Journal of Theoretical Biology 580 (2024) 111715

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Indirect reciprocity with abductive reasoning

Neel Pandula [a], Erol Akçay [a], Bryce Morsky [a,b,*]

[a] Department of Biology, University of Pennsylvania, Philadelphia, PA, USA
[b] Department of Mathematics, Florida State University, Tallahassee, FL, USA

ARTICLE INFO

Dataset link: https://github.com/bmorsky/indirectReciprocity-abduction

Keywords:
Abductive reasoning
Confirmation bias
Cooperation
Dempster–Shafer theory
Falsification bias
Indirect reciprocity

ABSTRACT

Indirect reciprocity is a reputational mechanism through which cooperative behavior can be promoted amongst a group of individuals. However, in order for this mechanism to effectively do so, cheating must be appropriately punished and cooperating appropriately rewarded. Errors in assessments and actions can hinder this process. In such a setting, individuals might try to reason about evidence to assign reputations given the possibility of errors. Here, we consider a well-established theory of reasoning used to combine evidence, abductive reasoning, as a possible means by which such errors can be circumvented. Specifically, we use Dempster–Shafer theory to model individuals who account for possible errors by combining information about their beliefs about the status of the population and the errors rates and then choose the simplest scenario that could explain their observations in the context of these beliefs. We investigate the effectiveness of abductive reasoning at promoting cooperation for five social norms: Scoring, Shunning, Simple Standing, Staying, and Stern Judging. We find that, generally, abductive reasoning can outperform non-reasoning models at ameliorating the effects of the aforementioned challenges and promote higher levels of cooperation under low-error conditions. However, for high-error conditions, we find that abductive reasoning can undermine cooperation. Furthermore, we also find that a degree of bias towards believing previously held reputations can help sustain cooperation.

1. Introduction

Cooperation has provided humans and various other species with a potent evolutionary advantage that has enabled them to weather a broad range of ecological scenarios (Dugatkin, 1997; Michod and Herron, 2006; Apicella et al., 2012; Apicella and Silk, 2019). In this sense, cooperation refers to altruistic behavior that comes at some cost at a lower socio-biological level, such as that of an individual, for the benefit of a higher socio-biological level, such as a population. The importance of such altruistic behavior to the success of human populations today cannot be understated. Cooperative behavior continues to play a vital role in many modern social systems, like those of innovation and business (Peña and de Arroyabe, 2002; de Faria et al., 2010).

There exists a number of well-studied mechanisms by which cooperation can be fostered amongst a group of players (Hauert et al., 2002; Rand et al., 2009; Okada, 2020). One such mechanism is that of indirect reciprocity, by which cooperation is rewarded indirectly, i.e. by those other than the individual receiving the benefit. A reputation system facilitates this mechanism. Individuals observe how others behave and then assign reputations in accordance with the social norm, which provides the metric by which individuals judge actions and subsequently assign reputations. Individuals are indirectly punished or rewarded for these reputations and, by extension, their behaviors by a class of players called Discriminators, who cooperate with (reward) those whom they deem "good" and defect against (punish) those whom they deem "bad". This mechanism and the role it plays in regard to encouraging cooperation have been studied thoroughly across a range of disciplines through both empirically and theoretically studies (Stanca, 2009; Akçay et al., 2010; Kato-Shimizu et al., 2013; Sasaki et al., 2017; Yoeli et al., 2013).

Theory shows that some social norms for assigning reputations are more effective than others at promoting cooperation (Ohtsuki and Iwasa, 2004). For instance, the Staying norm often leads to a higher degree of cooperation when compared to other norms, like Simple Standing (Sasaki et al., 2017; Okada et al., 2018). An important factor in the efficacy of norms in promoting cooperation is to what degree individuals agree on each others' reputations. If assessment errors can occur or reputations are held and assessed privately, then disagreements between Discriminators on reputations can arise (Hilbe et al., 2018). Players fitting the criteria of "good" can be mistakenly identified as "bad" by some observers and vice versa. This can lead to "good" players being punished and/or "bad" players being rewarded, both of

* Corresponding author at: Department of Mathematics, Florida State University, Tallahassee, FL, USA.
E-mail address: bmorsky@fsu.edu (B. Morsky).

https://doi.org/10.1016/j.jtbi.2023.111715
Received 7 July 2023; Received in revised form 12 December 2023; Accepted 15 December 2023
Available online 27 December 2023

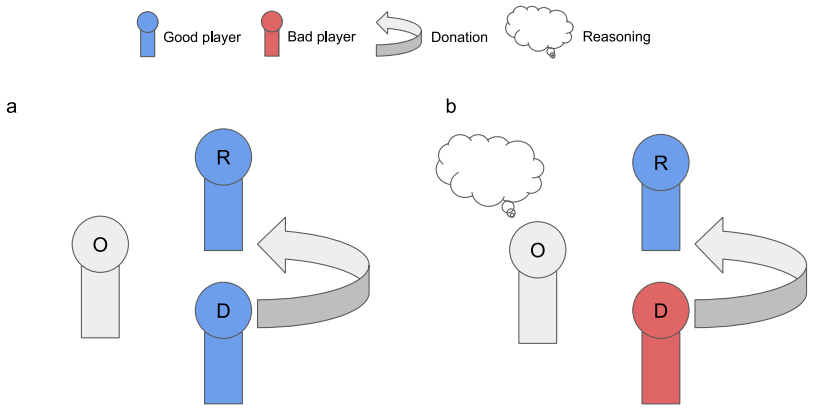

Fig. 1. Suppose that the social norm dictates that good players donate to good players while bad players do not. (a) An observer witnesses an interaction that matches their beliefs and thus makes no attempt to reason and account for errors. (b) An observer witnesses an interaction that does not match their beliefs. Therefore, they attempt to reason and account for errors that would explain the inconsistencies between their beliefs and what they observed.

which discourage cooperation. Recent research has focused on mechanisms that can overcome these problems, like empathy and action generosity (Radzvilavicius et al., 2019; Schmid et al., 2021; Kessinger et al., 2023).

Assigning reputation in the presence of errors creates a further inference problem for an individual observing an interaction, even under public information about reputations. Most indirect reciprocity literature sidesteps the inference problem, assuming individuals are unaware of the potential for errors and do not account for them. Here, we assume individuals know there are errors in behavior and observation, and try to reason about reputations based on this knowledge. Specifically, we focus on abductive reasoning (Douven, 2021; Paul, 1993; Thagard and Shelley, 1997), in which individuals aim to explain unexpected or unexplained phenomena by seeking the simplest conclusions possible, as one such remedy to the effects of the above challenges to indirect reciprocity.

In our model, players are aware of the possibility and probabilities of errors and use this information in conjunction with their beliefs about the makeup of the population. To do that, observers attempt to match interactions they observe to the reputation assessments of the given social norm in terms of the perceived reputations of the players involved and the donor's action in the "simplest" fashion possible. "Simplest" here refers to the fashion in which observers can assume the least number of errors. To do so, we employ Dempster–Shafer Theory, a framework for evaluating beliefs and uncertainty (Shafer, 1976a; Yager and Liu, 2008). Observers then update the reputation of the donor to fit the reputation assessment to which they matched the observed interaction.

For instance, consider an observer witnessing a donor they believe is good donating to a good recipient, and that the social norm dictates that you are a good person if you give to a good person. This observation makes sense within the context of the social norm, whereas a bad person donating to a good recipient is non-sensical (because someone who donates to a good recipient is by definition good). When a good person is perceived to be donating to a good person, it could be that the observer is wrong about the reputations and/or whether the donation occurred or not. However, the most parsimonious explanation is that there are no errors: what was observed is the truth. On the other hand, when a bad donor is observed to be giving to a good recipient, there *must* be at least one error assuming the truth of the social norm: one or both of the reputations are incorrect, or the donor did not donate. In such a case, observers use information about error rates and population composition to determine whether the donor being good is more feasible than them being bad, as is depicted in Fig. 1. We do this by combining the evidence from an observer's observation and their beliefs using Dempster's Rule of Combination (Dempster, 1967). Thus, in a sense, observers in our model perceive interactions they observe that match their beliefs (as determined by the social norm) as sensical, expected phenomena that do not call for reasoning, reevaluation, or explanation while perceiving interactions that do not as non-sensical, unexpected phenomena that do. Observers explain these unexpected phenomena by considering errors that might have caused a sensical, expected phenomena to appear so and readjust their beliefs to account for such errors. We show that players using such reasoning to judge actions and assign reputations can dramatically alter the outcomes of indirect reciprocity when compared to a system of non-reasoning players.

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

Social norm	Donor's action i and recipient's reputation j (ij)			
	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

2. Methods

2.1. Indirect reciprocity

Consider an infinite population in which individuals interact at random and in pairwise fashion. In each of these pairwise interactions, one individual is designated the donor player while the other is designated the recipient. The donor player may, at some cost to themselves, choose to transfer benefit unto the recipient player. We refer to such action as "cooperation". The donor player may also, at no cost to themselves, choose to not transfer benefit unto the recipient player. We refer to such action as "defection". Furthermore, a third individual, the observer player, bears witness to this interaction and updates their reputation of the donor based off the donor's action and the reputation of the recipient. How the observer judges the donor player based off this information depends on the social norm in play (as detailed in Table 1).

Errors can occur in this system, which can lead to observers misjudging donors. The first type of error is unintended defection, in which a donor who intends to cooperate instead defects. With probability $e_1 < 1/2$ such an error occurs. Note that there is no unintended cooperation. The second type of error is error in observation/assessment, in which

the donor's action is incorrectly observed or assessed as the opposite by the observer. With probability $e_2 < 1/2$ such an error occurs. We can consider this type of error as either an error in observation or assessment: only the interpretation of the observer's reasoning, which occurs after, changes. If the error is observational, then the observer will evaluate whether or not they saw something incorrect. If it is an error of assessment, then the observer will evaluate whether or not they made an initial error in their assessment of the donor's action. Given these two types of errors, the probability that a donor who intends to cooperate being observed as doing so is $\epsilon = (1-e_1)(1-e_2)+e_1e_2$. And, the probability that a donor who intends to cooperate being observed as not doing so is $1-\epsilon$. The probability that a donor who intends to defect being observed as not doing so is e_2 and the probability that a donor who intends to defect being observed as doing so is $1-e_2$. We define $e = e_2$ to simplify notation as is commonly done in the literature.

There are three strategies which players can adopt: Always Cooperate (AllC), Always Defect (AllD), and Discriminate (Disc). When acting as a donor player, an AllC player will always intend to cooperate with the recipient. Likewise, when acting as a donor player, an AllD player will always defect with the recipient. On the other hand, when acting as a donor player, a Discriminator will either donate or defect depending on the reputation the recipient player has with them. Discriminators will cooperate with recipients they believe to be good and defect with recipients they believe to be bad.

The payoffs of different strategies are a function of the composition of the population and the reputations of the different player strategy groups as a whole. Let π_i be the expected payoff to type i, then:

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

where x, y, and z are the frequencies of AllC, AllD, and Disc players, respectively, g_i is the frequency of good i individuals, and $g = g_x x + g_y y + g_z z$ is the frequency of good individuals. Also note that $r > 1$ represents the benefit to cost ratio of cooperating. g_i and g are equilibrium values for the reputation dynamics, since we assume that reputations reach equilibrium much more rapidly than strategies change. Additionally, note the cost to AllC players is 1, since they always cooperate, and the cost to Disc players is g, because they only cooperate with players they deem good. Over time, players shift their play styles to strategies that are more successful (those that have higher payoffs). To model the change in frequencies of strategies, we employ the replicator equations:

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

where $\bar{\pi} = \pi_x x + \pi_y y + \pi_z z$ is the average payoff in the population.

2.2. Abductive reasoning

Abductive reasoning is a form of reasoning in which observers aim to explain unexpected or unexplained phenomena by seeking the simplest conclusions possible. In a sense, abductive reasoners are "lazy". They use parsimonious reasoning, or Occam's razor, wherein the simplest explanation to the observation is assumed to be true. Here, simplest is the explanation that requires the fewest amount of errors to make an unexpected observation consistent with expectations, as we elaborate below.

We model abductive reasoning using Dempster–Shafer theory, which was developed to model belief, reasoning about uncertainty, and combining information (Shafer, 1976b). Here, we are combining individuals' beliefs about what can occur with what they observe. Individuals' beliefs about what can occur are informed by the social norm. For example, it cannot be that a bad donor gives to a good person and there are no errors, since under any norm such a donor is defined as good. However, it is possible that such a state was observed due to errors in reputation or observation. These possible states that explain the observation and are consistent with the social norm form a set of states that the observers believe can occur. The likelihood of each state is determined by the error rates, the observers' confidence in the reputations of others, and the frequency of good individuals in the population, which we assume are known by the observer. Using Dempster's rule of combination, an observer combines this information about what they believe are the possible states that can occur with what they have observed to determine whether or not a donor is good.

More formally, we define C as the set of possible states that can be observed in an interaction, and the errors — if any — that need to have happened for that observation to be consistent with a social norm. Specifically, each element of this set C includes what the observer saw with respect to the reputations of the donor and recipient and the action of the donor, and what, if any, errors needed to have occurred to generate that observation. The last element is determined under the assumption that reputations, intentions, and observations in the absence of any errors have to be consistent with the norm, so any combination of these inconsistent with the norm implies an error somewhere. To illustrate, suppose the population norm is Simple Standing and an observer observes a good donor giving to a good recipient. This observation is consistent with the expected behavior of a good donor under Simple Standing, and requires no errors anywhere to explain. This means that the state (Donor = Good, Action = Cooperate, Recipient = Good, no errors) $\in C$, i.e. observers believe it is a state of the world that could exist. Note that even when errors are not required to explain the observation, they might still have happened: the true state of the world might have been that the donor was bad and did not give to a good recipient, but the observer was wrong about their prior belief of the reputation of the donor and there was an error in the observation. Thus, another state of the world associated with the same observation, and therefore an element of the set C, is (Donor = Good, Action = Cooperate, Recipient = Good, errors in the reputation of donor and the observation). In this case, the observer would believe that they were initially wrong about the reputation of the donor: they are bad. There are, of course, additional states associated with this observation that includes different hypothesized "true" states and different sets of errors. Thus, each observation is associated with a set of errors that would be required for that observation to be consistent with the social norm.

Though there may be many explanations for an observation such that it is consistent with the norm, we assume under abductive reasoning that observers will assume the most parsimonious explanation, i.e. the one with the fewest amount of errors. Consider again the Simple Standing norm. If an observer sees a good person giving to a good person, then the simplest explanation — the one with the fewest errors from C — is that there are no errors, the state (Donor = Good, Action = Cooperate, Recipient = Good, no errors). Thus, an observer who sees this will assume that the donor is good, even though it is possible that the donor was bad and did not give, but the observer made an error in both the donor reputation and the action. This is what we call "lazy" reasoning: the observer will not consider errors if the observation is consistent with the norm at face value. On the other hand, consider an observer who sees a good donor *not* giving to a good person. Here, the observer must infer that there has been at least one error because this state should not happen under Simple Standing. The observer could be wrong about thinking the donor was good, or the observer might have mistaken the reputation of the recipient, or there might have been an error in the action or observation: the donor unintentionally defected or the observer made an error in assessing/observing the donation. Any single one of these errors is enough to explain the discrepancy between the observation and the expectation from the norm. However, there are additional explanations that requires two errors: for example the observer could be wrong about the reputation of the recipient (it was bad instead of good and the donor therefore did not give) and there was an error in the observation of the action. But this explanation (requiring two kinds of errors) is less parsimonious for this observation so the observer will not consider it. Note that for some norms, it is possible that the most parsimonious explanation includes two errors.

Since there may be several explanations that require the least amount of errors to make the observation consistent with the social norm, the observer will weigh the evidence of each to determine whether or not the donor is good. The weight of an explanation (different from the probability of it) is the product of the weights for each aspect of the observation, namely the donor's reputation, the recipient's reputation, and the action including whether or not there are errors. The weight for a reputation is simply the frequency of that reputation in the population multiplied by a parameter ψ, which denotes the relative weight the observer gives to their original beliefs about a reputation. Thus, the weight of a donor being good with no errors in reputation is $g\psi$, and the weight of a donor being good when the observer originally thought they were bad is $g(1-\psi)$. Similarly, the weight of an observing being bad with no errors is $(1-g)\psi$ and with errors is $(1-g)(1-\psi)$. The weights for reputations of recipients is calculated similarly. The parameter ψ modulates whether the observer has a confirmation or falsification bias. For $\psi > 1/2$, observers have a confirmation bias, meaning they give a larger weight to plausible explanations of their observations in which their original beliefs about the reputations were correct than in those in which they are incorrect. i.e. they have a bias towards believing that errors in observation and action (e_1 and e_2) better explain their observations than incorrect beliefs about the reputations of those they observe. For $\psi < 1/2$, the bias is in the other direction. For $\psi = 1/2$, observers have no bias either way. As for the weight of the action, it is the probability that the action is observed given the norm, the reputation of the donor, and whether or not there are errors. For example, the weight for a donor observed to be giving with no errors in the action (i.e. no errors on the part of the donor or the observer's assessment of the action) is ϵ. Returning to the previous example of a good donor not giving to a good recipient under Simple Standing, the weight of the explanation that there was only an error in the reputation of the donor is thus $(1-g)(1-\psi)(1-e)g\psi$. The frequency of bad individuals is $1-g$ and the observer was wrong about the reputation of the donor, so the first terms are $(1-g)(1-\psi)$. Given that the explanation of the observation is that the donor was bad, then the weight of observing defection from a bad individual is the probability that there are no errors of assessment/observation, i.e. $1-e$, and thus the next term in the weight. Finally, the recipient is, as originally believed, good. Thus, the final terms of the weight are g and ψ.

Given the weights of the possible explanations that are consistent with the norm (i.e. elements of C) for an observation, an observer will believe that the donor is good with a probability that is the sum of all the weights in which they are good divided by the sum of all the weights in which they are good or bad. Using the example of a good donor not giving to a good recipient under Simple Standing, we have the following sum of weights in which the donor is good: $g\psi(1-\epsilon)g\psi + g\psi(1-e)(1-g)(1-\psi)$. The first of which is where there was an error in the action, and the second where there was an error in the reputation of the recipient. Note that we assume that there has only been one error here, and thus in the second case the good donor intended to not give to the bad recipient (which is considered a good action under Simple Standing) and this was observed without an error. Thus, the weight for the action is simply $1-e$. The weight of the explanation in which the donor is bad is $(1-g)(1-\psi)(1-e)g\psi$, i.e. the error is in the reputation of the donor. These are the only explanations that have one error, and thus are the only considered by the observer. Therefore, the probability that the donor is good given these explanations is:

$$P(\text{Donor=Good}) = \frac{g\psi(1-\epsilon)g\psi + g\psi(1-e)(1-g)(1-\psi)}{g\psi(1-\epsilon)g\psi + g\psi(1-e)(1-g)(1-\psi) + (1-g)(1-\psi)(1-e)g\psi}. \tag{3}$$

This weighing of evidence is equivalent to applying Dempster's rule of combination from Dempster–Shafer Theory to two sets of evidence: the observation and the set of explanations consistent with the social norm. In the Appendix we provide these mathematical details.

Table 2
Possible observations.

Observation	Donor's reputation	Observed action	Recipient's reputation
$\mathcal{O}_1$	Bad	Defect	Bad
$\mathcal{O}_2$	Bad	Defect	Good
$\mathcal{O}_3$	Bad	Cooperate	Bad
$\mathcal{O}_4$	Bad	Cooperate	Good
$\mathcal{O}_5$	Good	Defect	Bad
$\mathcal{O}_6$	Good	Defect	Good
$\mathcal{O}_7$	Good	Cooperate	Bad
$\mathcal{O}_8$	Good	Cooperate	Good

Note that an assumption we make is that in both public and private assessment, players have accurate information about the frequency of good players within the population, even if under private assessment they might disagree over particular individuals' reputations. We also assume that players know the error rates accurately. We argue that is a reasonable assumption if the players can learn these rates (which do not change) from their observations and interactions, and through gossip and other methods of information transfer between individuals. We more thoroughly discuss these limitations in the Discussion.

2.3. Indirect reciprocity with abductive reasoning

Here we detail how we combine the model of indirect reciprocity with abductive reasoning. In the context of the reputations of the donor and recipient and the action, there are eight possible states that could be observed, depicted in Table 2. We let $P_i = P(G|\mathcal{O}_i)$ represent the probability that the donor is good given that the observer observed $\mathcal{O}_i$. Note that the observer will potentially update the donor's reputation given these probabilities.

The probability that the donor is assessed as good given the observer's beliefs about their actions, reputation, and the recipients' reputations is Q_{ijk}, where i is the donor's reputation, j is the intended action of the donor, and k is the recipient's reputation. For instance, Q_{BDB} represents the probability that a bad donor is good given that they intended to defect with a bad recipient. As such, Q_{BDB} is the probability the donor is good given they intended to defect and were observed correctly doing so (which occurs with probability $(1-e)P_1$) plus the probability the donor is good given they were observed incorrectly (which occurs with probability eP_3). Generally, Q_{ijk} is as follows:

$$\begin{aligned} Q_{BDB} &= (1-e)P_1 + eP_3, & Q_{BDG} &= (1-e)P_2 + eP_4, \\ Q_{BCB} &= (1-\epsilon)P_1 + \epsilon P_3, & Q_{BCG} &= \epsilon P_4 + (1-\epsilon)P_2, \\ Q_{GDB} &= (1-e)P_5 + eP_7, & Q_{GDG} &= (1-e)P_6 + eP_8, \\ Q_{GCB} &= \epsilon P_7 + (1-\epsilon)P_5, & Q_{GCG} &= (1-\epsilon)P_6 + \epsilon P_8. \end{aligned} \tag{4}$$

Recall that g_x, g_y, and g_z are the frequencies of good AllC players, AllD players, and Discriminators, respectively. Under public assessment of reputations, these reputations at equilibrium must satisfy:

$$\begin{aligned} g_x &= Q_{GCG}g_x g + Q_{GCB}g_x(1-g) + Q_{BCG}(1-g_x)g + Q_{BCB}(1-g_x)(1-g), \\ g_y &= Q_{GDG}g_y g + Q_{GDB}g_y(1-g) + Q_{BDG}(1-g_y)g + Q_{BDB}(1-g_y)(1-g), \\ g_z &= Q_{GCG}g_z g + Q_{GDB}g_z(1-g) + Q_{BCG}(1-g_z)g + Q_{BDB}(1-g_z)(1-g). \end{aligned} \tag{5}$$

Conversely, under private assessment, we need to keep track of not just the reputations of the different types, but also the probability that two randomly drawn individuals agree on the reputation of a third party. The latter we denote by g_{i2}, where i can be x, y, or z. We denote the probability that an individual of type i will newly acquire good reputation as g_i^+, and the probability they will lose it as g_i^- (and likewise, g_{i2}^+ and g_{i2}^- for the probabilities that a type i individual will gain and lose agreement over its good reputation, respectively). Given

Table 3
Definitions of parameters.

Parameters/variables	Definition
e_1	Prob. donor intending to cooperate defects.
$e = e_2$	Prob. donor's action is observed as its the opposite.
$\epsilon = (1-e_1)(1-e) + e_1 e$	Prob. a donor intending to cooperate is observed doing so.
ψ	Weighting of prior beliefs about reputations.
r	Cost to benefit ratio.
x	Frequency of AllC players.
y	Frequency of AllD players.
z	Frequency of Disc players.
g_i	Frequency of good i players.
g_{i2}	Prob. two random players agree on i's reputation.
$P_i = P(G\|\mathcal{O}_i)$	Prob. donor is good given observation $\mathcal{O}_i$.
Q_{ijk}	Prob. donor is good given reputations i (donor) and k (recipient), and intended action j.

these definitions, we can write the probabilities of change in reputation as:

$$\begin{aligned}
g_x^+ &= (1-g_x)(Q_{BCG}g + Q_{BCB}(1-g)),\\
g_x^- &= g_x((1-Q_{GCG})g + (1-Q_{GCB})(1-g)),\\
g_{x2}^+ &= (g_x - g_{x2})(Q_{BCG}g + Q_{BCB}(1-g)),\\
g_{x2}^- &= g_{x2}((1-Q_{GCG})g + (1-Q_{GCB})(1-g)),\\
g_y^+ &= (1-g_y)(Q_{BDG}g + Q_{BDB}(1-g)),\\
g_y^- &= g_y((1-Q_{GDG})g + (1-Q_{GDB})(1-g)),\\
g_{y2}^+ &= (g_y - g_{y2})(Q_{BDG}g + Q_{BDB}(1-g)),\\
g_{y2}^- &= g_{y2}((1-Q_{GDG})g + (1-Q_{GDB})(1-g)),\\
g_z^+ &= (1-g_z)(Q_{BCG}g_2 + (Q_{BCB}+Q_{BDG})(g-g_2) + Q_{BDB}(1-2g+g_2)),\\
g_z^- &= g_z(Q_{GCG}g_2 + (Q_{GCB}+Q_{GDG})(g-g_2) + Q_{GDB}(1-2g+g_2)),\\
g_{z2}^+ &= (g_z - g_{z2})(Q_{BCG}g_2 + (Q_{BCB}+Q_{BDG})(g-g_2) + Q_{BDB}(1-2g+g_2)),\\
g_{z2}^- &= g_{z2}(Q_{GCG}g_2 + (Q_{GCB}+Q_{GDG})(g-g_2) + Q_{GDB}(1-2g+g_2)).
\end{aligned} \tag{6}$$

For example, an individual of type AllC will newly acquire a good reputation when a bad AllC individual (which has frequency $1-g_x$) is evaluated as good. This occurs with probability Q_{BCG} when they interact with a good recipient (which occurs with probability g), and Q_{BCB} when they interact with a bad recipient (which occurs with probability $1-g$).

The steady state solutions are when $g_i^+ = g_i^-$, which are required to solve for the values of g_i. For numerical simulations and some proofs, we define the set of differential equations $\dot{g}_i = \tau(g_i^+ - g_i^-)$ with $\tau \gg 1$ to represent the dynamics of reputations. In other words (in line with other studies of indirect reciprocity), we assume that reputation dynamics operate much faster than the dynamics of type frequencies, given by the replicator dynamics, Eq. (2).

3. Results

Here we detail the simulation results and analysis of the qualitative behavior of the model under the different norms. Table 3 provides of a summary of the definitions of critical parameters and variables used throughout the results.

3.1. Scoring

Under the Scoring norm, the probabilities that the donor is assessed as being good given the state the observer sees are:

$$\begin{aligned}
&P_1 = 0, \quad P_3 = \frac{g(1-\psi)\epsilon}{(1-g)\psi e + g(1-\psi)\epsilon},\\
&P_5 = \frac{g\psi(1-\epsilon)}{(1-g)(1-\psi)(1-e) + g\psi(1-\epsilon)}, \quad P_7 = 1.
\end{aligned} \tag{7}$$

Note that the reputation of the recipient is not relevant to assessment under the Scoring norm and therefore is not a factor in determining the reputation of the donor. Therefore, $P_n = P_{n+1}$ for odd n. The probabilities that donors are good given there intentions are thus:

$$Q_{BD} = eP_3, \quad Q_{BC} = \epsilon P_3, \quad Q_{GD} = (1-e)P_5 + e, \quad Q_{GC} = \epsilon + (1-\epsilon)P_5. \tag{8}$$

Under public assessment, the reputations at equilibrium must satisfy:

$$\begin{aligned}
&(1-g_x)Q_{BC} = g_x(1-Q_{GC}),\\
&(1-g_y)Q_{BD} = g_y(1-Q_{GD}),\\
&(1-g_z)(gQ_{BC} + (1-g)Q_{BD}) = g_z(g(1-Q_{GC}) + (1-g)(1-Q_{GD})).
\end{aligned} \tag{9}$$

Now, under private assessment, the probabilities of reputation changes are:

$$\begin{aligned}
g_x^+ &= (1-g_x)Q_{BC}, & g_x^- &= g_x(1-Q_{GC}),\\
g_y^+ &= (1-g_y)Q_{BD}, & g_y^- &= g_y(1-Q_{GD}),\\
g_z^+ &= (1-g_z)(Q_{BC}g + Q_{BD}(1-g)), & g_z^- &= g_z((1-Q_{GC})g + (1-Q_{GD})(1-g)).
\end{aligned} \tag{10}$$

The steady state solutions to $g_i^+ = g_i^-$ are identical to the case of public assessment. Intuitively, the behavior here is the same whether the assessment is public or private because the reputation of the recipient is inconsequential to determining reputations. Therefore, in assigning reputations it does not matter whether or not an observer and Disc donor agree on the reputation of the recipient. The following analyses thus apply to both public and private assessment under Scoring.

Theorem 3.1.1. *For $\psi = 1$, reputations do not change. Therefore, the dynamics are dependent on the initial conditions of the reputations.*

Proof. Recall that $P_1 = 0$ and $P_7 = 1$. For $\psi = 1$, we have

$$\begin{aligned}
P_3 &= \frac{g(1-1)\epsilon}{(1-g)(1)e + g(1-1)\epsilon} = 0,\\
P_5 &= \frac{g(1)(1-\epsilon)}{(1-g)(1-1)(1-e) + g(1)(1-\epsilon)} = 1,
\end{aligned} \tag{11}$$

and thus $Q_{BD} = Q_{BC} = 0$ and $Q_{GD} = Q_{GC} = 1$. However, Eqs. (9) are then satisfied for all g_i, and Eq. (10) gives us $g_i^+ = g_i^- = 0$ for all i. Since reputations do not change, they are simply parameters in the replicator equation. □

When $\psi = 0$, observers believe that the reputations they assign must be wrong. The equations are then identical to those for Scoring in non-reasoning models (Sasaki et al., 2017; Okada et al., 2018): a continuum of unstable equilibria parallel to the AllD-AllC boundary emerges and AllD is the sole stable equilibrium.

Theorem 3.1.2. *For $\psi = 0$, we have the identical model as that of public/private assessment with no reasoning as in Sasaki et al. (2017).*

Proof. For $\psi = 0$, we have $P_3 = 1$ and $P_5 = 0 \implies Q_{BD} = Q_{GD} = e$ and $Q_{BC} = Q_{GC} = \epsilon$. At equilibrium, reputations must then satisfy $g_x = \epsilon$, $g_y = e$, and $g_z = \epsilon g + e(1-g)$. However, these are the same equalities as in Sasaki et al. (2017), and thus the reputation dynamics and equilibria are identical. □

Theorem 3.1.3. *If $\psi > \epsilon/(1+\epsilon-e)$, there is a continuum of equilibria along the AllD-Disc boundary that is stable to perturbations outside of the AllD-Disc boundary.*

Proof. On the AllD-Disc boundary, $x = 0$ and $y = 1-z$, and $g = g_y(1-z) + g_z z$. We can say that $g_z > g_y$ so long as $g \neq 0$, since you are only good if you give, and Discriminators will be observed cooperating at least as much as defectors. Further this implies that $g_z > g > 0$ if $g \neq 0$. If $\psi > \epsilon/(1+\epsilon-e)$ and $g > 0$, then

$$\begin{aligned}
g_z^+ - g_z^- &= (1-g_z)(gQ_{BC} + (1-g)Q_{BD}) - g_z(g(1-Q_{GC}) + (1-g)(1-Q_{GD}))\\
&< (1-g)(gQ_{BC} + (1-g)Q_{BD}) - g(g(1-Q_{GC}) + (1-g)(1-Q_{GD}))
\end{aligned}$$

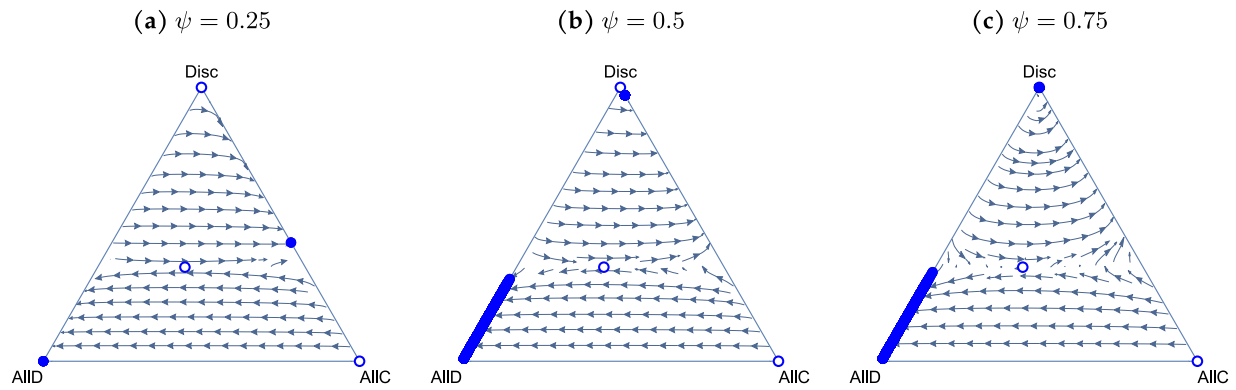

Fig. 2. Evolutionary dynamics of AllC, AllD, and Discriminators under the Scoring norm with abductive reasoning for $r = 3$, $e_1 = e_2 = 0.01$, and for different values of the weight of the observer's original beliefs, $\psi = 0.25, 0.5, 0.75$.

$$= -\frac{e_1(1-2e)(1-e)g^2(1-g)((\epsilon-e)^2(1-g)+(1-\epsilon)(1+\epsilon-e))}{(e+g(1-2e))(\epsilon(1-\epsilon)g+(1-e)^2(1-g))}$$
$$< 0,$$

and thus both reputations converge to 0.

To evaluate the stability, we consider the model as a system of differential equations for both reputation and replicator dynamics. We thus combine the replicator equations (Eq. (2)) with the differential equations $\dot{g}_i = \tau(g_i^+ - g_i^-)$, where $\tau \gg 1$ is the rate at which reputations change. The Jacobian for this system at $x = 0$ and $g = 0$ is:

$$J = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 \\ 1-z & 0 & 0 & z(1-z)(1+(r-1)z) & (r-1)z^2(z-1) \\ 0 & 0 & \tau(\epsilon-1) & \frac{\tau z(1-z)(1-\psi)\epsilon^2}{\psi e} & \frac{\tau z(1-\psi)\epsilon^2}{\psi e} \\ 0 & 0 & 0 & \tau(e-1+(1-z)(1-\psi)\epsilon/\psi) & \tau z(1-\psi)\epsilon/\psi \\ 0 & 0 & 0 & \tau(1-z)(1-\psi)\epsilon/\psi & \tau\,(e-1+z(1/\psi-1)\epsilon) \end{pmatrix}, \tag{12}$$

and the eigenvalues are:

$$\left\{0, -1, -\tau(1-e), -\tau(1-\epsilon), \frac{\tau(\epsilon-\psi(1+\epsilon-e))}{\psi}\right\}, \tag{13}$$

which, but for the zero eigenvalue, are negative for $\psi > \epsilon/(1+\epsilon-e)$. The zero eigenvalue corresponds to the eigenvector $(0, 1, 0, 0, 0)$ and thus indicates that the system is unstable along the AllD-Disc boundary. If $\psi \leq \epsilon/(1+\epsilon-e)$, $y^* = 1$ and $g^* = 0$ is stable. Since, the eigenvector for the positive eigenvalue is:

$$\left(0, \frac{z(z-1)\psi}{\tau(\psi(1-e)-(1-\psi)\epsilon)}, \frac{(1-\psi)\epsilon^2}{(\epsilon+\psi(e-2\epsilon))e}, 1, 1\right)$$
$$\xrightarrow{y^*=1} \left(0, 0, \frac{(1-\psi)\epsilon^2}{(\epsilon+\psi(e-2\epsilon))e}, 1, 1\right), \tag{14}$$

and thus the positive eigenvalues corresponds only to reputation space. □

We numerically solved the system of ODEs and discovered another equilibrium in which there is cooperation as depicted in Fig. 2. For $\psi = 0.25$, this equilibrium is on the AllC-Disc boundary (panel a). Increasing ψ (panels b and c), moves this equilibrium towards $z = 1$, the all Discriminators equilibrium. The population tends towards this equilibrium above a certain threshold of Discriminators. Below this threshold, the population goes towards entirely uncooperative equilibria. There, the population is composed of either all AllD players or a mix of AllD and Disc players on the Disc-AllD boundary. For $e_1 = e_2 = 0.01$, coexistence of Disc and AllD players occurs when $\psi > \epsilon/(1+\epsilon-e) = 0.497513$. When AllD players and Discriminators coexist, $g_z = g_y = g = 0$, and therefore Discriminators always defect. Thus, the system is bistable.

3.2. Shunning

For Shunning, $P_1 = P_2 = P_3 = P_5 = 0$, $P_8 = 1$, and

$$P_4 = \frac{\epsilon\psi(1-\psi)g^2}{e\psi^2 g(1-g)+\epsilon\psi(1-\psi)(1-g)^2+\epsilon\psi(1-\psi)g^2},$$
$$P_6 = \frac{(1-\epsilon)\psi^2 g^2}{(1-e)\psi(1-\psi)g(1-g)+(1-\epsilon)\psi^2 g^2},$$
$$P_7 = \frac{\epsilon\psi(1-\psi)g^2}{\epsilon\psi(1-\psi)(1-g)^2+\epsilon\psi(1-\psi)g^2} = \frac{g^2}{g^2+(1-g)^2}.$$

Thus, the probabilities that donors are good given their intentions and the recipients' reputations are:

$$Q_{BDB} = 0, \quad Q_{BDG} = eP_4, \quad Q_{BCB} = 0, \quad Q_{BCG} = \epsilon P_4,$$
$$Q_{GDB} = eP_7, \quad Q_{GDG} = (1-e)P_6 + e, \quad Q_{GCB} = \epsilon P_7, \quad Q_{GCG} = (1-\epsilon)P_6 + \epsilon. \tag{15}$$

Theorem 3.2.1. *$g = g_x = g_y = g_z = 0$ is a stable equilibrium, and Discriminators behave as defectors resulting in $\pi_z = \pi_y > \pi_x$. Thus, the AllD-Disc boundary is asymptotically stable.*

Proof. Let $g \leq 1/2$, and therefore $P_4 \leq P_7 \leq 1/2$. Now consider the dynamics for g_x^+ and g_x^- from Eq. (6) to show that $g_x \to 0$ recalling that $g_x \geq g$:

$$\begin{aligned} g_x^+ - g_x^- &= Q_{GCG}g_x g + Q_{GCB}g_x(1-g) + Q_{BCG}(1-g_x)g - g_x \\ &= Q_{GCG}g_x g + \epsilon P_7 g_x(1-g) + \epsilon P_4(1-g_x)g - g_x g \\ &\quad - g_x(1-g)(1+\epsilon P_7 - \epsilon P_7) \\ &= (Q_{GCG}-1)g_x g + \epsilon P_4(1-g_x)g - g_x(1-g)(1-\epsilon P_7) \\ &\leq \epsilon P_7(1-g_x)g_x - g_x(1-g_x)(1-\epsilon P_7) \\ &= (2\epsilon P_7 - 1)g_x(1-g_x) \leq (\epsilon-1)g_x(1-g_x) < 0. \end{aligned}$$

Therefore, $g = g_x = g_y = g_z = 0$, and Discriminators behave as defectors resulting in $\pi_z = \pi_y > \pi_x$. Note that this holds for both public and private assessment. □

The degree and type of bias with respect to reputations ψ nor the type of assessment affects this theorem. The AllD-Disc boundary is always the sole stable set of equilibria, because reputation dynamics always drive reputations to 0, and thus Discriminators play identically to AllD players. Fig. 3 depicts this result. This behavior is qualitatively identical to that of the Shunning norm in a model in which Bayes rule is used to account for observers' inference problem (Morsky et al., 2023).

3.3. Simple Standing, Staying, and Stern Judging

Fig. 4 depicts ternary plots for Simple Standing, Staying, and Stern Judging under both public and private assessment. In all but Stern

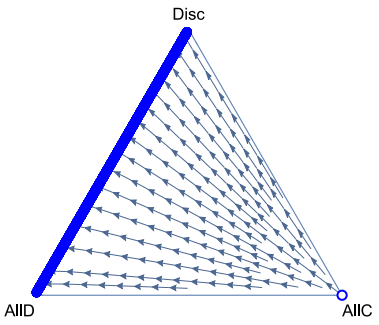

Fig. 3. Ternary figure for the Shunning norm for any value of ψ and either public or private assessment of reputation.

Judging under private assessment (wherein $y^* = 1$ is the sole stable equilibrium), a bistable system can emerge with the population either tending towards a stable state of no cooperation or a relatively cooperative one. The default error rates of the figures are $e_1 = e_2 = 0.01$ and $\psi = 0.5$. These results are similar to those of non-reasoning models (Sasaki et al., 2017; Okada et al., 2018). To compare these results to such non-reasoning models and to explore the effect of different error rates and degrees of bias, we plot the average degree of intended cooperation (i.e. the frequency of AllC players and Discriminators that cooperate), $x + gz$, at equilibrium initialized across strategy space for the abductive reasoning model and the non-reasoning model in heat maps in Figs. 5 and 6.

Fig. 5 plots the intended degree of cooperation for public assessment. In the case of Simple Standing, abductive reasoning fails to promote cooperation entirely for $\psi = 0.25$ and thus is less effective than non-reasoning with respect to promoting cooperation. However, when ψ is not low ($\psi = 0.5, 0.75$ in the first row of the figure), abductive reasoning can outperform non-reasoning when errors are low, but under performs when they are high. In the case of Staying, abductive reasoning outperforms non-reasoning for all ψ. Further, ψ has no appreciable effect on the degree of cooperation. In the case of Stern Judging, abductive reasoning outperforms non-reasoning and is more robust against errors when ψ is high (i.e. there is confirmation bias with respect to reputations). However, for $\psi = 0.25$ there is no cooperation, and thus abductive reasoning is inferior to non-reasoning. Unlike Simple Standing, Stern Judging with abductive reasoning is more tolerant of errors than non-reasoning for $\psi = 0.75$. Note that in all cases, abductive reasoning is more tolerant of errors in observation/assessment, e_2, than of errors in action, e_1.

Fig. 6 depicts the heat maps for the average degree of intended cooperation under private assessment. Note that we do not include heat maps for Stern Judging, since in the case of both abductive reasoning and non-reasoning, $y^* = 1$, a population of only AllD players, is the sole stable equilibrium. Similarly to public assessment, abductive reasoning under Simple Standing with private assessment of errors is superior in promoting intended cooperation relative to non-reasoning when ψ is intermediate to high ($\psi = 0.5, 0.75$). Further, so long as error rates are low, no bias results in more cooperation than confirmation bias. Abductive reasoning is also more robust against errors compared to non-reasoning in these cases. $\psi = 0.25$, however, results in no cooperation for abductive reasoning, while non-reasoning can still support cooperation. Under Staying, however, abductive reasoning does worse than non-reasoning when errors are large across the ψ values explored. Though abductive reasoning under Staying does promote cooperation, it is more undermined by errors than non-reasoning. As is the case with public assessment, we find that in all cases of private assessment explored, abductive reasoning is more tolerant of errors in observation/assessment, e_2, than of errors in action, e_1.

4. Discussion

We find that abductive reasoning can help to promote cooperation through indirect reciprocity. In the case of the Scoring norm, abductive reasoning can produce stable cooperative states whereas non-reasoning fails to do so entirely. Furthermore, abductive reasoning can outperform non-reasoning with respect to promoting cooperation under certain conditions in the case of other norms as well, namely Staying and Stern Judging under public assessment as well as Simple Standing under both public and private assessment. On the other hand, abductive reasoning results in only defection under Shunning whereas cooperation can be promoted under Shunning and non-reasoning (Okada et al., 2018). We also observe that regardless of norm or whether or not assessments are privately held, abductive reasoning is more tolerant of errors in observation/assessment, e_2, than of errors in action, e_1.

We further find that the effect of confirmation and falsification bias on our results varies widely. Existing literature typically regards confirmation bias as being an "epistemically problematic" tendency that hampers decision making and the development of accurate beliefs (Jones and Sugden, 2001; Peters, 2020). While this may be true on an individual level, we find that in the context of our study, confirmation bias can prove beneficial to a population as a whole, at least with regards to promoting cooperation and under certain conditions. For instance, in the case of Stern Judging under public assessment, confirmation bias not only improves the effectiveness of abductive reasoning at promoting cooperation but also its robustness against errors while falsification bias erodes them. On the other hand, confirmation bias undermines cooperation relative to no bias under Simple Standing, and has no large effect for Staying. Understanding why and how confirmation and falsification bias produce such widely varying results might provide yet a new direction for future research.

It is interesting to compare our model with abductive reasoning to the more conventional Bayesian reasoning about errors, recently studied by Morsky et al. (2023). Firstly, both abductive and Bayesian reasoning can produce stable cooperative states in the case of the Scoring norm while non-reasoning cannot. This suggests that the use of inference to resolve errors is particularly useful in promoting cooperation under Scoring. Without reasoning, errors of execution and observation under Scoring lead discriminators to defect against other discriminators because they are perceived as bad given errors. This leads either AllC or AllD to initially take over depending on the initial frequency of discriminators, and AllD to be the only stable equilibrium (Okada et al., 2018). Reasoning avoids this outcome, as accounting for the possibility of errors avoids mistakenly assigning bad reputations discriminators and can make them–and cooperation–part of a stable population.

Conversely, both abductive and Bayesian reasoning perform poorly under the Shunning norm, producing only entirely uncooperative stable states. This is because Shunning is a draconian norm, and most things an individual can do lands them with a bad reputation. Therefore, when considering errors, reasoners are in effect pessimistic about the reputation of the donor, which disadvantages the discriminators. In the cases of the other three norms we investigated, Simple Standing, Staying, and Stern Judging, abductive reasoning produces results that are similar to those produced by Bayesian reasoning under private assessment but not under public assessments (Morsky et al., 2023). Under public assessment, the Bayesian model is less robust against errors and produces less cooperation than non-reasoning. Our model, too, is generally less robust against errors than non-reasoning. However, for sufficiently low error rates, we find that abductive reasoning can outperform non-reasoning at promoting cooperation. The fact that abductive reasoning, which could be considered a relatively "simpler" reasoning mechanism, is able to outperform Bayesian reasoning in this regard even without the consideration of an explicit cognitive cost

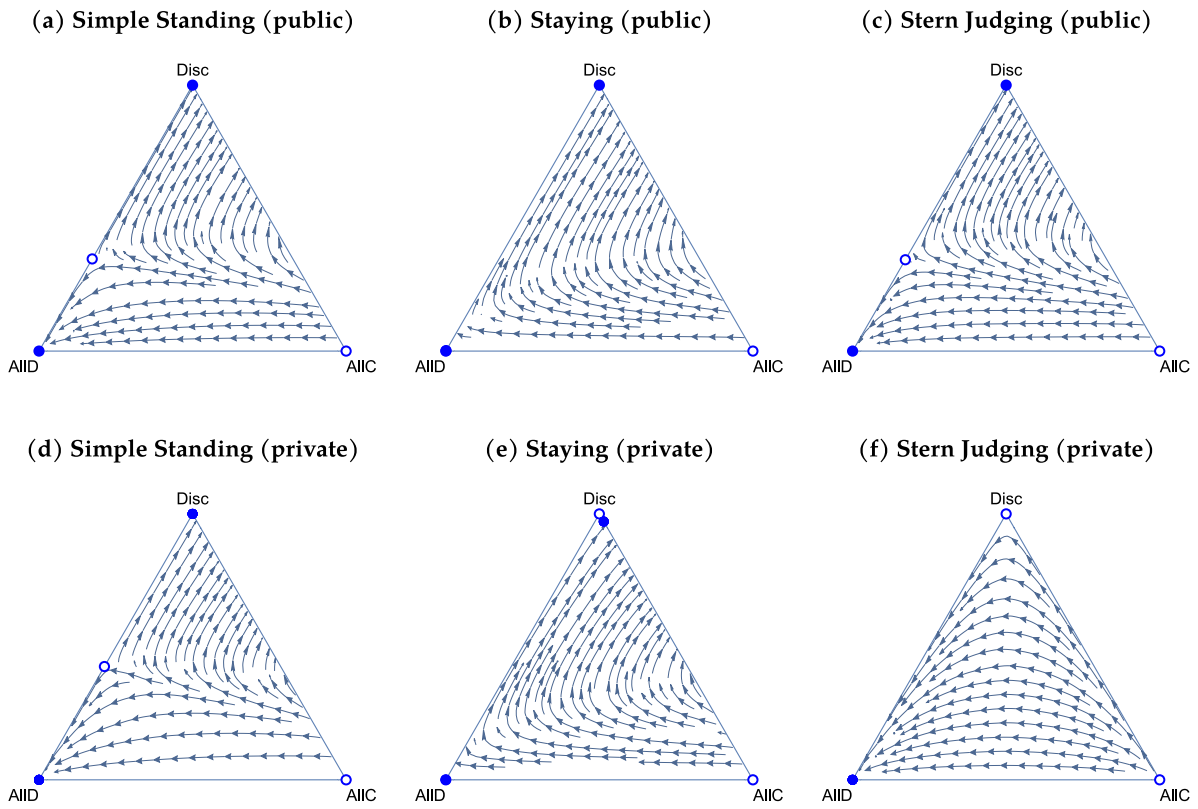

Fig. 4. Ternary figures for Simple Standing, Staying, and Stern Judging under private and public assessment of reputations. For all figures, $r = 3$, $e_1 = e_2 = 0.01$, and $\psi = 0.5$. Note that there is an unstable equilibrium on the AllD-Disc boundary very near $y = 1$ for Staying (public) that does not show in the figure.

highlights its effectiveness at dealing with the challenges of indirect reciprocity's inference problem.

One of the crucial assumptions we make is that in both public and private assessment, players were aware of and provided with accurate information about the error rates and the frequency of good players within the population, even if under private assessment they might disagree over particular individuals' reputations. These assumptions are relatively strong but not entirely unrealistic even under private reputations: mechanisms like learning and gossiping can and do plausibly provide individuals with a sufficiently large sample of the population from which they could make reasonably accurate estimates of error rates and the reputational makeup of the population (Sommerfeld et al., 2007). For example, a recent study by Dores Cruz et al. (2021) found that over a 10 day period, a Dutch sample of just over 300 individuals exchanged more than 5000 pieces of gossip about each other. Depending on the rate of interaction and the distribution of gossiped information, this rate can be enough to give each player a good estimate about the reputational composition of the population. Public institutions (Radzvilavicius et al., 2021) are another mechanism in which individuals could obtain such information. Public institutions could act as a centralized reputation tracking system, even if they do not broadcast individual reputations. The only information they would need to provide players would be information about the distribution of reputations (e.g. the current fraction of good players in the population).

In this paper, we consider abductive reasoning under what are termed zeroth or first order norms, where the reputation assigned to the donor at most depends on the action and the recipient's reputation. Indirect reciprocity literature has also considered higher order norms, such as a second order norm that also considers the donor's (starting) reputation to assign the donor a new reputation. In a sense, this is exactly what abductive reasoning does as well, but instead of being a simple assignment function, it works through the potential sources of errors given the expectation that behavior should be consistent a first or zeroth order norm. In a sense, then, abductive reasoning can be seen as a way to reason about higher order norms.

We have not explicitly considered any cognitive costs of reasoning. Given that abductive reasoning is only used when there is a conflict in the information received, making it a relatively simple reasoning mechanism in relation to some other forms of probabilistic reasoning. As we show, abductive reasoning can sustain higher levels of cooperation, but it is an open question whether this population level benefit will be individually beneficial. Nonetheless, given the relative frugality of abductive reasoning, we hypothesize that it might be more common than more complicated methods of reasoning.

Finally, applying our model in the context of a finite population presents a direction for future research. Although intuition from deterministic models do not always carry over to finite population models, we are reasonably confident that at least reputation dynamics would not be affected unduly in large enough populations. In other work, we find that global convergence of reputations is generally robust and that the dynamics are therefore similar in a sufficiently large finite population (Morsky et al. unpublished results). That being said, dynamics under Scoring may be the most promising area for future research with finite populations as it displays an interesting bistability. Additionally, investigating additional reasoning mechanisms within the framework of indirect reciprocity, perhaps such as inductive reasoning or analogical reasoning, against which our study might be compared might also provide new directions for future research.

Abduction is widely recognized by both philosophers and psychologists as being one of three major types of logical inference, the other two being deduction and induction. Of these three, abduction is regarded by many as being the most ubiquitous, playing a major role in everyday reasoning (Douven, 2021). Given this prevalence and its ability to promote cooperation within the framework of indirect reciprocity, a well-established mechanism by which cooperation can be fostered amongst a group of players, even in the presence of errors and

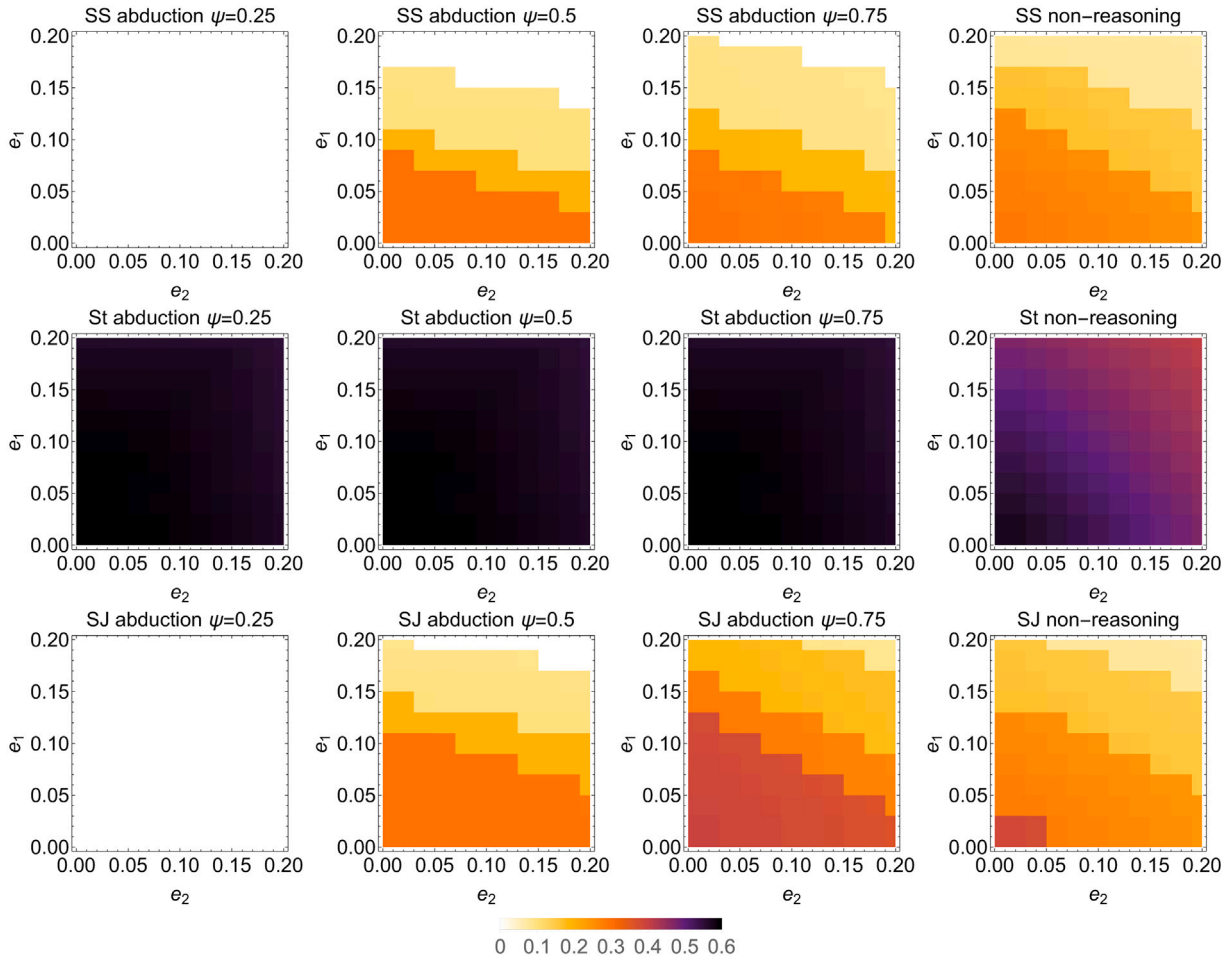

Fig. 5. Heat maps of average degree of intended cooperation $(x+gz)$ for Simple Standing (SS), Staying (St), and Stern Judging (SJ) averaged over the simplex for public assessment of reputation and $r = 3$.

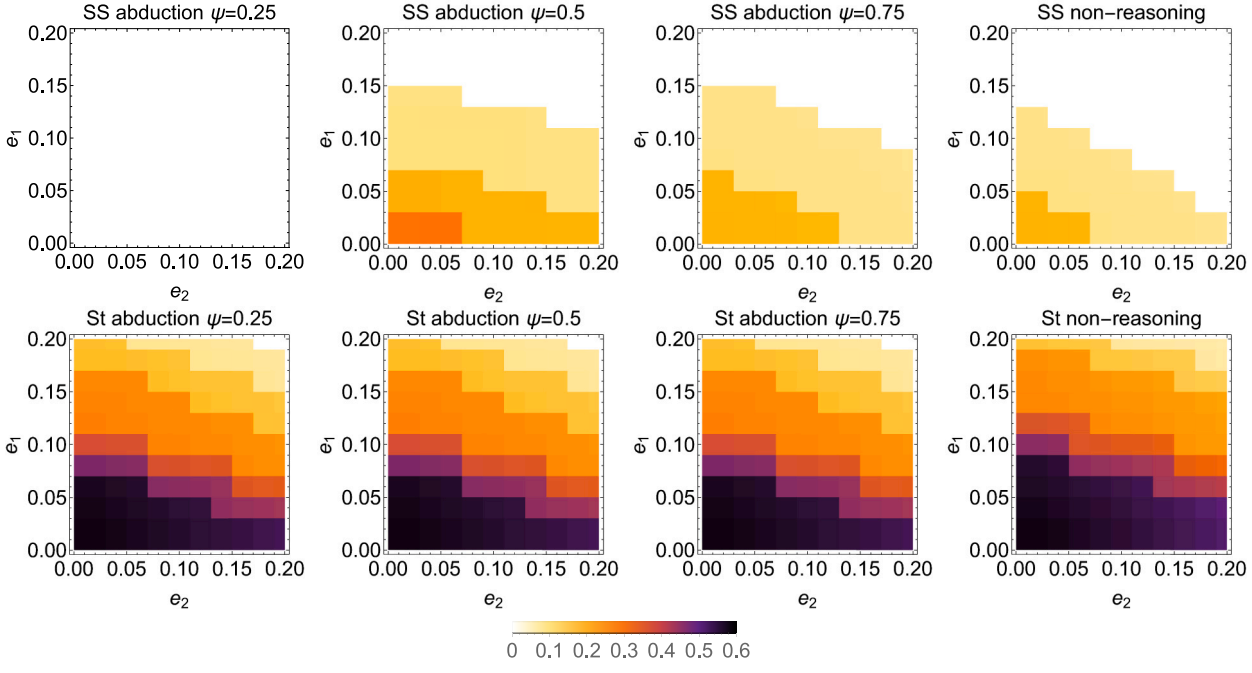

Fig. 6. Heat maps of average degree of intended cooperation $(x + gz)$ for Simple Standing (SS) and Staying (St) averaged over the simplex for private assessment of reputation and $r = 3$. Note that Stern Judging is not presented, because it always leads to defection.

noise, it is possible that abductive reasoning plays a crucial role in the development and maintenance of cooperative social networks in both the human and animal worlds.

CRediT authorship contribution statement

Neel Pandula: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Erol Akçay:** Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Bryce Morsky:** Conceptualization, Formal analysis, Investigation, Methodology, Software, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

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

Appendix. Dempster–Shafer Theory formalism

Here we provide mathematical details on the Dempster–Shafer Theory formalism of the model. Let $\Theta = \{\theta_1, \theta_2, \dots \theta_n\}$ be the set of states of the world. Each state θ_i is a tuple that includes an observation and a belief about any errors: $\theta_i = (\text{Rep. of donor}, \text{Observed action}, \text{Rep. of recipient}, \text{Errors})$. More succinctly, we can write $\theta_i = (\mathcal{O}_j, \text{Errors})$ where $\mathcal{O}_j$ are the possible observations listed in Table 2. Note that these are the observed reputations before errors have been accounted for and for Scoring we do not need the reputation of the recipient as it is irrelevant. Let $\mathcal{C}, \mathcal{E}_i \subset \mathcal{P}(\Theta)$, where $\mathcal{P}(\Theta)$ is the power set of Θ. $\mathcal{C}$ contains all single element sets $\{\theta_i\}$ where θ_i is consistent under the social norm; these are the possible states that observers believe can occur. For example, under Simple Standing, a bad donor giving to a good recipient where there is an error in the reputation of the donor is consistent with the social norm, while a bad donor giving to a good recipient with no errors is not. $\mathcal{E}_i$ contains a single element, namely the set of the most parsimonious θ_i's that are consistent with the observation, i.e. those that have the fewest changes from states within $\mathcal{C}$. This modeling choice comes from assuming a computation cost/opportunity cost of reasoning about where an error may be. Thus, observers will only consider errors if they must; i.e. if the observation does not make sense given their beliefs. Further, they will only consider alternative explanations that have the fewest number of errors.

Let $m : \mathcal{P}(\Theta) \to [0, 1]$ be a belief mass function. $m(U)$ for $U \in \mathcal{P}(\Theta)$ represents the degree to which the evidence supports a state within U, i.e. its belief mass. Note however that $m(U)$ has no bearing on any subsets of U, which would have their own belief masses. Thus, it is possible that $m(U) > 0$ and yet $m(V) = 0$ for all $V \subset U$. We assign different belief mass functions m and m' for our two different sources of information represented by the sets $\mathcal{C}$ and $\mathcal{E}$, respectively. The sum of all of the masses for an information set equals one, and the mass of the empty set is zero. m supplies belief masses as a function of g, ϵ, e_2, and the parameter ψ to every $\{\theta_i\}$ that is consistent with the social norm (i.e. every element of $\mathcal{C}$) as described in the main text, and a belief mass of zero to all other elements of $\mathcal{P}(\Theta)$. m' supplies a belief mass of 1 to the single element of $\mathcal{E}_i$ and zero to everything else in $\mathcal{P}(\Theta)$. To combine these belief mass functions, we apply Dempster's rule of combination via the fusion operator $\oplus$, which is used to combine information from different sources with different belief mass functions as follows:

$$m \oplus m'(U) = \frac{\sum_{V \cap W = U \neq \emptyset} m(V) m'(W)}{1 - \sum_{V \cap W = \emptyset} m(V) m'(W)}, \tag{16}$$

where U is the set whose mass we are computing via the fusion of both belief mass functions and $W, V \in \mathcal{P}(\Theta)$. The probability that the donor is good given the social norm and the observation $\mathcal{O}_i$ is thus:

$$P(G|\mathcal{O}_i) = \sum_{A \subset \mathcal{E}_i(G)} m \oplus m'(A) \tag{17}$$

where $\mathcal{E}_i(G) \subset \mathcal{E}_i$ is the set of all states within $\mathcal{E}_i$ in which the explanation concludes that the donor is good.

For an example, consider the case where we observe a good donor not giving to a good recipient (observation $\mathcal{O}_6$ from Table 2) under the Simple Standing norm. The simplest explanations consistent with the norm given the observation is that there was an error in the action, an error in the reputation of the recipient, or an error in the reputation of the donor. These possibilities all have only one error each. The set of the most parsimonious explanations given this observation is thus:

$$\mathcal{E}_6 = \{\{(\mathcal{O}_6, \text{act.}), (\mathcal{O}_6, \text{rep. of donor}), (\mathcal{O}_6, \text{rep. of recipient})\}\}. \tag{18}$$

It is composed of only one element that is a set of these three possibilities, and we have the mass $m'(\tilde{\mathcal{E}}_6) = 1$ where $\tilde{\mathcal{E}}_6$ is the single non-empty subset of $\mathcal{E}_6$. Given this, the donor is good if there was an error in the action or reputation of the recipient. Then, the probability that the donor is good given the observation is:

$$\begin{aligned}
&P(G|\mathcal{O}_6) = m \oplus m'(\{(\mathcal{O}_6, \text{act.})\}) + m \oplus m'(\{(\mathcal{O}_6, \text{rep. of recipient})\}) \\
&= \frac{m(\{(\mathcal{O}_6, \text{act.})\}) m'(\tilde{\mathcal{E}}_6)}{1 - \left(1 - m(\{(\mathcal{O}_6, \text{act.})\}) - m(\{(\mathcal{O}_6, \text{rep. of donor})\}) - m(\{(\mathcal{O}_6, \text{rep. of recipient})\})\right) m'(\tilde{\mathcal{E}}_6)} \\
&+ \frac{m(\{(\mathcal{O}_6, \text{rep. of recipient})\}) m'(\tilde{\mathcal{E}}_6)}{1 - \left(1 - m(\{(\mathcal{O}_6, \text{act.})\}) - m(\{(\mathcal{O}_6, \text{rep. of donor})\}) - m(\{(\mathcal{O}_6, \text{rep. of recipient})\})\right) m'(\tilde{\mathcal{E}}_6)} \\
&= \frac{g\psi(1-\epsilon)g\psi}{g\psi(1-\epsilon)g\psi + (1-g)(1-\psi)(1-e)g\psi + g\psi(1-e)(1-g)(1-\psi)} \\
&+ \frac{g\psi(1-e)(1-g)(1-\psi)}{g\psi(1-\epsilon)g\psi + (1-g)(1-\psi)(1-e)g\psi + g\psi(1-e)(1-g)(1-\psi)} \\
&= \frac{g\psi(1-\epsilon)g\psi + g\psi(1-e)(1-g)(1-\psi)}{g\psi(1-\epsilon)g\psi + (1-g)(1-\psi)(1-e)g\psi + g\psi(1-e)(1-g)(1-\psi)}.
\end{aligned} \tag{19}$$

References

Akçay, Çağlar, Reed, Veronica A, Campbell, S Elizabeth, Templeton, Christopher N, Beecher, Michael D, 2010. Indirect reciprocity: song sparrows distrust aggressive neighbours based on eavesdropping. Anim. Behav. 80 (6), 1041–1047.

Apicella, Coren L, Marlowe, Frank W, Fowler, James H, Christakis, Nicholas A, 2012. Social networks and cooperation in hunter-gatherers. Nature 481 (7382), 497–501.

Apicella, Coren L., Silk, Joan B., 2019. The evolution of human cooperation. Curr. Biol. 29 (11), R447–R450.

Dempster, A.P., 1967. Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Stat. 38 (2), 325–339. http://dx.doi.org/10.1214/aoms/1177698950.

Dores Cruz, Terence D, Thielmann, Isabel, Columbus, Simon, Molho, Catherine, Wu, Junhui, Righetti, Francesca, De Vries, Reinout E, Koutsoumpis, Antonis, Van Lange, Paul AM, Beersma, Bianca, et al., 2021. Gossip and reputation in everyday life. Phil. Trans. R. Soc. B 376 (1838), 20200301.

Douven, Igor, 2021. Abduction. In: Zalta, Edward N. (Ed.), The Stanford Encyclopedia of Philosophy, Summer 2021 ed. Metaphysics Research Lab, Stanford University.

Dugatkin, Lee Alan, 1997. Cooperation Among Animals. Oxford University Press.

de Faria, Pedro, Lima, Francisco, Santos, Rui, 2010. Cooperation in innovation activities: The importance of partners. Res. Policy 39 (8), 1082–1092.

Hauert, Christoph, de Monte, Silvia, Hofbauer, Josef, Sigmund, Karl, 2002. Volunteering as red queen mechanism for cooperation in public goods games. Science 296 (5570), 1129–1132.

Hilbe, Christian, Schmid, Laura, Tkadlec, Josef, Chatterjee, Krishnendu, Nowak, Martin A, 2018. Indirect reciprocity with private, noisy, and incomplete information. Proc. Natl. Acad. Sci. 115 (48), 12241–12246.

Jones, Martin, Sugden, Robert, 2001. Positive confirmation bias in the acquisition of information. Theory and Decision 50 (1), 59–99.

Kato-Shimizu, Mayuko, Onishi, Kenji, Kanazawa, Tadahiro, Hinobayashi, Toshihiko, 2013. Preschool children's behavioral tendency toward social indirect reciprocity. PLoS One 8 (8), e70915.

Kessinger, Taylor A., Tarnita, Corina E., Plotkin, Joshua B., 2023. Evolution of norms for judging social behavior. Proc. Natl. Acad. Sci. 120 (24), e2219480120.

Michod, Richard E., Herron, Matthew D., 2006. Cooperation and conflict during evolutionary transitions in individuality. J. Evol. Biol. 19 (5), 1406–1409.
Morsky, Bryce, Plotkin, Joshua B., Akçay, Erol, 2023. Indirect reciprocity with Bayesian reasoning and biases. URL https://osf.io/preprints/socarxiv/ustqg.
Ohtsuki, Hisashi, Iwasa, Yoh, 2004. How should we define goodness?—reputation dynamics in indirect reciprocity. J. Theoret. Biol. 231 (1), 107–120.
Okada, Isamu, 2020. A review of theoretical studies on indirect reciprocity. Games 11 (3), 27.
Okada, Isamu, Sasaki, Tatsuya, Nakai, Yutaka, 2018. A solution for private assessment in indirect reciprocity using solitary observation. J. Theoret. Biol. 455, 7–15.
Paul, Gabriele, 1993. Approaches to abductive reasoning: an overview. Artif. Intell. Rev. 7 (2), 109–152.
Peña, Nieves Arranz, de Arroyabe, Juan Carlos Fernández, 2002. Business Cooperation. Palgrave Macmillian UK.
Peters, Uwe, 2020. What is the function of confirmation bias? Erkenntnis 87 (3), 1351–1376. http://dx.doi.org/10.1007/s10670-020-00252-1.
Radzvilavicius, Arunas L, Kessinger, Taylor A, Plotkin, Joshua B, 2021. Adherence to public institutions that foster cooperation. Nat. Commun. 12 (1), 1–14.
Radzvilavicius, Arunas L, Stewart, Alexander J, Plotkin, Joshua B, 2019. Evolution of empathetic moral evaluation. eLife 8, e44269.
Rand, David G., Ohtsuki, Hisashi, Nowak, Martin A., 2009. Direct reciprocity with costly punishment: Generous tit-for-tat prevails. J. Theoret. Biol. 256 (1), 45–57.
Sasaki, Tatsuya, Okada, Isamu, Nakai, Yutaka, 2017. The evolution of conditional moral assessment in indirect reciprocity. Sci. Rep. 7 (1), 1–8.
Schmid, Laura, Shati, Pouya, Hilbe, Christian, Chatterjee, Krishnendu, 2021. The evolution of indirect reciprocity under action and assessment generosity. Sci. Rep. 11.
Shafer, Glenn, 1976a. A Mathematical Theory of Evidence, Vol. 42. Princeton University Press.
Shafer, Glenn, 1976b. A Mathematical Theory of Evidence. Princeton University Press.
Sommerfeld, Ralf D, Krambeck, Hans-Jürgen, Semmann, Dirk, Milinski, Manfred, 2007. Gossip as an alternative for direct observation in games of indirect reciprocity. Proc. Natl. Acad. Sci. 104 (44), 17435–17440.
Stanca, Luca, 2009. Measuring indirect reciprocity: Whose back do we scratch? J. Econ. Psychol. 30 (2), 190–202.
Thagard, Paul, Shelley, Cameron, 1997. Abductive reasoning: Logic, visual thinking, and coherence. In: Logic and Scientific Methods. Springer, pp. 413–427.
Yager, Ronald R., Liu, Liping, 2008. Classic Works of the Dempster-Shafer Theory of Belief Functions, Vol. 219. Springer.
Yoeli, Erez, Hoffman, Moshe, Rand, David G., Nowak, Martin A., 2013. Powering up with indirect reciprocity in a large-scale field experiment. Proc. Natl. Acad. Sci. 110 (Supplement 2), 10424–10429.