Author for correspondence:
Tatsuya Sasaki
e-mail: tatsuya.sasaki@univie.ac.at

THE ROYAL SOCIETY
PUBLISHING

Indirect reciprocity can overcome free-rider problems on costly moral assessment

Tatsuya Sasaki[1], Isamu Okada[2] and Yutaka Nakai[3]

[1]Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria
[2]Department of Business Administration, Soka University, 1-236 Tangi-cho, Hachioji-city, Tokyo 192-8577, Japan
[3]Faculty of Systems Engineering and Science, Shibaura Institute of Technology, Fukasaku 2307, Minuma-ku, Saitama-city, Saitama 337-8570, Japan

TS, 0000-0002-4635-1389; IO, 0000-0002-9943-2668

Indirect reciprocity is one of the major mechanisms of the evolution of cooperation. Because constant monitoring and accurate evaluation in moral assessments tend to be costly, indirect reciprocity can be exploited by cost evaders. A recent study crucially showed that a cooperative state achieved by indirect reciprocators is easily destabilized by cost evaders in the case with no supportive mechanism. Here, we present a simple and widely applicable solution that considers pre-assessment of cost evaders. In the pre-assessment, those who fail to pay for costly assessment systems are assigned a nasty image that leads to them being rejected by discriminators. We demonstrate that considering the pre-assessment can crucially stabilize reciprocal cooperation for a broad range of indirect reciprocity models. In particular for the most leading social norms, we analyse the conditions under which a prosocial state becomes locally stable.

1. Introduction

Natural selection disfavours indirect reciprocity unless the costs of observation and assessment are negligible [1]. According to proper social norms that distinguish good from evil, such as image-scoring norm [2], indirect reciprocity can promote cooperation even in large populations [3]. However, making moral assessments takes time and effort. Discriminators who incur no assessment cost thus appear as free-riders that erode a cooperative state achieved by discriminators who incurs the costs. Despite the advance of indirect reciprocity, the crucial question remains unsolved [4]: how can cooperation through indirect reciprocity be maintained when considering the costs associated with the assessment system?

To address this crucial question, we focus on the fact that this puzzling situation is closely related to the second-order free-rider problem in costly punishment [1]. The evolution of costly punishment, in striking contrast to indirect reciprocity, has been given much more attention over the last decades. In tackling the second-order free-rider problem, previous study significantly examined pool punishment [5–10]. The key aspect of pool punishment is its proactive mechanism to detect the second-order free-riders through unconditional prepayment. The mechanism paves the way for effectively punishing the second-order free-riders.

In this paper, we apply the essence of the pool-punishment mechanism to fix the issue of the costly moral assessment. In the next section, we introduce a basic model of indirect reciprocity and the known negative outcome from considering the assessment costs. In the Results section, we show how adopting a proactive assessment mechanism can improve the outcome.

Table 1. How the second-order rules make moral assessments in giving games with pre-assessment. 'G' and 'B' describe a good and bad image, respectively. In the donor's action, 'C' and 'D' describe giving help and refusing help, respectively.

conditions	recipient's image	G and nice	G and nice	B or nasty	B or nasty
	donor's action	C	D	C	D
assessment rule: what does the	simple standing	G	B	G	G
donor's image look like?	stern judging	G	B	B	G

2. Material and methods

We build upon the standard framework for the evolution of indirect reciprocity by reputation [11,12]. Using the framework, a strategy for discriminators is given by an assessment rule combined with an action rule. We base indirect reciprocity on the giving game, which is a two-player game in which one player acts as a donor and the other a recipient. The donor can choose to help the recipient by giving benefits $b > 0$ at personal cost $c > 0$ or not to help. We consider the following implementation error: a player who has intended to help involuntarily fails to do so with a probability e [1,13].

We start with a basic model in which each individual is endowed with a binary image score of 'good' or 'bad'. It is assumed that the discriminator's action rule is to help a good recipient or not to help a bad recipient. After observing every giving game, a unique assessment system assigns the donor's image by following a specific assessment rule. We assume that all discriminators share the same list of individual image scores provided by the assessment system. We later consider in particular the second-order assessment rule, which is a function of the donor's last action and the recipient's last image (table 1).

To study the evolution of discriminators, we consider a continuous-entry model: an individual's birth and death sometimes happen, and this changes the strategy distribution in the population [14]. We assume that in one's lifetime an individual infinitely plays the one-round giving game with different opponents. We consider infinitely large populations and analyse the replicator dynamics [15] for the following four strategies: (i) paying discriminator [Z] is willing to help a good recipient and refuses to help a bad recipient in the giving game. Also s/he is willing to pay for the assessment cost $k > 0$. (ii) Evading discriminator [W] similarly acts as a paying discriminator in the giving game, except that s/he is not willing to pay for the assessment cost. (iii) Cooperator [X] unconditionally intends to help a potential recipient and (iv) defector [Y] unconditionally intends not to help a potential recipient. Both cooperator and defector are not willing to pay for the assessment cost. We denote by x, y, z and w the frequencies of cooperators, defectors and paying and evading discriminators, respectively. The replicator dynamics for these strategies are described as $\mathrm{d}n/\mathrm{d}t = n(P_S - P)$, where n is the frequency of strategy S ($=X, Y, Z, W$), P_S is the expected pay-off given by the limit in the mean of the pay-off per round for strategy S and P is the average pay-off over the population, given by $xP_X + yP_Y + zP_Z + wP_W$.

To formalize the expected pay-offs, we denote by g_S the probability that a recipient with strategy S is helped by a given discriminator. In the basic model, this is identical to the fraction of good players within all S strategists. Let g be the population average of g_S, thus $g = xg_X + yg_Y + zg_Z + wg_W$. The population size is very large, so we may assume that the population configuration for g_S does not change between the consecutive one-round giving games [16]. Thus, the expected pay-offs are described as

$$\left.\begin{aligned} P_X &= (1-e)b[x+g_X(z+w)] - (1-e)c, \\ P_Y &= (1-e)b[x+g_Y(z+w)], \\ P_Z &= (1-e)b[x+g_Z(z+w)] - (1-e)cg - k \\ \text{and} \quad P_W &= (1-e)b[x+g_W(z+w)] - (1-e)cg. \end{aligned}\right\} \quad (2.1)$$

We note that in the basic model either paying or evading discriminators intend to help a potential recipient who has a good image, thus leading to $g_z = g_W$. This results in paying discriminators being worse off than evading discriminators. Substituting this into equation (2.1) yields

$$P_Z - P_W = -k < 0. \quad (2.2)$$

For any degree of the assessment cost k, evading discriminators dominate paying discriminators in the interior state space, and thus the population in the end attains a state that excludes paying discriminators. In the absence of cost payers, the assessment system cannot be established. Consequently, cooperation in that case would vanish without discrimination.

3. Results

The basic model reveals that considering cost evaders destabilizes indirect reciprocity irrespective of the assessment rule, as shown in previous work [1]. To stabilize indirect reciprocity, we examine an institutional variant of the basic model. As a first step, we extend the basic model to a two-stage game in which one round consists of the stage of payment for the observation costs followed by the stage of the giving game, which is the same as in the basic model. The first stage offers an opportunity to transfer some fees to a central account as in automatic utility payments.

The essential idea is to specifically assess the second-order free-rider. We consider a different binary moral code 'nice' or 'nasty'. The (unique) assessment system assigns a nice image to an individual if s/he pays the costs in the first stage, otherwise that individual is assigned a nasty image. In evaluating the donor's action of the giving game, as the first step we simply apply the existing assessment framework to the second stage, as in the basic model.

We keep the four strategies, cooperators, defectors, paying discriminators and evading discriminators, as before and assume that in the first stage, paying discriminators are willing to pay but the remaining cooperators, defectors and evading discriminators are not. We also modify the discriminator's action rule for the giving game as follows: either paying or evading discriminators give help if a potential recipient has a good *and* nice image, or otherwise (if bad *or* nasty), refuse help.

The extra assessment by the utilities payment system seriously lowers the image score for the second-order free-riders. For analytical simplicity, we assume that the utilities payment system is so perfect that no assessment error occurs for the first stage. All of the evading strategies: cooperators, defectors and evading discriminators (X, Y and W), therefore, are necessarily assessed as nasty. This yields $g_X = g_Y = g_W = 0$. (Note that with the extra assessment, g_S equals the probability of good and nice players.) Thus, equation
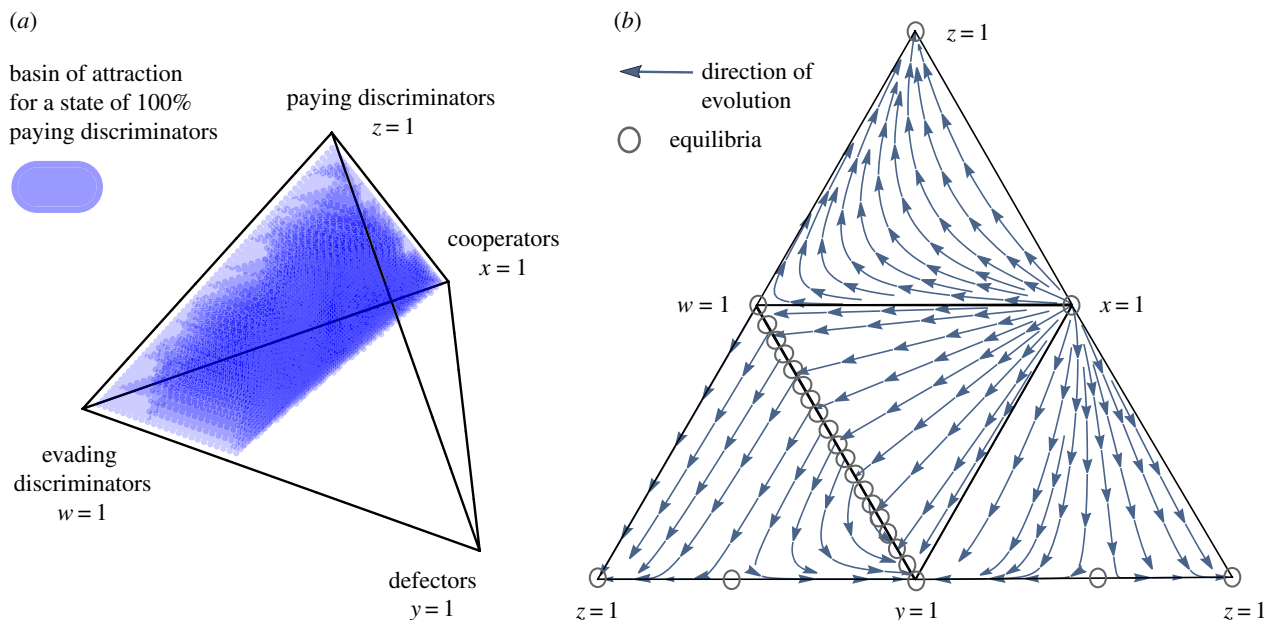
Figure 1. Pre-assessment of cost evaders stabilizes costly indirect reciprocity. (a) The tetrahedron describes a simplex of the state space $\{(x, y, z, w): x + y + z + w = 1; x, y, z, w \geq 0\}$. Each corner corresponds to the homogeneous state of each specific strategy. The basin of attraction for paying discriminators covers approximately 61.5% of the whole space. (b) The flow diagrams depict the direction of evolution on the boundary faces of the state-space simplex. The state space has no interior equilibrium, and all interior orbits converge to the boundary. Any mixed state of defectors and evading discriminators forms an equilibrium point on the edge $y + w = 1$. Parameters: $c = 1$, $b = 1.5$, $e = 0.01$, $k = 0.3$ and simple standing rule. (Online version in colour.)

(2.1) becomes

$$
\left.
\begin{aligned}
P_X &= (1 - e)bx - (1 - e)c, \\
P_Y &= (1 - e)bx, \\
P_Z &= (1 - e)b[x + g_Z(z + w)] - (1 - e)cg_Zz - k \\
\text{and} \quad P_W &= (1 - e)bx - (1 - e)cg_Zz.
\end{aligned}
\right\}
$$

(3.1)

It is clear that $P_Y \geq P_W \geq P_X$. To understand when the homogeneous state of paying discriminators $z = 1$ becomes locally stable, it is enough to check if $P_Z - P_Y > 0$ in the vicinity of $z = 1$ on the face $x = w = 0$. This yields

$$P_Z - P_Y = (1 - e)(b - c)g_Zz - k. \tag{3.2}$$

With suitable assessment rules, it is possible to have that $g_Z > 0$ in the vicinity of $z = 1$. In this case, the node $z = 1$ turns into a locally stable equilibrium when the net benefit $b - c$ is sufficiently large compared to the assessment cost k.

Finally, we demonstrate how the model with the extra assessment improves the results for some of the most leading assessment rules. We examine simple standing [13,16] and stern judging [17], the only two second-order assessment rules in the leading eight norms [10,11]. According to the discriminator's action rule in the variant, we extend simple standing and stern judging as in table 1. These rules assign a good image to those who help a good and nice recipient with no implementation error (probability $(1 - e)g$) and also a good image to those who refuse to help a bad or nasty recipient (probability $1 - g$). By assumption of the image dynamics, the sum of these probabilities should equal g_Z. Considering also $g = g_Zz$ then leads to the recursive equation for g_Z, $g_Z = (1 - e)g_Zz + (1 - g_Zz)$. This yields $g_Z = 1/(1 + ez)$. Hence, the necessary and sufficient condition for the homogeneous state of paying discriminators ($z = 1$) to be locally stable either

under the simple standing or stern judging rule is

$$\frac{1 - e}{1 + e}(b - c) - k > 0. \tag{3.3}$$

Figure 1a shows the basin of attraction for $z = 1$, and figure 1b depicts the flow on the boundary faces of the state space under simple standing. If we assume assessment errors in the first stage, the image dynamics become more complicated but the main results remain qualitatively unchanged—paying discriminators can stabilize with the pre-assessment of cost evaders (electronic supplementary material, S1).

4. Discussion

Since the definitive 2013 work by Suzuki & Kimura [1], the evolution of indirect reciprocity relying on costly assessment systems has been explicitly recognized as one of the inevitable issues that challenge the advance of indirect reciprocity [4]. To address the issue, we considered a simple pre-assessment mechanism that is set prior to the primary game in order to detect and label cost evaders. We then demonstrated that the mechanism considered leads to stabilizing costly indirect reciprocity under the most leading social norms, simple standing and stern judging.

Our results are potentially applicable to a broad range of existing indirect reciprocity models, such as tolerant scoring [18], group scoring [19], reputation-based punishment [20], mixed public and private interactions [21], optional interactions [22] and finite populations [23]. On the one hand, managing more complicated assessment systems, such as in [18–21], would be more costly, and thus it is worth considering pre-assessment mechanisms for reducing the temptation to evade cost sharing. On the other hand, as in the case of pool punishment [8], jointly considering optional interactions [22] and finite populations [23] might facilitate establishing pre-assessment mechanisms.

Another promising avenue for future studies would be to explore costly indirect reciprocity on more realistic structured populations. Recent studies using structured populations suggest the importance of cooperator assortment based on reputation [24,25]. However, little is known about how information cost affects reputation-based reciprocity on a network. In the case of the second-order free-rider problem in costly punishment, considering the locality of interactions among players can solve the problem by separating costly punishers from the second-order free-riders [26]. Similarly, the extension to structured populations may lead to significantly different outcomes for paying and evading discriminators.

We left out an advanced issue of analysing non-linking discriminators [27] who act as paying discriminators yet are willing to help cost evaders with a good image. Non-linking discriminators can invade paying discriminators by neutral drift. The preliminary results indicate that considering implementation or assessment errors for the first stage can lead paying discriminators to become better off than non-linking ones, as in fixing neutral drift between conditional and unconditional cooperators [13]. Further investigation is planned in future work.

We note that prepayment for assessment systems can be viewed as a kind of contribution to collective action. Thus, our results corroborate those of previous studies on two-stage games in which reciprocal behaviours in the second stage are linked to a collective action in the first stage. For instance, Panchanathan & Boyd showed that collective action in the first stage can be maintained by considering a shunning strategy that in the second stage withholds help for those who failed to contribute in the first stage [28]. Together, the present results further imply that such a pro-active social mechanism, which can discriminate those who deserve to enter social exchange, and reciprocal norms within social exchange may evolve jointly.

References

1. Suzuki S, Kimura H. 2013 Indirect reciprocity is sensitive to costs of information transfer. *Sci. Rep.* **3**, 1435. (doi:10.1038/srep01435)

2. Nowak MA, Sigmund K. 1998 Evolution of indirect reciprocity by image scoring. *Nature* **282**, 462–466. (doi:10.1038/31225)

3. Nowak MA, Sigmund K. 2005 Evolution of indirect reciprocity. *Nature* **437**, 1291–1298. (doi:10.1038/nature04131)

4. Rand DG, Nowak MA. 2013 Human cooperation. *Trends Cogn. Sci.* **17**, 413–425. (doi:10.1016/j.tics.2013.06.003)

5. Yamagishi T. 1986 The provision of a sanctioning system as a public good. *J. Personal. Soc. Psychol.* **51**, 110–116. (doi:10.1037/0022-3514.51.1.110)

6. Traulsen A, Röhl T, Milinski M. 2012 An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proc. R. Soc. B* **279**, 3716–3721. (doi:10.1098/rspb.2012.0937)

7. Zhang B, Li C, De Silva H, Bednarik P, Sigmund K. 2014 The evolution of sanctioning institutions: an experimental approach to the social contract. *Exp. Econ.* **17**, 285–303. (doi:10.1007/s10683-013-9375-7)

8. Sigmund K, De Silva H, Traulsen A, Hauert C. 2010 Social learning promotes institutions for governing the commons. *Nature* **466**, 861–863. (doi:10.1038/nature09203)

9. Schoenmakers S, Hilbe C, Blasius B, Traulsen A. 2014 Sanctions as honest signals—the evolution of pool punishment by public sanctioning institutions. *J. Theor. Biol.* **356**, 36–46. (doi:10.1016/j.jtbi.2014.04.019)

10. Szolnoki A, Szabó G, Perc M. 2011 Phase diagrams for the spatial public goods game with pool punishment. *Phys. Rev. E* **83**, 036101. (doi:10.1103/PhysRevE.83.036101)

11. Ohtsuki H, Iwasa Y. 2004 How should we define goodness?—reputation dynamics in indirect reciprocity. *J. Theor. Biol.* **231**, 107–120. (doi:10.1016/j.jtbi.2004.06.005)

12. Brandt H, Sigmund K. 2004 The logic of reprobation: assessment and action rules for indirect reciprocation. *J. Theor. Biol.* **231**, 475–486. (doi:10.1016/j.jtbi.2004.06.032)

13. Panchanathan K, Boyd R. 2003 A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *J. Theor. Biol.* **224**, 115–126. (doi:10.1016/S0022-5193(03)00154-1)

14. Brandt H, Sigmund K. 2005 Indirect reciprocity, image scoring, and moral hazard. *Proc. Natl Acad. Sci. USA* **102**, 2666–2670. (doi:10.1073/pnas.0407370102)

15. Hofbauer J, Sigmund K. 1998 *Evolutionary games and population dynamics*. Cambridge, UK: Cambridge University Press.

16. Leimar O, Hammerstein P. 2001 Evolution of cooperation through indirect reciprocity. *Proc. R. Soc. Lond. B* **268**, 745–753. (doi:10.1098/rspb.2000.1573)

17. Pacheco JM, Santos FC, Chalub FAC. 2006 Stern-judging: a simple, successful norm which promotes cooperation under indirect reciprocity. *PLoS Comput. Biol.* **2**, e178. (doi:10.1371/journal.pcbi.0020178)

18. Berger U. 2011 Learning to cooperate via indirect reciprocity. *Games Econ. Behav.* **72**, 30–37. (doi:10.1016/j.geb.2010.08.009)

19. Nax HH, Perc M, Szolnoki A, Helbing D. 2015 Stability of cooperation under image scoring in group interactions. *Sci. Rep.* **5**, 12145. (doi:10.1038/srep12145)

20. Ohtsuki H, Iwasa Y, Nowak MA. 2009 Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature* **457**, 79–82. (doi:10.1038/nature07601)

21. Ohtsuki H, Iwasa Y, Nowak MA. 2015 Reputation effects in public and private interactions. *PLoS Comput. Biol.* **11**, e1004527. (doi:10.1371/journal.pcbi.1004527)

22. Olejarz J, Ghang W, Nowak MA. 2015 Indirect reciprocity with optional interactions and private information. *Games* **6**, 438–457. (doi:10.3390/g6040438)

23. Santos FP, Santos FC, Pacheco JM. 2016 Social norms of cooperation in small-scale societies. *PLoS Comput. Biol.* **12**, e1004709. (doi:10.1371/journal.pcbi.1004709)

24. Wang Z, Wang L, Yin Z-Y, Xia C-Y. 2012 Inferring reputation promotes the evolution of cooperation in spatial social dilemma games. *PLoS ONE* **7**, e40218. (doi:10.1371/journal.pone.0040218)

25. Cuesta JA, Gracia-Lázaro C, Ferrer A, Moreno Y, Sánchez A. 2015 Reputation drives cooperative behaviour and network formation in human groups. *Sci. Rep.* **5**, 7843. (doi:10.1038/srep07843)

26. Helbing D, Szolnoki A, Perc M, Szabó G. 2010 Evolutionary establishment of moral and double moral standards through spatial interactions. *PLoS Comput. Biol.* **6**, e1000758. (doi:10.1371/journal.pcbi.1000758)

27. Yamagishi T, Takahashi N. 1994 Evolution of norms without metanorms. In *Social dilemmas and cooperation* (eds U Schulz, W Albers, U Mueller), pp. 311–326. Berlin, Germany: Springer.

28. Panchanathan K, Boyd R. 2004 Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* **432**, 499–502. (doi:10.1038/nature02978)