

Biases in prior beliefs about reputations of others can have a dramatic impact on the dynamics of the strategies under Simple Standing. Fig 6 depicts the most cooperative equilibrium (corresponding to the AllD-Disc mixture described above without bias), and the amount of intended cooperation at equilibrium (i.e. $x^* + g^* z^*$). Fig 6A shows that increasing optimism bias decreases the frequency of Disc players at equilibrium and reduces the total amount of cooperation. This happens because optimism bias causes Discriminators to evaluate more individuals as good and cooperate with them when they should not, which favours AllD strategists. This effect is relatively mild except for very high degree of optimism bias, where the gullibility of Discriminators drives them extinct and the population consists entirely of AllD

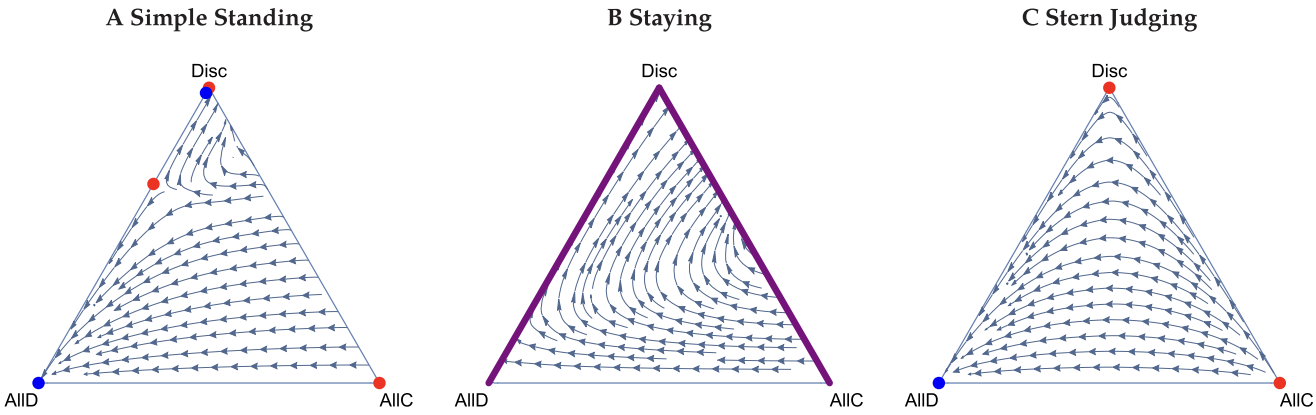

Fig 5. Ternary plots for Simple Standing, Staying, and Stern Judging under private assessment. Stable and unstable equilibria are plotted in blue and red, respectively. Circles are singular equilibria and lines sets of them. A: Simple Standing gives qualitatively similar results to Public assessment. The system either evolves to AllD or a point on the interior of the AllD-Disc boundary. For this panel, $e_1 = e_2 = 0.05$ so that the plotting of the equilibria are clear. For $e_1 = e_2 = 0.01$, the interior stable equilibrium is nearly on the Disc vertex. B: Under Staying, the AllC-Disc and AllD-Disc boundaries are sets of equilibria. Further, at $z^* = 1$, any reputation is an equilibria. Thus, the trajectory towards $z^* = 1$ will determine reputations at it. C: Under Stern Judging, AllD is globally asymptotically stable. $r = 3$ for all figures, and $e_1 = e_2 = 0.01$ for panels B and C.

https://doi.org/10.1371/journal.pcbi.1011979.g005

strategists. Negative bias (Fig 6B) displays a qualitatively different pattern: here, even for small amount of bias, discriminator frequency at equilibrium goes down significantly, but now they are replaced by AllC strategists. This happens because pessimism bias causes Discriminators to defect against individuals judged as good by others, which makes them subject to punishment by other Discriminators. AllC strategists, who always cooperate, do not suffer from this punishment and can increase in frequency. Strikingly, when pessimism bias increases further, this effect is reversed: AllC strategists start getting punished as much as Discriminators, and lose their advantage, which causes Discriminators to increase in frequency at high pessimism biases. This exchange of Disc and AllC strategists causes little change in the overall level of cooperation for most degrees of pessimism bias, except for when pessimism bias is so high that

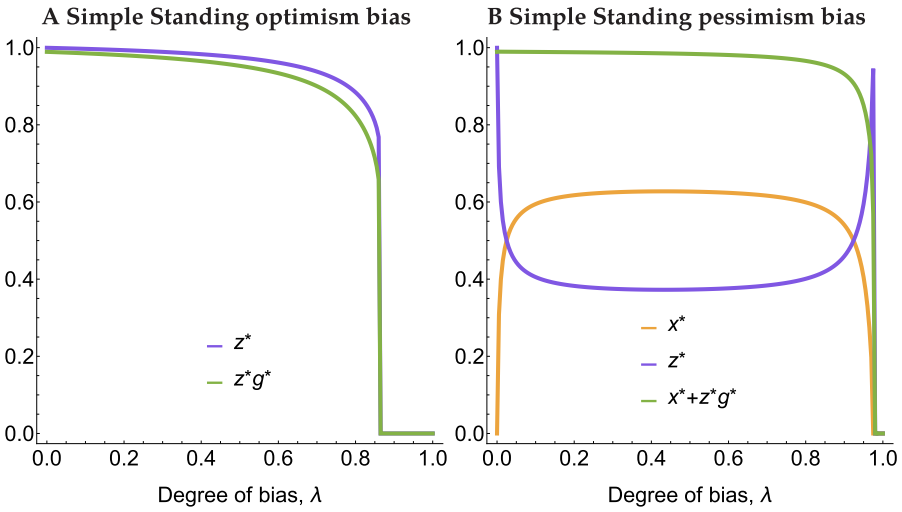

Fig 6. Equilibria x^* (orange), z^* (violet), and the equilibrium degree of cooperation $x^* + g^*z^*$ (green) for varying degrees of optimism and pessimism biases under Simple Standing.

https://doi.org/10.1371/journal.pcbi.1011979.g006

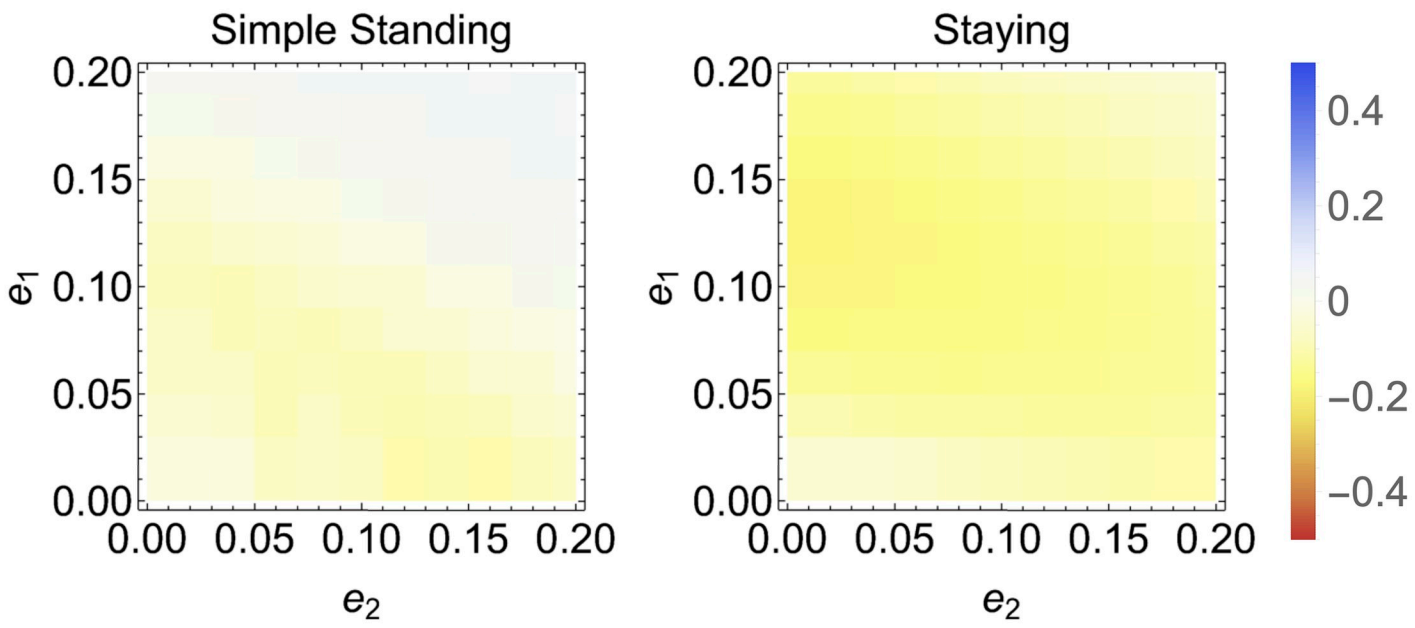

Fig 7. Heat maps of relative cooperation for Simple Standing and Staying, i.e. cooperation under Bayesian reasoning minus cooperation under non-reasoning. Results are averaged over initial conditions evenly spread across the simplex, with $r = 3$.

https://doi.org/10.1371/journal.pcbi.1011979.g007

Discriminators regard everyone to have bad reputations and defect against them. At this point (the far right hand side of Fig 6B), Discriminators get replaced by AllD strategists.

Fig 5B depicts a ternary plot for Staying. The AllC-Disc and AllD-Disc boundaries both form sets of equilibria since $g^* = 1$ and $g^* = 0$ on each, respectively. $y^* = 1$ is therefore semi-stable, since perturbations from it along the AllD-Disc boundary are neutral and trajectories with sufficiently low z converge to it. Perturbations to $z^* = 1$ may also occur neutrally along the AllD-Disc boundary in addition to along the AllC-Disc boundary. $z^* = 1$ is attracting in the interior of strategy space. Further, all reputations are equilibria at $z^* = 1$. Therefore, g^* at $z^* = 1$ will depend on the trajectory approaching it. Both this result for Staying and that for Simple Standing are in contrast to non-reasoning models of private assessment, where there is a polymorphic stable equilibrium of Discriminators and AllC players [18], and public assessment, where $z^* = 1$ is stable [39].

Finally, Bayesian reasoning makes no impact on Stern Judging when there is no bias. Like the non-Bayesian case under private assessment [18], the equilibrium reputations for all strategies are always $\frac{1}{2}$. Therefore, $\pi_y > \pi_z > \pi_x$, which results in all players playing AllD as depicted in Fig 5C. Even moderate bias has no discernible effect on outcomes (see S4 Fig for ternary plots of Stern Judging along with Staying and Simple Standing with $\lambda = 0.25$).

Fig 7 shows heat maps for varying error rates under private assessment for Simple Standing and Staying (Scoring is not presented as the results are identical to public assessment of reputations, and Shunning and Stern Judging are not presented because they always lead to defection). Bayesian reasoning generally under-performs compared to non-reasoning, although this effect is small and it is lessened when error rates are low. One exception, however, is that for high error rates, Simple Standing has approximately the same degree of cooperation as or more than non-reasoning. Since these results show a marginal difference between Bayesian reasoning and non-reasoning and since they are generated from discretizing the space and numerical simulations, we cannot conclude that Bayesian reasoning can significantly promote

cooperation. Rather, error rates have a marginal impact on the relative outcomes particularly with respect to Simple Standing. We also explored varying errors rates and biases (results in S5 Fig).

Discussion

Indirect reciprocity is one of the main mechanisms that can sustain cooperation among strangers [3]. But this mechanism requires keeping track of reputations in the population, and it also requires some degree of mutual agreement about reputations in the population. In particular, Discriminators must agree on who the "bad" people are so that they can effectively punish them while rewarding those who are "good." When evaluations of others are privately held, this coordination can break down, and with it cooperation maintained by indirect reciprocity. Public institutions that decide and promulgate reputations centrally [16] or individuals putting themselves in others' shoes, i.e., empathy [19], can solve this problem, however even in publicly shared reputations, individuals still have to contend with errors of action and observation. There is a large body of evidence suggesting that humans, when they need to make decisions or learn in the presence of errors and noise, employ Bayesian reasoning to take into account the sources of uncertainty [24, 25, 40]. We have shown that Bayesian reasoning about two sources of error, in action and perception, can substantially alter social behaviour in the context of reputations.

Bayesian reasoning about the sources of error can sometimes promote cooperation that would otherwise be suppressed. The most striking example of this happens with Scoring, which is an attractive social norm since it is first order (individuals' reputations only depend on their own behaviour, not the recipients) and thus presents low cognitive and informational requirements. In most theoretical treatments, however, Scoring is plagued by the fact that, in the presence of AllC, errors gradually erode reputations and, subsequently, cooperation [5, 39, 41]. Particularly, inaccurate information can prevent the evolution of social norms and cooperation [41]. Bayesian reasoning can counteract this erosion, because it allows individuals to overlook errors provided they have sufficiently high prior expectations for good reputations in the population. In this way, individuals can avoid mistakenly assigning bad reputations to cooperators and Discriminators and maintain the relative reward these types enjoy. This process can maintain a mixed equilibrium with all strategic types. Specifically, AllD can also exist in these equilibria because they are given a measure of relief as defection is sometimes ascribed to errors. Notably, however, this only works if the population starts from a sufficient fraction of Discriminators; otherwise the AllD-Disc boundary is the only attractive equilibrium. [42] also found attractive interior equilibria in a model that has no observation error but potentially incomplete observation of past history. Such interior stable sets of equilibria disappear when there are both observation and action errors [18], yet they return in a different form when Bayesian reasoning is introduced.

Bayesian reasoning under public information has more complicated effects for second-order norms, where the assessment of a donor's action depends on the recipient's reputation. Under the very strict Shunning norm (where interacting with bad individuals at all is bad, regardless of outcome) for example, Bayesian reasoning precludes a cooperative equilibrium that exists with public reputations and no reasoning [18, 39]. That equilibrium exists in the non-reasoning case because errors about recipient reputation ameliorates the erosion of reputation for Disc and AllC strategists that interact with bad individuals, so that a reputational equilibrium with positive probably of being good is possible for these types. Bayesian reasoning removes this refuge and all reputations evolve to bad. The other three norms we have studied (Simple Standing, Staying, and Stern Judging) show relatively smaller difference with the

non-reasoning case, all showing bistability with one AllD and one cooperative equilibrium. However under Bayesian reasoning, the cooperative equilibrium is a mixture of Disc and AllD, instead of Disc and AllC, which all things being equal results in lower cooperation. And so reasoning can be a double=edged sword when it comes to maintaining cooperation, depending on the social norm of judgement.

Considering probabilistic reasoning about reputations also allows us to model the potential for biased beliefs about others. Classic results in social psychology show that humans attend to and retain negative social information or signals more readily [43, 44]. These biases might lead actors to underestimate the frequency of good individuals in their prior beliefs (pessimism bias). On the other hand, other well-established biases in perception and decision-making such as optimism bias [45] or ego-centric bias [46] can induce actors to overestimate the frequency of good individuals. While the social psychology research on biases in social perception and inference is vast, to our knowledge its effect on the dynamics of indirect reciprocity had not been studied.

We have shown that biases in beliefs of others can sometimes help sustain cooperators (though not necessarily cooperation). And yet excessive biases tend to unravel cooperative behaviour, resulting in negative reputations for everyone and defection rampant. In both cases, the effects are dependent on the norm: with Scoring, for example, optimism bias (thinking others are good more often than they really are) can increase the basin of attraction of the internal equilibrium curve with cooperation, but pessimism bias shrinks it. With Simple Standing (under public reputations), however, optimism bias lowers the degree of cooperation as it makes Discriminators susceptible to cooperating with AllD strategists. Pessimism bias however, can help Discriminators. Both positive and negative bias, however, destroys cooperative equilibria when it is extreme, as Discriminators with highly biased beliefs make uninformative judgments.

There are several limitations and assumptions we have made that may provide interesting avenues for future research. For one, we have assumed an infinite population and no stochasticity. We expect that a model featuring stochasticity and a large but finite population would behave similarly to our model. Except there may be key differences with respect to the continuous sets of equilibria. A finite population experiencing stochasticity could move along these sets of equilibria, which could act as a bridge from one region of phase space to another. We have also assumed no cognitive constraints or complexity costs [47] on individuals, which is an important factor [48]. Previous research on indirect reciprocity that has explored complexity costs for Discriminators finds that such costs can undermine cooperation [42, 47, 49]. Were we to impose such costs here, they may exacerbate the failures of Bayesian reasoning to promote cooperation since Discriminators would earn lower payoffs. On the other hand, this effect may promote cooperators relative to Discriminators, which in turn could promote overall cooperation. It is also of note that Bayesian reasoning tends to be beneficial under the simplest norm and thus arguably the norm with the lowest complexity cost (since observers need not factor in the reputation of the recipient). Whether or not Bayesian reasoning might be advantageous with a complexity cost is thus an open question. Though cognitive costs for reciprocity may not be high [50], they may be substantial for the type of reasoning employed here and sufficient enough to impact the qualitative behaviour of the system. In addition to these points, the ability to use Bayesian reasoning can vary between individuals [51, 52], whereas we have assumed all agents reason in the same way.

Another simplifying assumption is that even with private assessments of individual reputations, beliefs about the error rates and the overall frequency of good individuals are homogeneous and, with respect to the error rates, accurate. Under private assessment of reputations, individuals could obtain such accurate and homogeneous information through their own and

others' past personal experience and repeated interactions (or, under public reputation, through institutions). But what happens if such learning is not feasible or happens only slowly remains an open question. Under the norms where there is bistability, if a population initially has a high belief in the frequency of good individuals, that might be enough to bootstrap cooperative equilibria [53]. For error rates, individuals who intend to cooperate may know the rate at which they fail to cooperate, e_1, through their own involuntary failures to cooperate, which could be transmitted via gossip to others. Similarly, the observational error, e_2, may be learned through gossip when individuals learn they have been inaccurately assessed. At the same time, another well-documented cognitive bias, overconfidence [54], might lead to error rates to be underestimated.

Rationality and probabilistic reasoning have an important role in economic theory [55]. And, it has been argued that humans engage in Bayesian rationality to reason about uncertainty [56]. Our model of indirect reciprocity can be interpreted in this sense. However, surprisingly, reasoning in this matter is not always effective in promoting cooperation, in our analysis. For cooperation to be fostered, AllD players must receive relatively low payoffs, which is not the same as uncovering the true frequencies of good individuals and agreeing on reputations.

Supporting information

S1 Text. Analytical results.
(PDF)

S1 Fig. Ternary plots for Simple Standing, Staying, and Stern Judging under public assessment of reputations with bias λ = 0.25. The benefit to cost ratio is $r = 3$ and the error rates are $e_1 = e_2 = 0.01$. The results are not qualitatively different from when there is no bias.
(PDF)

S2 Fig. Bifurcation diagrams for optimism and pessimism biases under public assessment of reputations and for Staying and Stern Judging. Here AllC strategists are excluded and thus $1 - z = y$. Violet curves are stable equilibria and orange curves are unstable. The results are qualitatively similar to that of Simple Standing in the main text.
(PDF)

S3 Fig. Heat maps of relative cooperation for Scoring, Simple Standing, Staying, and Stern Judging under public assessment of reputations and for different error rates and degrees of bias λ. Each cell represents the average amount of cooperation under Bayesian reasoning minus the average amount of cooperation under non-reasoning. The average is over initial conditions evenly spread across the simplex and $r = 3$. We observe a slight synergy between the error rates and bias for Scoring under optimism bias, Simple Standing and Stern Judging under pessimism bias, and public assessment of reputations: Bayesian reasoning generally has relatively lower cooperation when biases are large and errors low.
(PDF)

S4 Fig. Ternary plots for Simple Standing, Staying, and Stern Judging under private assessment of reputations with bias λ = 0.25. The benefit to cost ratio is $r = 3$ and the error rates are $e_1 = e_2 = 0.01$. Bias has no great impact on the qualitative outcome, except that private Staying no longer has unstable equilibria on two boundaries. For optimism bias and Staying, the plot is qualitatively similar to public assessment of reputations. Negative bias results in a stable equilibrium along the AllC-Disc boundary.
(PDF)

S5 Fig. Heat maps of relative cooperation for Scoring, Simple Standing, Staying, and Stern Judging under private assessment of reputations and for different error rates and degrees of bias λ. Each cell represents the average amount of cooperation under Bayesian reasoning minus the average amount of cooperation under non-reasoning. The average is over initial conditions evenly spread across the simplex and $r = 3$.
(PDF)

Author Contributions

Conceptualization: Bryce Morsky, Joshua B. Plotkin, Erol Akçay.

Formal analysis: Bryce Morsky.

Investigation: Bryce Morsky, Joshua B. Plotkin, Erol Akçay.

Software: Bryce Morsky.

Writing – original draft: Bryce Morsky.

Writing – review & editing: Bryce Morsky, Joshua B. Plotkin, Erol Akçay.

References

1. Alexander Richard D. *The biology of moral systems*. Routledge, 2017.
2. Leimar Olof and Hammerstein Peter. Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1468):745–753, 2001. https://doi.org/10.1098/rspb.2000.1573 PMID: 11321064
3. Nowak Martin A and Sigmund Karl. Evolution of indirect reciprocity. *Nature*, 437(7063):1291–1298, 2005. https://doi.org/10.1038/nature04131 PMID: 16251955
4. Sigmund Karl. *The calculus of selfishness*. Princeton University Press, 2010.
5. Okada Isamu. A review of theoretical studies on indirect reciprocity. *Games*, 11(3):27, 2020. https://doi.org/10.3390/g11030027
6. Engelmann Dirk and Fischbacher Urs. Indirect reciprocity and strategic reputation building in an experimental helping game. *Games and Economic Behavior*, 67(2):399–407, 2009. https://doi.org/10.1016/j.geb.2008.12.006
7. Kato-Shimizu Mayuko, Onishi Kenji, Kanazawa Tadahiro, and Hinobayashi Toshihiko. Preschool children's behavioral tendency toward social indirect reciprocity. *PloS one*, 8(8):e70915, 2013. https://doi.org/10.1371/journal.pone.0070915 PMID: 23951040
8. Nava Elena, Croci Emanuela, and Turati Chiara. 'I see you sharing, thus I share with you': indirect reciprocity in toddlers but not infants. *Palgrave Communications*, 5(1):1–9, 2019. https://doi.org/10.1057/s41599-019-0268-z
9. Seinen Ingrid and Schram Arthur. Social status and group norms: Indirect reciprocity in a repeated helping experiment. *European economic review*, 50(3):581–602, 2006. https://doi.org/10.1016/j.euroecorev.2004.10.005
10. Yoeli Erez, Hoffman Moshe, Rand David G, and Nowak Martin A. Powering up with indirect reciprocity in a large-scale field experiment. *Proceedings of the National Academy of Sciences*, 110(Supplement 2):10424–10429, 2013. https://doi.org/10.1073/pnas.1301210110
11. Akçay Çağlar, Reed Veronica A, Campbell S Elizabeth, Templeton Christopher N, and Beecher Michael D. Indirect reciprocity: song sparrows distrust aggressive neighbours based on eavesdropping. *Animal Behaviour*, 80(6):1041–1047, 2010. https://doi.org/10.1016/j.anbehav.2010.09.009
12. Kandori Michihiro. Social norms and community enforcement. *The Review of Economic Studies*, 59(1):63–80, 1992. https://doi.org/10.2307/2297925
13. Tomasello Michael and Vaish Amrisha. Origins of human cooperation and morality. *Annual review of psychology*, 64:231–255, 2013. https://doi.org/10.1146/annurev-psych-113011-143812 PMID: 22804772
14. Ohtsuki Hisashi and Iwasa Yoh. The leading eight: social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, 239(4):435–444, 2006. https://doi.org/10.1016/j.jtbi.2005.08.008 PMID: 16174521

15. Milinski Manfred, Semmann Dirk, and Krambeck Hans-Jürgen. Reputation helps solve the 'tragedy of the commons'. *Nature*, 415(6870):424–426, 2002. https://doi.org/10.1038/415424a PMID: 11807552

16. Radzvilavicius Arunas L, Kessinger Taylor A, and Plotkin Joshua B. Adherence to public institutions that foster cooperation. *Nature communications*, 12(1):1–14, 2021. https://doi.org/10.1038/s41467-021-24338-8

17. Sommerfeld Ralf D, Krambeck Hans-Jürgen, Semmann Dirk, and Milinski Manfred. Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the national academy of sciences*, 104(44):17435–17440, 2007. https://doi.org/10.1073/pnas.0704598104 PMID: 17947384

18. Okada Isamu, Sasaki Tatsuya, and Nakai Yutaka. A solution for private assessment in indirect reciprocity using solitary observation. *Journal of theoretical biology*, 455:7–15, 2018. https://doi.org/10.1016/j.jtbi.2018.06.018 PMID: 29997059

19. Radzvilavicius Arunas L, Stewart Alexander J, and Plotkin Joshua B. Evolution of empathetic moral evaluation. *eLife*, 8:e44269, 2019. https://doi.org/10.7554/eLife.44269 PMID: 30964002

20. Oaksford Mike, Chater Nick, and Larkin Joanne. Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(4):883, 2000. PMID: 10946369

21. Courville Aaron C, Daw Nathaniel D, and Touretzky David S. Bayesian theories of conditioning in a changing world. *Trends in cognitive sciences*, 10(7):294–300, 2006. https://doi.org/10.1016/j.tics.2006.05.004 PMID: 16793323

22. Eguíluz Víctor M, Masuda Naoki, and Fernández-Gracia Juan. Bayesian decision making in human collectives with binary choices. *PLoS One*, 10(4):e0121332, 2015. https://doi.org/10.1371/journal.pone.0121332 PMID: 25867176

23. Gopnik Alison. Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, 337(6102):1623–1627, 2012. https://doi.org/10.1126/science.1223416 PMID: 23019643

24. Gopnik Alison, Glymour Clark, Sobel David M, Schulz Laura E, Kushnir Tamar, and Danks David. A theory of causal learning in children: causal maps and Bayes nets. *Psychological review*, 111(1):3, 2004. https://doi.org/10.1037/0033-295X.111.1.3 PMID: 14756583

25. Griffiths Thomas L and Tenenbaum Joshua B. Optimal predictions in everyday cognition. *Psychological science*, 17(9):767–773, 2006. https://doi.org/10.1111/j.1467-9280.2006.01780.x PMID: 16984293

26. Jacobs Robert A and Kruschke John K. Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(1):8–21, 2011. PMID: 26301909

27. Marchetti Gionni, Patriarca Marco, and Heinsalu Els. A bayesian approach to the naming game model. *Frontiers in Physics*, 8:10, 2020. https://doi.org/10.3389/fphy.2020.00010

28. Pérez Toni, Zamora Jordi, and Eguíluz Víctor M. Collective intelligence: Aggregation of information from neighbors in a guessing game. *PloS one*, 11(4):e0153586, 2016. https://doi.org/10.1371/journal.pone.0153586 PMID: 27093274

29. Pérez-Escudero Alfonso and de Polavieja Gonzalo. Collective animal behavior from bayesian estimation and probability matching. *Nature Precedings*, pages 1–1, 2011. https://doi.org/10.1371/journal.pcbi.1002282 PMID: 22125487

30. Tenenbaum Joshua B, Griffiths Thomas L, and Kemp Charles. Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, 10(7):309–318, 2006. https://doi.org/10.1016/j.tics.2006.05.009 PMID: 16797219

31. Denison Stephanie, Bonawitz Elizabeth, Gopnik Alison, and Griffiths Thomas L. Rational variability in children's causal inferences: The sampling hypothesis. *Cognition*, 126(2):285–300, 2013. https://doi.org/10.1016/j.cognition.2012.10.010 PMID: 23200511

32. Sanborn Adam N and Chater Nick. Bayesian brains without probabilities. *Trends in cognitive sciences*, 20(12):883–893, 2016. https://doi.org/10.1016/j.tics.2016.10.003 PMID: 28327290

33. Okasha Samir. The evolution of Bayesian updating. *Philosophy of Science*, 80(5):745–757, 2013. https://doi.org/10.1086/674058

34. Pandula Neel, Akçay Erol, and Morsky Bryce. Indirect reciprocity with abductive reasoning. *Journal of Theoretical Biology*, 580:111715, 2024. https://doi.org/10.1016/j.jtbi.2023.111715 PMID: 38154522

35. Shafer Glenn. A mathematical theory of evidence. In *A mathematical theory of evidence*. Princeton university press, 1976.

36. Fishman Michael A. Indirect reciprocity among imperfect individuals. *Journal of Theoretical Biology*, 225(3):285–292, 2003. https://doi.org/10.1016/S0022-5193(03)00246-7 PMID: 14604582

37. Sandholm William H. *Population games and evolutionary dynamics*. MIT press, 2010.

38. Taylor Peter D and Jonker Leo B. Evolutionary stable strategies and game dynamics. *Mathematical biosciences*, 40(1-2):145–156, 1978. https://doi.org/10.1016/0025-5564(78)90077-9

39. Sasaki Tatsuya, Okada Isamu, and Nakai Yutaka. The evolution of conditional moral assessment in indirect reciprocity. *Scientific reports*, 7(1):1–8, 2017. https://doi.org/10.1038/srep41870 PMID: 28150808

40. Körding Konrad P and Wolpert Daniel M. Bayesian decision theory in sensorimotor control. *Trends in cognitive sciences*, 10(7):319–326, 2006. https://doi.org/10.1016/j.tics.2006.05.003 PMID: 16807063

41. Hilbe Christian, Schmid Laura, Tkadlec Josef, Chatterjee Krishnendu, and Nowak Martin A. Indirect reciprocity with private, noisy, and incomplete information. *Proceedings of the national academy of sciences*, 115(48):12241–12246, 2018. https://doi.org/10.1073/pnas.1810565115 PMID: 30429320

42. Brandt Hannelore and Sigmund Karl. The good, the bad and the discriminator—errors in direct and indirect reciprocity. *Journal of theoretical biology*, 239(2):183–194, 2006. https://doi.org/10.1016/j.jtbi.2005.08.045 PMID: 16257417

43. Fiske Susan T. Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of personality and Social Psychology*, 38(6):889, 1980. https://doi.org/10.1037/0022-3514.38.6.889

44. Skowronski John J and Carlston Donal E. Negativity and extremity biases in impression formation: A review of explanations. *Psychological bulletin*, 105(1):131, 1989. https://doi.org/10.1037/0033-2909.105.1.131

45. Sharot Tali. The optimism bias. *Current biology*, 21(23):R941–R945, 2011. https://doi.org/10.1016/j.cub.2011.10.030 PMID: 22153158

46. Krueger Joachim. On the perception of social consensus. In *Advances in experimental social psychology*, volume 30, pages 163–240. Elsevier, 1998.

47. Suzuki Shinsuke and Kimura Hiromichi. Indirect reciprocity is sensitive to costs of information transfer. *Scientific reports*, 3(1):1435, 2013. https://doi.org/10.1038/srep01435 PMID: 23486389

48. Stevens Jeffrey R, Cushman Fiery A, and Hauser Marc D. Evolving the psychological mechanisms for cooperation. *Annu. Rev. Ecol. Evol. Syst.*, 36:499–518, 2005. https://doi.org/10.1146/annurev.ecolsys.36.113004.083814

49. Imhof Lorens A, Fudenberg Drew, and Nowak Martin A. Evolutionary cycles of cooperation and defection. *Proceedings of the National Academy of Sciences*, 102(31):10797–10800, 2005. https://doi.org/10.1073/pnas.0502589102 PMID: 16043717

50. Berra Irene. An evolutionary Ockham's razor to reciprocity. *Frontiers in psychology*, 5:1258, 2014. https://doi.org/10.3389/fpsyg.2014.01258 PMID: 25414681

51. McDowell Michelle and Jacobs Perke. Meta-analysis of the effect of natural frequencies on bayesian reasoning. *Psychological bulletin*, 143(12):1273, 2017. https://doi.org/10.1037/bul0000126 PMID: 29048176

52. Simpson Brent and Willer Robb. Altruism and indirect reciprocity: The interaction of person and situation in prosocial behavior. *Social Psychology Quarterly*, 71(1):37–52, 2008. https://doi.org/10.1177/019027250807100106

53. Morsky Bryce and Akçay Erol. False beliefs can bootstrap cooperative communities through social norms. *Evolutionary Human Sciences*, 3, 2021. https://doi.org/10.1017/ehs.2021.30 PMID: 37588567

54. Brenner Lyle A, Koehler Derek J, Liberman Varda, and Tversky Amos. Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes*, 65(3):212–219, 1996. https://doi.org/10.1006/obhd.1996.0021

55. Doyle Jon. Rationality and its roles in reasoning. *Computational Intelligence*, 8(2):376–409, 1992. https://doi.org/10.1111/j.1467-8640.1992.tb00371.x

56. Oaksford Mike, Chater Nick, et al. *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press, 2007.