Journal of Theoretical Biology 595 (2024) 111947

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Convergence of reputations under indirect reciprocity

Bryce Morsky [a,b,*], Joshua B. Plotkin [b], Erol Akçay [b]

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

ARTICLE INFO

Keywords:
Cooperation
Evolutionary game theory
Indirect reciprocity
Reputations
Social norms

ABSTRACT

Previous research has shown how indirect reciprocity can promote cooperation through evolutionary game theoretic models. Most work in this field assumes a separation of time-scales: individuals' reputations equilibrate at a fast time scale for given frequencies of strategies while the strategies change slowly according to the replicator dynamics. Much of the previous research has focused on the behaviour and stability of equilibria for the replicator dynamics. Here we focus on the underlying reputational dynamics that occur on a fast time scale. We describe reputational dynamics as systems of differential equations and conduct stability analyses on their equilibria. We prove that reputations converge to a unique equilibrium under a solitary observer model for each of the five standard norms and whether assessments are public or private. These results confirm a crucial but previously understudied assumption underlying the theory of indirect reciprocity for the most studied set of norms.

1. Introduction

Indirect reciprocity is an important mechanism to foster cooperation. Theoretical studies have extended an evolutionary game theoretic model of the Prisoner's Dilemma game (also termed the Donation game) by adding a system of reputations and a population of Discriminators who defect against "bad" individuals and cooperate with "good" ones (Fishman, 2003; Leimar and Hammerstein, 2001; Nowak and Sigmund, 2005; Ohtsuki and Iwasa, 2006; Okada, 2020; Okada et al., 2018; Sasaki et al., 2017). To determine who is good and who is bad, interactions between pairs of individuals are observed. To assess a donor's reputation (as good or bad), an observer may consider the action the donor took (to cooperate or defect), the reputation of the recipient with the observer (either good or bad), and a social norm. The social norms provide the rules of how to assess what the observer observed. For all norms, a donor is assessed as good if they cooperate with a good donor. This is a minimum requirement to promote cooperation. However, the norms will differ on their recommendations for other scenarios. There are five social norms that are frequently studied: Scoring (Wedekind and Milinski, 2000), Shunning (Takahashi and Mashima, 2006), Staying (Nakai and Muto, 2008), Simple Standing (Milinski et al., 2001), and Stern Judging (Santos et al., 2016). Scoring was the first norm studied, and under it a donor is considered good if they cooperate and bad if they defect. Assessments under Scoring do not depend on the reputation of the recipient, but this norm leads to a population that only ever defects. Thus, attention shifted to higher order norms that factor in the reputation of the recipient (Shunning, Staying, Simple Standing, and Stern Judging). Under the Stern Judging norm, donors are assessed as good if they cooperate with good recipients and defect with bad ones. Conversely, they are assessed as bad if they defect with good recipients and cooperate with bad ones. In the initial work studying these norms, reputations were generally assumed to be assessed publicly wherein there is a shared reputational system (Brandt and Sigmund, 2004; Chalub et al., 2006; Ohtsuki and Iwasa, 2004) and all individuals agree on each other's reputation. Private reputations were later explored (Okada et al., 2018), which allow for individuals to hold private information and disagree about the reputations of others. Conflict between different opinions of individuals can undermine the reputational system and thus cooperation. These five norms and two methods of assessment (public and private) form the core set of theoretical models of indirect reciprocity. Additionally, there are a great many models extending this framework including noisy and incomplete information (Hilbe et al., 2018), individuals' emotions (Radzvilavicius et al., 2019) and individuals' reasoning to account for errors and discrepancies in reputations (Morsky et al., 2024, Pandula et al., 2024) to name a few.

Despite a wide range of assumptions about assessment and game play, the above models generally share the assumption that the dynamics of reputations and strategies operate at two different time scales. Reputations are assumed to equilibrate rapidly while strategies change much more slowly. This assumption can be justified if each individual undergoes many interactions during its lifetime (if replicator dynamics

* Corresponding author at: Department of Mathematics, Florida State University, Tallahassee, FL, USA.
E-mail addresses: bmorsky@fsu.edu (B. Morsky), jplotkin@sas.upenn.edu (J.B. Plotkin), eakcay@sas.upenn.edu (E. Akçay).

https://doi.org/10.1016/j.jtbi.2024.111947
Received 15 April 2024; Received in revised form 7 August 2024; Accepted 9 September 2024
Available online 18 September 2024

model a birth–death process, or simply update infrequently). A further justification is that individuals cannot fully assess the payoffs of their strategies (and thus not compare and imitate) until reputations have reached an equilibrium. As a theory building strategy, this assumption allows the models to account for the full incentive effects through reputations of a strategy composition of the population. It also makes the theory analytically tractable.

Under this separation of time-scale assumption, the strategies of individuals change in response to payoffs that are computed for the frequencies of the types of individuals while assuming that their reputations are at the equilibrium levels given the strategy frequencies. Though equilibria of the reputational system were found previously and used to analyse the replicator dynamics that govern the change in strategies, whether or not they converged to these equilibria has been understudied. Convergence of reputations to unique equilibria has only been proved in a few cases such as for homogeneous populations (Ohtsuki and Iwasa, 2006) and for models that incorporate reasoning (Morsky et al., 2024; Pandula et al., 2024). However, convergence in some situations is conditional on the parameter values chosen (Pandula et al., 2024), and thus need not hold generally. Thus, it is an open and key question as to whether or not reputations converge. Here, we have proven that reputations in the standard indirect reciprocity model do indeed converge to unique equilibria for the five common norms and both public and private assessments of reputations under a solitary observation model. We do this by representing the reputational dynamics as a system of ordinary differential equations and analyse the stability of the equilibria of these systems. This setup places some assumptions on the reputation dynamics such as the population being infinite and time being continuous. However, these are frequently assumed in mathematical models of indirect reciprocity. So long as the population is large and dynamics occur in short intervals of time, this is a reasonable assumption. In the methods section, we define the system of ordinary differential equations that model the reputational dynamics, and in the results section analyse the stability of their equilibria.

2. Methods

Consider three types of players each playing a specific strategy in the donation game. AllC (always cooperate) players are those who always *intend* to cooperate regardless of the reputation of the recipient, and x is their frequency in the population. Note that AllC players only intend to cooperate: they do not always successfully do so. As discussed below, we assume — as is standard in the literature — that errors may occur when players attempt to cooperate by which they unintentionally defect. AllD players always defect, and their frequency in the population is y. Note that there is no possibility for an AllD player to accidentally cooperate. The third and final type of player are Discriminators, who intend to cooperate with good recipients and defect against bad ones. They thus act as punishers of "bad behaviour" (as determined by the social norm). Their frequency in the population is z, and we have $x + y + z = 1$. Since reputations converge rapidly, i.e. before strategies can change, x, y, and z will be constants in our analyses.

Errors in action and assessment are assumed in many models of indirect reciprocity. Let $\frac{1}{2} > e_1 > 0$ be the probability that a donor who intends to cooperate defects by mistake. Further, let $\frac{1}{2} > e_2 > 0$ be the probability that there is an error in the assessment of the reputation of the donor. That is to say, with probability e_2, the observer assigns the *opposite* reputation to the donor than they intended to. Define $\epsilon = (1 - e_1)(1 - e_2) + e_1 e_2$ as the probability that an individual who intends to cooperate is observed doing so. We will use ϵ throughout our analysis rather than e_1. Also, we write $e = e_2$ to further simplify our notation. The parameters ϵ and e along with the social norm and whether or not assessments are public will thus determine reputations.

Table 1
Assessments of the donor (either G or B for good or bad) given the donor's action (either C or D for cooperate or defect), recipient's reputation (G or B), and the social norm. The dash under Staying implies that the reputation of the donor is not updated when they interact with a recipient with a bad reputation.

Social norm	Donor's action/recipient's reputation			
	C/G	D/G	C/B	D/B
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

Social norms determine what actions are good and what bad given the reputation of the donor. Five important norms frequently studied in the literature are: Scoring, Shunning, Staying, Simple Standing, and Stern Judging. The rules for these norms are represented in Table 1. For example, under Simple Standing, a donor is assessed as good if they cooperate with a good recipient, and bad if they do not. And, they are assessed as good when they interact with a bad recipient, regardless of whether or not they cooperate. Note that due to the error in assessment, it is possible for an observer to assess a donor who interacts with a bad recipient as bad. However, in models where players can factor in error rates and thereby reason about the intention of the donor, this cannot happen (Morsky et al., 2024; Pandula et al., 2024).

The state variables in our analyses of the reputational dynamics are the fraction of players of each type that are good. Thus, g_x and $1-g_x$ are the frequencies of AllC players with good and bad reputations, respectively. Likewise g_y and $1 - g_y$ are the frequencies of AllD players with good and bad reputations, respectively. Finally, g_z and $1 - g_z$ are the frequencies of Discriminators with good and bad reputations, respectively. The total frequency of good players in the population is $g = xg_x + yg_y + zg_z$.

Reputations can be assessed publicly or privately. Under public assessments of reputations, each interaction is observed by a single observer and all players follow their assessment of the donor's reputation. For private assessments of reputations we use the solitary observation model proposed by Okada et al. (2018). Under this model, each interaction is observed by only one observer and thus reputations are statistically independent of one another, which simplifies the analysis. Private assessment of reputations particularly impacts the assessment of the reputations of Discriminators. Because, Discriminators' behaviours are determined by the reputation the recipient has with them. Thus, if this differs between the observer and a donor Discriminator, we need to know the frequency with which two players agree that a recipient is good, denoted g_2. The probability that two AllC players are good is defined as g_{x2}. g_{y2} and g_{z2} for AllD players and Discriminators are defined similarly. Thus, $g_2 = xg_{x2}+yg_{y2}+zg_{z2}$. Under private assessment of reputations, g_{x2}, g_{y2}, and g_{z2} will be state variables in addition to g_x, g_y, and g_z.

As in previous models, we assume that reputations change by a good individual being reassessed as bad or a bad individual being reassessed as good (Okada et al., 2018; Sasaki et al., 2017). Let g_i^+ be the probability that a bad individual of type i is reassessed as good, and g_i^- be the probability that a good individual of type i is reassessed as bad. To represent this process as a system of differential equations, we define $\dot{g}_i = g_i^+ - g_i^-$ by assuming a limiting process. We are thus able to convert a discrete process into a continuous one. We use the same values of g_i^+ and g_i^- from Okada et al. (2018) and Sasaki et al. (2017). The reputational systems under public assessment for the five norms are thus:

$$\left.\begin{aligned} \dot{g}_x &= \epsilon - g_x \\ \dot{g}_y &= e - g_y \\ \dot{g}_z &= \epsilon g + e(1-g) - g_z \end{aligned}\right\} \text{Scoring,} \tag{1a}$$

$$\left.\begin{aligned}\dot{g}_x &= \epsilon g + e(1-g) - g_x\\ \dot{g}_y &= e - g_y\\ \dot{g}_z &= \epsilon g + e(1-g) - g_z\end{aligned}\right\}\text{Shunning,} \tag{1b}$$

$$\left.\begin{aligned}\dot{g}_x &= (\epsilon - g_x)g\\ \dot{g}_y &= (e - g_y)g\\ \dot{g}_z &= (\epsilon - g_z)g\end{aligned}\right\}\text{Staying,} \tag{1c}$$

$$\left.\begin{aligned}\dot{g}_x &= \epsilon g + (1-e)(1-g) - g_x\\ \dot{g}_y &= eg + (1-e)(1-g) - g_y\\ \dot{g}_z &= \epsilon g + (1-e)(1-g) - g_z\end{aligned}\right\}\text{Simple Standing,} \tag{1d}$$

$$\left.\begin{aligned}\dot{g}_x &= \epsilon g + (1-\epsilon)(1-g) - g_x\\ \dot{g}_y &= eg + (1-e)(1-g) - g_y\\ \dot{g}_z &= \epsilon g + (1-e)(1-g) - g_z\end{aligned}\right\}\text{Stern Judging.} \tag{1e}$$

And for private assessment of reputations, the systems are as follows:

$$\left.\begin{aligned}\dot{g}_x &= \epsilon g + e(1-g) - g_x\\ \dot{g}_y &= e - g_y\\ \dot{g}_z &= \epsilon g_2 + e(1-g_2) - g_z\\ \dot{g}_{x2} &= (\epsilon g + e(1-g))g_x - g_{x2}\\ \dot{g}_{y2} &= eg_y - g_{y2}\\ \dot{g}_{z2} &= (\epsilon g_2 + e(1-g_2))g_z - g_{z2}\end{aligned}\right\}\text{Shunning,} \tag{2a}$$

$$\left.\begin{aligned}\dot{g}_x &= (\epsilon - g_x)g\\ \dot{g}_y &= (e - g_y)g\\ \dot{g}_z &= \epsilon g_2 + e(g-g_2) - g_z g\\ \dot{g}_{x2} &= (\epsilon g_x - g_{x2})g\\ \dot{g}_{y2} &= (eg_y - g_{y2})g\\ \dot{g}_{z2} &= (\epsilon g_2 + e(g-g_2))g_z - g_{z2}g\end{aligned}\right\}\text{Staying,} \tag{2b}$$

$$\left.\begin{aligned}\dot{g}_x &= \epsilon g + (1-e)(1-g) - g_x\\ \dot{g}_y &= eg + (1-e)(1-g) - g_y\\ \dot{g}_z &= \epsilon g_2 + e(g-g_2) + (1-e)(1-g) - g_z\\ \dot{g}_{x2} &= (\epsilon g + (1-e)(1-g))g_x - g_{x2}\\ \dot{g}_{y2} &= (eg + (1-e)(1-g))g_y - g_{y2}\\ \dot{g}_{z2} &= (\epsilon g_2 + e(g-g_2) + (1-e)(1-g))g_z - g_{z2}\end{aligned}\right\}\text{Simple Standing,} \tag{2c}$$

$$\left.\begin{aligned}\dot{g}_x &= \epsilon g + (1-\epsilon)(1-g) - g_x\\ \dot{g}_y &= eg + (1-e)(1-g) - g_y\\ \dot{g}_z &= (1-2g)(1-e) + g + (\epsilon - e)(2g_2 - g) - g_z\\ \dot{g}_{x2} &= (\epsilon g + (1-\epsilon)(1-g))g_x - g_{x2}\\ \dot{g}_{y2} &= (eg + (1-e)(1-g))g_y - g_{y2}\\ \dot{g}_{z2} &= ((1-2g)(1-e) + g + (\epsilon - e)(2g_2 - g))g_z - g_{z2}\end{aligned}\right\}\text{Stern Judging.}$$

(2d)

Since the reputation of the recipient is irrelevant for Scoring, the reputation dynamics are the same whether assessments are public or private.

To be more explicit on how these equations are derived, consider the dynamics for AllC under Simple Standing. g_x increases when a bad AllC player is reassessed as good, which occurs with probability $g_x^+ = (1-g_x)(\epsilon g + (1-e)(1-g))$. A bad AllC player is selected with probability $1-g_x$. With probabilities g and $1-g$ they pair with a good and bad recipient, respectively. When paired with a good recipient, the bad AllC player is reassessed as good with probability ϵ. When paired with a good AllC player, they are reassessed as good with probability $1-e$. In a similarly fashion a good AllC player is reassessed as bad with probability $g_x^- = g_x((1-\epsilon)g + e(1-g))$. Simplifying we have $g_x^+ - g_x^- = \epsilon g + (1-e)g - g_x$, to which we define $\dot{g}_x$. The reputational dynamics for g_y can be found similarly. Under public assessment of reputations, g_z is also found similarly. However, under private assessment, one has to take care of how the Discriminator donor and observer view the reputations of the recipient. If they both agree that they are good, then the donor will intend to cooperate and the observer will evaluate them as if they are interacting with a good recipient. This occurs with probability g_2 and the donor will then be assessed as good with probability ϵ. With probability $g - g_2$ the donor believes that the recipient is bad and so defects, but the observer believes that they are good. A donor who intends to defect against a good recipient will be assessed as good only if an error in assessment occurs, i.e. with probability e. Finally, with probability $1-g$, the observer believes that the recipient is bad, and thus will assess the donor as good so long as there is no error in assessment, i.e. with probability $1-e$. Thus, a bad Discriminator will be reassessed as good with probability $g_z^+ = (1-g_z)(\epsilon g_2 + e(g-g_2) + (1-e)(1-g))$. g_z^- is found in a similar way. However, we consider the probabilities that the donor is bad. For example, if both donor and observer believe that the recipient is good, which happens with probability g_2, then the donor is assessed as bad with probability $1-\epsilon$. The other terms are found in a similar way giving us $g_z^- = g_z((1-\epsilon)g_2 + (1-e)(g-g_2) + e(1-g))$ and $\dot{g}_z = g_z^+ - g_z^- = \epsilon g_2 + e(g_2 - g) + (1-e)(1-g)$.

Continuing with the example of Simple Standing under private assessment of reputations, $g_x - g_{x2}$ is the probability that one player believes that an AllC player is good and another believes that they are bad. Thus, g_{x2} increases when such an AllC player is reassessed as good, which happens with probability $g_{x2}^+ = (g_x - g_{x2})(\epsilon g + (1-e)(1-g))$. In a similar way, we can calculate the probability that g_{x2} decreases as $g_{x2}^- = g_{x2}((1-\epsilon)g + e(1-g))$, which gives us $\dot{g}_{x2} = g_{x2}^+ - g_{x2}^- = (\epsilon g + (1-e)(1-g))g_x - g_{x2}$. g_{y2}^+, g_{y2}^-, $\dot{g}_{y2}$, g_{z2}^+, g_{z2}^-, and $\dot{g}_{z2}$ are all computed similarly.

Here, we have assumed the solitary observation model of Okada et al. (2018), which in the long run will lead reputations to be independent. Indeed, at equilibrium we generally have $g_{i2} = g_i^2$. This suggests that we could analyse reduced systems for private assessment in three dimensions rather than six as we do in our analysis. However, this reduced system would not cover cases where reputations are initialized non-independently. For example, the system could be initialized where a given focal individual has the same reputation in everyone's eye's, but that reputations differ from focal individual to focal individual. Such a scenario can happen, e.g., in a population where a public reputation institution breaks down and individuals have to rely on private reputations. In such a case, g_z could be $\frac{1}{2}$, for example, but g_{z2} would also be $\frac{1}{2}$, because everyone initially agrees about the reputations. To prove convergence in such cases, the full six dimensional model is still required. Therefore, we show below that even in such a scenario, the system will return to independent reputations. These proofs also hold for the scenario where reputations are always strictly independent.

3. Results

3.1. Public assessment

First consider an analysis of the case of public assessment of reputations. Equilibria have previously been found in the literature (Sasaki et al., 2017), but here we show that reputations converge to a unique equilibrium for each norm. Beginning with Scoring, there is only one equilibrium frequency of good individuals, namely

$$g^* = \frac{\epsilon x + e(1-x)}{1-(\epsilon - e)z}, \tag{3}$$

which is defined across the simplex and for all error rates (as will all equilibria here). The Jacobian of the system of ODEs is

$$\mathbf{J} = \begin{pmatrix} -1 & 0 & 0\\ 0 & -1 & 0\\ (\epsilon - e)x & (\epsilon - e)y & (\epsilon - e)z - 1\end{pmatrix}, \tag{4}$$

which has eigenvalues $\lambda_1 = \lambda_2 = -1$ and $\lambda_3 = (\epsilon - e)z - 1 < 0$. Note that $\mathbf{J}$ is not a function of g, which will be the case under public assessment for all norms but Staying. Since the eigenvalues are negative and g^* is the sole equilibrium, g^* is stable. Additionally, since assessments under Scoring do not depend upon the reputation of the recipient, there is no difference in Scoring between public and private assessments and thus these results hold for both.

For Shunning, the sole equilibrium frequency of good individuals is

$$g^* = \frac{e}{1 - (\epsilon - e)(1 - y)}, \tag{5}$$

and the Jacobian of the system is

$$\mathbf{J} = \begin{pmatrix} (\epsilon - e)x - 1 & (\epsilon - e)y & (\epsilon - e)z \\ 0 & -1 & 0 \\ (\epsilon - e)x & (\epsilon - e)y & (\epsilon - e)z - 1 \end{pmatrix}. \tag{6}$$

It has eigenvalues $\lambda_1 = \lambda_2 = -1$ and $\lambda_3 = (\epsilon - e)(1 - y) - 1 < 0$, and thus g^* is stable.

For Staying, there are two equilibria: $g^* = 0$ and $g^* = \epsilon(1 - y) + ey$, and in the latter case $g_x^* = g_z^* = \epsilon$ and $g_y^* = e$. Subbing these solutions into the Jacobian gives us the following matrices:

$$\mathbf{J}(0) = \begin{pmatrix} \epsilon x & \epsilon y & \epsilon z \\ ex & ey & ez \\ \epsilon x & \epsilon y & \epsilon z \end{pmatrix}, \tag{7a}$$

$$\mathbf{J}(\epsilon(1-y)+ey) = \begin{pmatrix} -\epsilon(1 - y) - ey & 0 & 0 \\ 0 & -\epsilon(1 - y) - ey & 0 \\ 0 & 0 & -\epsilon(1 - y) - ey \end{pmatrix}. \tag{7b}$$

The eigenvalues for Eq. (7a) are $\lambda_1 = \lambda_2 = 0$ and $\lambda_3 = \epsilon(1 - y) + ey > 0$, and thus $g^* = 0$ is unstable. The eigenvalues for Eq. (7b) are $\lambda_1 = \lambda_2 = \lambda_3 = -(1 - y)\epsilon - ey < 0$, and thus $g^* = \epsilon(1 - y) + ey$ is stable.

For Simple Standing, we have the sole equilibrium

$$g^* = \frac{1 - e}{1 - e + 1 - \epsilon(1 - y) - ey}. \tag{8}$$

The Jacobian of the system is

$$\mathbf{J} = \begin{pmatrix} (\epsilon + e - 1)x - 1 & (\epsilon + e - 1)y & (\epsilon + e - 1)z \\ (2e - 1)x & (2e - 1)y - 1 & (2e - 1)z \\ (\epsilon + e - 1)x & (\epsilon + e - 1)y & (\epsilon + e - 1)z - 1 \end{pmatrix}, \tag{9}$$

which has eigenvalues $\lambda_1 = \lambda_2 = -1$ and $\lambda_3 = e - 1 + \epsilon(1 - y) + ey - 1 < 0$. Therefore, g^* is stable.

And, finally, for Stern Judging, we have the sole equilibrium

$$g^* = \frac{1 - \epsilon x - e(1 - x)}{1 - \epsilon x - e(1 - x) + 1 - \epsilon(1 - y) - ey}. \tag{10}$$

The Jacobian of the system is

$$\mathbf{J} = \begin{pmatrix} (2\epsilon - 1)x - 1 & (2\epsilon - 1)y & (2\epsilon - 1)z \\ (2e - 1)x & (2e - 1)y - 1 & (2e - 1)z \\ (\epsilon + e - 1)x & (\epsilon + e - 1)y & (\epsilon + e - 1)z - 1 \end{pmatrix}, \tag{11}$$

which has eigenvalues $\lambda_1 = \lambda_2 = -1$ and $\lambda_3 = \epsilon x + e(1 - x) - 1 + \epsilon(1 - y) + ey - 1 < 0$. Therefore, g^* is stable.

We can also show convergence by noting that we may find a closed form equation for $\dot{g}$ and analyse it. For Scoring, for example, $\dot{g} = \epsilon x + ey + (\epsilon g + e(1 - g))z - g$. Since there is a single interior equilibrium point g^* and since $\dot{g} > 0$ for $g = 0$ and $\dot{g} < 0$ for $g = 1$, g^* is stable. Similar arguments hold for the other norms.

3.2. Private assessment

For Shunning, the sole equilibrium frequency of good individuals is

$$g^* = \frac{\epsilon g_2^* z + e(1 - g_2^* z)}{1 - (\epsilon - e)x}. \tag{12}$$

The Jacobian of the system evaluated at this equilibrium is

$$\mathbf{J}(g^*) = \begin{pmatrix} (\epsilon - e)x - 1 & (\epsilon - e)y & (\epsilon - e)z & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & (\epsilon - e)x & (\epsilon - e)y & (\epsilon - e)z \\ ((\epsilon - e)x + 1)g_x^* & (\epsilon - e)g_x^* y & (\epsilon - e)g_x^* z & -1 & 0 & 0 \\ 0 & e & 0 & 0 & -1 & 0 \\ 0 & 0 & g_z^* & (\epsilon - e)g_z^* x & (\epsilon - e)g_z^* y & (\epsilon - e)g_z^* z - 1 \end{pmatrix}. \tag{13}$$

The characteristic equation is $(1 + \lambda)^3(\lambda^3 + c_2\lambda^2 + c_1\lambda + c_0) = 0$ with the following coefficients:

$$c_2 = 3 - (\epsilon - e)(x + g_z^* z) > 2, \tag{14a}$$

$$c_1 = 3 - (\epsilon - e)(3g_z^* z + (2 + (\epsilon - e)(g_x^* - g_z^*)z)x) > 0, \tag{14b}$$

$$\begin{aligned} c_0 &= 1 - (\epsilon - e)(x + 2g_z^* z + 2(\epsilon - e)(g_x^* - g_z^*)xz) \\ &= 1 - (\epsilon - e)(2ez + x(1 + 2g^* z(\epsilon - e)^2)) - 2z(\epsilon - e)^2(1 - (\epsilon - e)x)g_2^* \\ &\geq 1 - (\epsilon - e)(2ez + x(1 + 2g^* z(\epsilon - e)^2)) - 2z(\epsilon - e)^2(1 - (\epsilon - e)x)g^* \\ &= 1 - (\epsilon - e)x - 2(\epsilon - e)((\epsilon - e)g^* + e)z \\ &\geq 1 - (\epsilon - e)x - \frac{2e(\epsilon - e)}{1 - \epsilon + e}z \\ &> 1 - (\epsilon - e)x - z \geq 0. \end{aligned} \tag{14c}$$

The inequality for c_0 follows for the following reasons: $g_x^* = (\epsilon - e)g^* + e$ and $g_z^* = (\epsilon - e)g_2^* + e$; $g^* \geq g_2^*$; $g_x^* = (\epsilon - e)g^* + e \leq (\epsilon - e)g_x^* + e \implies g_x^* \leq e/(1 - \epsilon + e)$; and $1 - \epsilon + e - 2e(\epsilon - e) = e_1 + 4(1 - e_1)e^2 > 0$. The first three eigenvalues are negative as can be seen from the first factor of the characteristic equation. The second factor is a cubic equation of λ. Note that all of the coefficients of this cubic are positive. The Routh–Hurwitz criterion for stability requires that all coefficients of this cubic to be positive and $c_2c_1 - c_0 > 0$. Checking this last condition gives us

$$c_2c_1 - c_0 > 2c_1 - c_0 = 5 - (\epsilon - e)(3x + 4g_z^* z) > 0. \tag{15}$$

Therefore, g^* is stable.

There are two equilibria for Staying under private assessment. The first of which is $g^* = 0$, and evaluating the Jacobian at this equilibrium gives us

$$\mathbf{J}(0) = \begin{pmatrix} x\epsilon & y\epsilon & z\epsilon & 0 & 0 & 0 \\ ex & ey & ez & 0 & 0 & 0 \\ ex & ey & ez & (\epsilon - e)x & (\epsilon - e)y & (\epsilon - e)z \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{16}$$

The eigenvalues are $\lambda_i = 0$ for $i = 1 \ldots, 5$ and $\lambda_6 = \epsilon x + e(1 - x) > 0$. Thus, $g^* = 0$ is unstable. At the other equilibrium, $g_x^* = \epsilon$, $g_y^* = e$, $g_z^* = (\epsilon - e)g_2^*/g^* + e$, and the Jacobian evaluated at it is

$$\mathbf{J}(g^*) = \begin{pmatrix} \mathbf{J}_1 & \mathbf{J}_2 \\ \mathbf{J}_3 & \mathbf{J}_4 \end{pmatrix}, \tag{17a}$$

$$\mathbf{J}_1 = \begin{pmatrix} -g^* & 0 & 0 \\ 0 & -g^* & 0 \\ (e - g_z^*)x & (e - g_z^*)y & (e - g_z^*)z - g^* \end{pmatrix}, \tag{17b}$$

$$\mathbf{J}_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ (\epsilon - e)x & (\epsilon - e)y & (\epsilon - e)z \end{pmatrix}, \tag{17c}$$

$$\mathbf{J}_3 = \begin{pmatrix} \epsilon g^* & 0 & 0 \\ 0 & eg^* & 0 \\ (e - g_z^*)g_z^* x & (e - g_z^*)g_z^* y & (e - g_z^*)g_z^* z + g_z^* g^* \end{pmatrix}, \tag{17d}$$

$$\mathbf{J}_4 = \begin{pmatrix} -g^* & 0 & 0 \\ 0 & -g^* & 0 \\ (\epsilon - e)g_z^* x & (\epsilon - e)g_z^* y & (\epsilon - e)g_z^* z - g^* \end{pmatrix}. \tag{17e}$$

The characteristic equation is $(\lambda + g^*)^4(\lambda^2 + c_1\lambda + c_0) = 0$. Since $g^* \geq g_y^* = e$,

$$c_1 = 2g^* - ez + (1 - \epsilon + e)g_z^* z \geq e(2 - z) + (1 - \epsilon + e)g_z^* z > 0, \tag{18a}$$

$$c_0 = g^*(g^* + ((1 - 2(\epsilon - e))g_z^* - e)z) \geq g^*(g_z^* z + ((1 - 2(\epsilon - e))g_z^* - e)z)$$
$$= g^*\sqrt{k_1 + k_2 + k_3} > 0, \tag{18b}$$

$$k_1 = k1 = e^2(y - z)^2 + 2\epsilon exy + \epsilon^2 x^2 > 0, \tag{18c}$$

$$k_2 = 2e^2 yz(2(1 - \epsilon^2) + 2e(2\epsilon - e)) > 0, \tag{18d}$$

$$k_3 = 2\epsilon exz + 4\epsilon(1 - \epsilon)(\epsilon - e)^2 xz > 0. \tag{18e}$$

We obtain the inequality for Eq. (18b) by solving $g^* = \epsilon x + ey + g_z^* z$ and $g_2^* = \epsilon^2 x + e^2 y + (g_z^*)^2 z = \epsilon^2 x + e^2 y + ((\epsilon - e)g_2^*/g^* + e)^2 z$ for g^* and g_2^* and then plugging these solutions into c_0. Note also that the radicand is positive. Thus, the last two eigenvalues must be negative and so g^* is stable.

Next consider Simple Standing. The equilibrium frequency of good individuals is

$$g^* = \frac{1 - e + (\epsilon - e)zg_2^*}{1 - 2e + 1 - (\epsilon - e)}, \tag{19}$$

and the Jacobian evaluated at it is

$$\mathbf{J}(g^*) = \begin{pmatrix} \mathbf{J}_1 & \mathbf{J}_2 \\ \mathbf{J}_3 & \mathbf{J}_4 \end{pmatrix}, \tag{20a}$$

$$\mathbf{J}_1 = \begin{pmatrix} (\epsilon + e - 1)x - 1 & (\epsilon + e - 1)y & (\epsilon + e - 1)z \\ (2e - 1)x & (2e - 1)y - 1 & (2e - 1)z \\ (2e - 1)x & (2e - 1)y & (2e - 1)z - 1 \end{pmatrix}, \tag{20b}$$

$$\mathbf{J}_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ (\epsilon - e)x & (\epsilon - e)y & (\epsilon - e)z \end{pmatrix}, \tag{20c}$$

$$\mathbf{J}_3 = \begin{pmatrix} (\epsilon + e - 1)g_x^* x + g_x^* & (\epsilon + e - 1)g_x^* y & (\epsilon + e - 1)g_x^* z \\ (2e - 1)g_y^* x & (2e - 1)g_y^* y + g_y^* & (2e - 1)g_y^* z \\ (2e - 1)g_z^* x & (2e - 1)g_z^* y & (2e - 1)g_z^* z + g_z^* \end{pmatrix}, \tag{20d}$$

$$\mathbf{J}_4 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ (\epsilon - e)g_z^* x & (\epsilon - e)g_z^* y & (\epsilon - e)g_z^* z - 1 \end{pmatrix}. \tag{20e}$$

The characteristic equation is $(\lambda + 1)^3(\lambda^3 + c_2\lambda^2 + c_1\lambda + c_0) = 0$ with positive coefficients:

$$c_2 = 2 + 1 - (\epsilon - e)g_z^* z + (1 - \epsilon - e)x + (1 - 2e)y + (1 - 2e)z > 2, \tag{21a}$$

$$c_1 = (1 + (1 - \epsilon - e)(2 + (\epsilon - e)(g_x^* - g_z^*)z))x + 2(1 - 2e)$$
$$\times (1 - (\epsilon - e)(g_z^* - g_y^*)z)y$$
$$+ 3(1 - (\epsilon - e)g_z^* z) + 2(1 - 2e)z > 0, \tag{21b}$$

$$c_0 = (1 + (1 - \epsilon - e)(1 + 2(\epsilon - e)(g_x^* - g_z^*)z))x + 2(e + (1 - 2e)$$
$$\times (1 - (\epsilon - e)(g_z^* - g_y^*)z))y$$
$$+ 2(1 - \epsilon g_z^* - e(1 - g_z^*))z > 0, \tag{21c}$$

since $1 - \epsilon - e = e_1(1 - 2e) > 0$. The first three eigenvalues are negative. Since all of the coefficients of the cubic are positive, we need only to confirm that $c_2c_1 - c_0 > 2c_1 - c_0 > 0$ to prove stability. Checking this last condition gives us

$$2c_1 - c_0 = 8 - 6e - 3x(\epsilon - e) - 4(\epsilon - e)(1 - x - y)g_z^*$$
$$\geq 8 - 6e - 3x(\epsilon - e) - 4(\epsilon - e)(1 - x - y)$$
$$= 4 - 2e + 4(1 - \epsilon) + (\epsilon - e)(x + 4y) > 0. \tag{22}$$

Therefore, it is stable.

Finally, consider Stern Judging, which has the equilibrium $g^* = g_x^* = g_y^* = g_z^* = \frac{1}{2}$ (Okada et al., 2018). Evaluating the Jacobian at this equilibrium gives us

$$\mathbf{J}(g^*) = \begin{pmatrix} \mathbf{J}_1 & \mathbf{J}_2 \\ \mathbf{J}_3 & \mathbf{J}_4 \end{pmatrix}, \tag{23a}$$

$$\mathbf{J}_1 = \begin{pmatrix} (2\epsilon - 1)x - 1 & (2\epsilon - 1)y & (2\epsilon - 1)z \\ (2e - 1)x & (2e - 1)y - 1 & (2e - 1)z \\ (2e - 1 + e - \epsilon)x & (2e - 1 + e - \epsilon)y & (2e - 1 + e - \epsilon)z - 1 \end{pmatrix}, \tag{23b}$$

$$\mathbf{J}_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 2(\epsilon - e)x & 2(\epsilon - e)y & 2(\epsilon - e)z \end{pmatrix}, \tag{23c}$$

$$\mathbf{J}_3 = \begin{pmatrix} (2\epsilon - 1)g_x^* x + g_x^* & (2\epsilon - 1)g_x^* y & (2\epsilon - 1)g_x^* z \\ (2e - 1)g_y^* x & (2e - 1)g_y^* y + g_y^* & (2e - 1)g_y^* z \\ (2e - 1 + e - \epsilon)g_z^* x & (2e - 1 + e - \epsilon)g_z^* y & (2e - 1 + e - \epsilon)g_z^* z + g_z^* \end{pmatrix}, \tag{23d}$$

$$\mathbf{J}_4 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 2(\epsilon - e)g_z^* x & 2(\epsilon - e)g_z^* y & 2(\epsilon - e)g_z^* z - 1 \end{pmatrix}. \tag{23e}$$

The characteristic equation is $(\lambda + 1)^2(\lambda^4 + c_3\lambda^3 + c_2\lambda^2 + c_1\lambda + c_0) = 0$ with coefficients:

$$c_3 = 4 - \epsilon x + (1 - \epsilon)x + (1 - 2e)(1 - x) > 0, \tag{24a}$$

$$c_2 = 3 + 3(1 - \epsilon x) + 3(1 - \epsilon)x + 3(1 - 2e)y + (2(1 - 2e) + 1 - \epsilon - e)z > 3, \tag{24b}$$

$$c_1 = 3(1 - \epsilon x) + 3(1 - \epsilon)x + 3(1 - 2e)y + 1 - \epsilon z + (2(1 - 2e) + 1 - \epsilon)z > 0 \tag{24c}$$

$$c_0 = 1 - \epsilon x + (1 - \epsilon)x + (1 - 2e)y + (1 - \epsilon - e)z > 0, \tag{24d}$$

Further, we have the following inequalities:

$$c_3c_2 - c_1 \geq 3c_3 - c_1 = 8 - 2(\epsilon - e)z > 0, \tag{25a}$$

$$c_3c_2c_1 - c_3^2c_0 - c_1^2 = k_1k_2 > 0, \tag{25b}$$

$$k_1 = 2(2 - (2\epsilon - 1)x + (1 - 2e)(y + z)) > 0, \tag{25c}$$

$$k_2 = (4 + 2(1 - 2\epsilon)x + 2(1 - 2e)y + (1 - 2e + 1 - \epsilon - e)z)^2 > 0. \tag{25d}$$

Therefore, g^* is stable by the Routh–Hurwitz criteria.

4. Discussion

Indirect reciprocity is a key mechanism to promote cooperation and has been well studied in the literature with both theoretical models and experiments. Models have shown how indirect reciprocity can evolve and how it can promote cooperation. Additionally, experimental evidence of indirect reciprocity has been found in both humans and other animals (Akçay et al., 2010; Nava et al., 2019; Seinen and Schram, 2006; Sommerfeld et al., 2007; Yoeli et al., 2013). Many of the mathematical models of indirect reciprocity assume a fast dynamic for reputations and a slow dynamic for strategies. That is to say, reputations of individuals are assessed and reach an equilibrium relatively quickly. Expected payoffs are calculated given these reputations. Then, individuals can change their strategies by imitating those who have greater payoffs. Reputations converge to an equilibrium quickly again, and so on. These models assume that reputations converge to a unique equilibrium, but whether or not they do is a crucial and understudied assumption. Here, we closed this gap, and have shown that the reputational dynamics that occur rapidly do converge to unique equilibria for each of the five standard norms and two assessment rules, which provides a basis for the previous analysis of the strategical dynamics in the literature.

There are several limitations of these results due to several assumptions we have made with these models, particularly in how reputations are assessed. Here, we have assumed that only one individual observes

each interaction. This simplifies the analyses somewhat as in Okada et al. (2018). Other private assessment models have been explored in the literature where more than one individual can observe the same interaction (either probabilistically, as in Hilbe et al. (2018), or the entire population, as in Fujimoto and Ohtsuki (2022) and Fujimoto and Ohtsuki (2023)). Simultaneous observation of the same interaction can create dependencies in the reputations, and therefore require different mathematical machinery that we do not consider in this paper. Additionally, we have assumed that norms are homogeneous in a population, and thus it is an open question as to whether or not reputations converge in populations with a mix of different norms.

There are several further modifications and extensions to the models explored here that would be interesting to study with respect to convergence of reputations. For example, the norms we considered are zeroth order (Scoring) and first order (Shunning, Staying, Simple Standing, and Stern Judging). Assignment of reputations under Scoring only depends upon the action of the donor while assignments of the other norms also depends on the reputation of the recipient. Higher order norms — those that use other information such as the previous reputation of the donor or multiple observations of the donor — may not lead to convergence to a unique set of reputations. For third order norms, the reputational system can be bistable, and when observers make multiple observations, reputations may not converge (unpublished research). We note that our systems of ODEs contain at most cubic polynomials with respect to the variables. The case of multiple observations is quartic, while the abductive reasoning model — which has conditional convergence — involves rational functions (Pandula et al., 2024). Finding higher order norms that are relevant to behaviour and that do not converge to reputational equilibria is an area for future research. Another research area that may lead to non-convergence is finite populations whether well-mixed or on a network. It is possible that small populations do not converge to unique equilibria or that there are network configurations that also impede convergence. Finally, we could also extend the model to include stochastic norms (Murase and Hilbe, 2023) and study their convergence.

Code and data availability

Code to verify analytical results is available at github.com/bmorsky/indirectReciprocity-convergence.

CRediT authorship contribution statement

Bryce Morsky: Writing – review & editing, Writing – original draft, Software, Investigation, Formal analysis, Conceptualization. **Joshua B. Plotkin:** Writing – review & editing, Conceptualization. **Erol Akçay:** Writing – review & editing, Conceptualization.

Declaration of competing interest

None.

Acknowledgements

JBP acknowledges support from the John Templeton Foundation, USA (grant #62281) and the Army Research Office (grant #W911NF-23-S-0001). EA acknowledges support from the U.S.-Israel 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.

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.

Brandt, Hannelore, Sigmund, Karl, 2004. The logic of reprobation: assessment and action rules for indirect reciprocation. J. Theoret. Biol. 231 (4), 475–486.

Chalub, Fabio A.C.C., Santos, Francisco C., Pacheco, Jorge M., 2006. The evolution of norms. J. Theoret. Biol. 241 (2), 233–240.

Fishman, Michael A., 2003. Indirect reciprocity among imperfect individuals. J. Theoret. Biol. 225 (3), 285–292.

Fujimoto, Yuma, Ohtsuki, Hisashi, 2022. Reputation structure in indirect reciprocity under noisy and private assessment. Sci. Rep. 12 (1), 10500.

Fujimoto, Yuma, Ohtsuki, Hisashi, 2023. Evolutionary stability of cooperation in indirect reciprocity under noisy and private assessment. Proc. Natl. Acad. Sci. 120 (20), e2300544120.

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.

Leimar, Olof, Hammerstein, Peter, 2001. Evolution of cooperation through indirect reciprocity. Proc. R. Soc. Lond. Ser. B: Biol. Sci. 268 (1468), 745–753.

Milinski, Manfred, Semmann, Dirk, Bakker, Theo CM, Krambeck, Hans-Jürgen, 2001. Cooperation through indirect reciprocity: image scoring or standing strategy? Proc. R. Soc. Lond. Ser. B: Biol. Sci. 268 (1484), 2495–2501.

Morsky, Bryce, Plotkin, Joshua B., Akçay, Erol, 2024. Indirect reciprocity with Bayesian reasoning and biases. PLoS Comput. Biol. 20 (4), e1011979.

Murase, Yohsuke, Hilbe, Christian, 2023. Indirect reciprocity with stochastic and dual reputation updates. PLoS Comput. Biol. 19 (7), e1011271.

Nakai, Yutaka, Muto, Masayoshi, 2008. Emergence and collapse of peace with friend selection strategies. J. Artif. Soc. Soc. Simul. 11 (3), 6.

Nava, Elena, Croci, Emanuela, Turati, Chiara, 2019. 'I see you sharing, thus I share with you': indirect reciprocity in toddlers but not infants. Palgrave Commun. 5 (1), 1–9.

Nowak, Martin A., Sigmund, Karl, 2005. Evolution of indirect reciprocity. Nature 437 (7063), 1291–1298.

Ohtsuki, Hisashi, Iwasa, Yoh, 2004. How should we define goodness?—reputation dynamics in indirect reciprocity. J. Theoret. Biol. 231 (1), 107–120.

Ohtsuki, Hisashi, Iwasa, Yoh, 2006. The leading eight: social norms that can maintain cooperation by indirect reciprocity. J. Theoret. Biol. 239 (4), 435–444.

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.

Pandula, Neel, Akçay, Erol, Morsky, Bryce, 2024. Indirect reciprocity with abductive reasoning. J. Theoret. Biol. 580, 111715.

Radzvilavicius, Arunas L, Stewart, Alexander J, Plotkin, Joshua B, 2019. Evolution of empathetic moral evaluation. eLife 8, e44269.

Santos, Fernando P., Pacheco, Jorge M., Santos, Francisco C., 2016. Evolution of cooperation under indirect reciprocity and arbitrary exploration rates. Sci. Rep. 6 (1), 37517.

Sasaki, Tatsuya, Okada, Isamu, Nakai, Yutaka, 2017. The evolution of conditional moral assessment in indirect reciprocity. Sci. Rep. 7 (1), 1–8.

Seinen, Ingrid, Schram, Arthur, 2006. Social status and group norms: Indirect reciprocity in a repeated helping experiment. Eur. Econ. Rev. 50 (3), 581–602.

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.

Takahashi, Nobuyuki, Mashima, Rie, 2006. The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity. J. Theoret. Biol. 243 (3), 418–436.

Wedekind, Claus, Milinski, Manfred, 2000. Cooperation through image scoring in humans. Science 288 (5467), 850–852.

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.