RESEARCH ARTICLE

Justified defection is neither justified nor unjustified in indirect reciprocity

Hitoshi Yamamoto [1,2]*, Takahisa Suzuki[3], Ryohei Umetani[2]

1 Faculty of Business Administration, Rissho University, Tokyo, Japan, **2** Graduate School of Business Administration, Rissho University, Tokyo, Japan, **3** College of Policy Studies, Tsuda University, Tokyo, Japan

* hitoshi@ris.ac.jp

Abstract

Indirect reciprocity is one of the major mechanisms in the evolution of human cooperation. In indirect reciprocity, social norms with which individuals distinguish good people from bad people play essential roles. Despite extensive studies on the evolution of cooperation in indirect reciprocity, little is known about which social norms people actually adopt. Here we reveal what kind of norms are adopted by people in indirect reciprocal situations in daily life by using scenario-based experiments. The results showed that people evaluated "justified defection" as neither good nor bad and withheld their evaluation. Theoretically, social norms that evaluate justified defection as good are required for cooperation to be stable. However, the norm that people actually adopted deviates from the theoretical predictions. Our results indicate the necessity to reconsider the justification of "justified defection" in the evolution of cooperation.

Introduction

Mutual cooperation in a competitive situation has been an essential driver in the development of human society, but its mechanism is a difficult puzzle to solve. Especially, indirect reciprocity [1–8] is a significant solution for the large-scale and highly flexible systems often observed in modern society. For indirect reciprocity to work, norms are needed that distinguish good people from bad people so that cooperation is selectively directed to good people. To distinguish a good person from a bad person, two types of information about him are required. One is information about his past helpfulness (first-order information), and the other is information about the helpfulness of his recipient (second-order information).

The simplest norm of indirect reciprocity is image-scoring [3,4], which refers only to first-order information. However, image-scoring is not evolutionarily stable because it is vulnerable to errors and invasion by perfect defectors [9]. Therefore, second-order information is needed to sustain cooperation in indirect reciprocity. The major norms that use second-order information consist of stern-judging [10,11], simple-standing [1,5,12], and shunning [13]. As for the second-order information, it is crucial to consider the assessment rule when the player's recipient has a bad reputation because evaluating the action of the player to a good recipient is easy. Naturally, cooperation with and defection against the good should respectively be evaluated as good and bad.

In particular, defection against a bad recipient (called justified defection) is a key factor in establishing cooperation [5,6,12,14–16]. This is because the main reason for the vulnerability of image-scoring is that it cannot distinguish justified defections from unjustified ones [9]. Furthermore, eight norms that can stabilize cooperation identified from 4,096 candidate norms by theoretical approaches (called the "leading eight") [17,18] have a common relevant feature—if a good donor defects against a bad recipient, the donor is evaluated as good. In this study, we focus on the evaluation of justified defection in real human behavior.

Despite numerous experiments on indirect reciprocity [19–24] having been conducted, it remains unclear how humans actually use second-order information. On one hand, a study [19] has reported that people tend to use only first-order information because people do not like to process complex information when they can assess others on the basis of less complex information. On the other hand, other studies [25,26] have shown that even if there is a cost associated with obtaining information, people make decisions on the basis of both first- and second-order information. Furthermore, an assessment rule that ignores actions against bad recipients has been proposed as an evolutionarily stable norm for cooperation [27,28]. In addition, experimental evidence shows that the evaluations about bad people are more uncertain than those about good ones [29].

As mentioned above, there is a deviation between predictions from theories and experimental results about the assessment rule for the actions against bad recipients. In particular, the distributions of assessment rules toward bad recipients in humans are unknown. By analyzing how second-order information, which is theoretically required to sustain cooperation, is used by humans in evaluating others, the function of the omission of punishment and the justified defection in human society will be revealed.

Here, we design experiments to clarify the distribution of people's assessment rules in indirect reciprocity. In the scenario-based experiments, two people appeared in the scenario. One was a donor who does or does not cooperate with his/her recipient. The other was the recipient who asks the donor for help. Participants were asked to rate the donor as good or bad by observing when the donor chose to cooperate or not with the recipient. The recipient is labeled either good or bad. Participants evaluate four scenes in the combination of donor behaviors and recipient labels. Since the purpose of this study is to investigate which norms people adopt in real society, we adopted scenario-based experiments rather than an economic experiment.

Most theoretical studies have adopted an assumption in which a single norm pervades throughout a population resulting in all individuals using the same norm. However this assumption clearly oversimplifies the reality. In recent years, researchers have tackled this assumption by developing a model in which various norms coexist in the population [30,31]. However, if any of the norms proposed in the theoretical studies are actually adopted by people, the assessment of the donor by observers should be widely distributed. For example, if there is a mixture of good and bad evaluations for "justified defections" (i.e., not cooperating with bad recipients), the variance in the evaluation of justified defections should increase. This study explores whether people's norms remain within the norms proposed in previous theoretical studies or whether people adopt norms that are different from those assumed in theoretical research.

Methods

We designed three scenario-based experiments to explore how humans evaluate the donor's action using the first and second-order information in indirect reciprocity. Participants were recruited through a Japanese crowdsourcing service (Yahoo! Crowdsourcing). To broaden the generalizability of findings, each experiment had completely different participants. All

experiments were approved by the ethics committee of Rissho University, application number 29–3 and 30–15. Informed consent was obtained from all subjects.

Experiment 1

We recruited 200 participants (female = 71, mean age = 46.31 (S.D. = 9.96)). Participants evaluated a donor's behavior using first- and second-order information. First-order information is the donor's behavior (cooperation/defection), and second-order information is recipient reputation (good/bad). Combining both types of information, participants rated four scenes: cooperation with the good recipient (CtoG), defection against the good recipient (DtoG), cooperation with the bad recipient (CtoB), and defection against the bad recipient (DtoB) (Fig 1).

Participants first evaluated the scenes in order of CtoG and DtoG. To eliminate order effects on the cases of a bad recipient, the order of CtoB and DtoB was randomized among participants. The specific scenarios are as follows. In the first scenario, the participants were assumed to be workers in a restaurant. Assume that a colleague, Alice (recipient), asks another colleague, Bob (Donor), to take over the night shift, and Bob agrees (cooperation) or refuses (defection). We also controlled Alice's reputation. The cases in which Alice has a good reputation are as follows.

> Alice works hard and is always willing to take over when others cannot come to do the night shift. That is why Alice is liked very much by colleagues in the restaurant including you.

The cases in which Alice has a bad reputation are as follows.

> Alice is not serious about her work and often takes time off from the night shift for various reasons. She often takes time off for selfish reasons and is not liked by her colleagues including you.

In the actual experiment, Alice and Bob's names were converted into common Japanese names. After reading the scenario, participants rated how they assessed Bob's behavior from three viewpoints using a 5-point scale: "Bob is a reliable person", "Do you like Bob?", and "Do you feel sympathetic to Bob?" The evaluation scores for the donor's action were obtained by simply adding together the scores of these three statements (See S1–S3 Tables).

Experiment 2

Experiment 2 was conducted using different scenarios in order to generalize the results by eliminating the dependence on a specific scenario. We recruited 200 participants (female = 93, mean age = 45.71 (S.D. = 10.45)). The basic settings are the same as for Experiment 1. The differences from Experiment 1 are as follows. The donor and recipient were changed from restaurant workers to neighborhood residents, and the recipient (Alice) asks the donor (Bob) for personal advice about distress. The cases in which Alice's has a good reputation are as follows.

> Alice is always willing to respond to other people's requests and give them advice in a friendly manner. For this reason, Alice is well-liked by her acquaintances, including you.

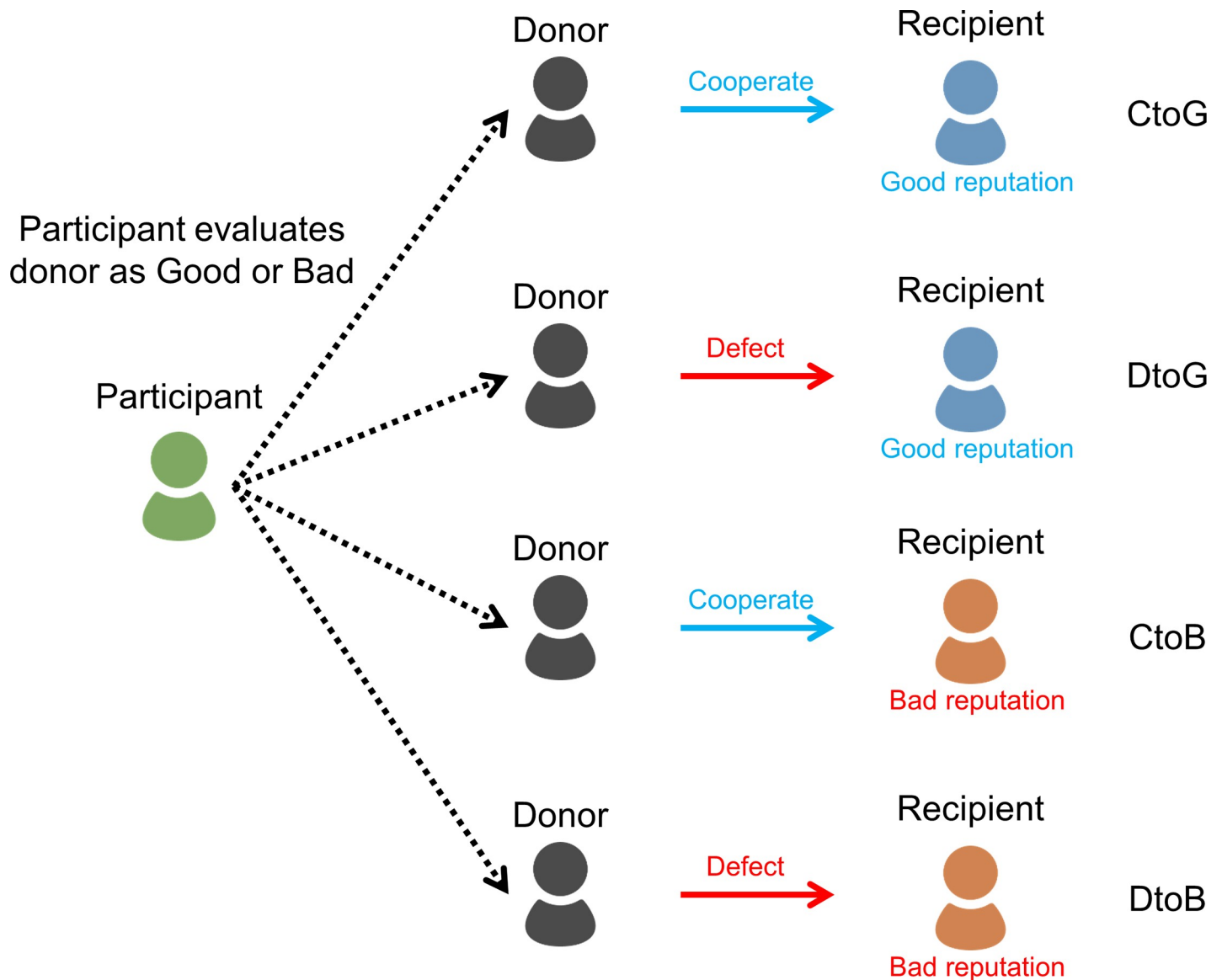The cases in which Alice has a bad reputation are as follows.

Fig 1. Overview of the experiment: Participants evaluate four scenes that consist of a combination of donor behavior (cooperate / defect) and recipient reputation (good / bad).

Alice has a habit of talking negatively about other people behind their backs. Also, she always thinks that she is right and sometimes forces her opinions on others, so her acquaintances, including you, feel that Alice is somewhat unpleasant.

Experiment 3

We set up the scene of a giving game to examine not only the evaluation of behaviors in daily life but also behavior in an economic experiment. We recruited 200 participants (female = 87, mean age = 43.37 (S.D. = 12.31)). Participants were given an explanation of the rules of a giving game, in which a donor and a recipient were randomly matched in an iterated game, and

then asked to rate the behavior of the donor. Participants rated donor behavior (donate / not donate) in the giving game the same as in Experiments 1 and 2. Recipient reputation was controlled by how many times he/she chose to donate during his/her past five rounds as a donor (5 (good), 0 (bad)). To show the participants that it was an economic game in an anonymous situation, we described the donor and the recipient as Player A and Player B. The cases in which the recipient (Player B) has a good reputation are as follows.

> Player A is now the "donor" and Player B is now the "recipient". Player B selected "donate" all five times when he/she was previously the donor.

> The cases in which Player B has a bad reputation are as follows.

> Player A is now the "donor" and Player B is now the "recipient". Player B selected "not donate" all five times when he/she was previously the donor.

Results

We first analyze distributions of donor ratings in each scenario. The curves in all diagrams in Fig 2 are Gaussian kernel density estimations of the distributions of the evaluation of a donor's action in each experiment. Kernel density estimation allows us to smooth over the distribution to avoid harsh edges caused by a sampling size, as found in the histogram method. Consistent with intuitive expectations, CtoG is rated as good and DtoG is rated as bad. CtoB is also rated as good. These results suggest that many participants adopt either image-scoring or simple-standing norms. DtoB, on the other hand, shows remarkable unique results. The distribution of ratings is centralized, and participants rate DtoB as neither good nor bad. If image-scoring and simple-standing are mixed in the participants, two peaks should appear, so it can be said that many participants do not adopt either image-scoring or simple-standing. In other words, participants adopt a norm of holding back on the evaluation to DtoB and not updating the reputation of donors.

To analyze these differences in distributions statistically, we employed quantile regression [32]. The ordinary linear regression assumes that the mean value represents the whole of the distribution and so enables us to estimate the effects of independent variables on the mean value of a dependent variable. The quantile regression enables us to estimate how independent variables affect each quantile of a dependent variable. Thus, the quantile regression could be used to explore the difference in the shape of the distribution of a dependent variable between conditions.

In this study, each quantile of evaluation scores was estimated for three experiments separately. To compare CtoB and other conditions, we used the CtoG dummy, DtoG dummy, and DtoB dummy as independent variables.

Table 1 shows the results of the quantile regression. The difference between CtoB and DtoB is small and not significant in the .25th quantile, but the .50th and .75th quantiles of CtoB are larger than those of DtoB, and these differences were significant. These results meant that the left half of the distribution was similar between CtoB and DtoB, and in another area, CtoB was more right-skewed than DtoB. The whole distribution of CtoB was similar to that of CtoG in that both were right-skewed and there was no significant difference at the 5 percent level, except the .25th and .50th quantiles of Experiment 2. DtoG is more left-skewed than CtoB in all experiments. To summarize these results, relative to CtoB, CtoG has a similar distribution, DtoG shifts to the left, and DtoB shifts to the center.
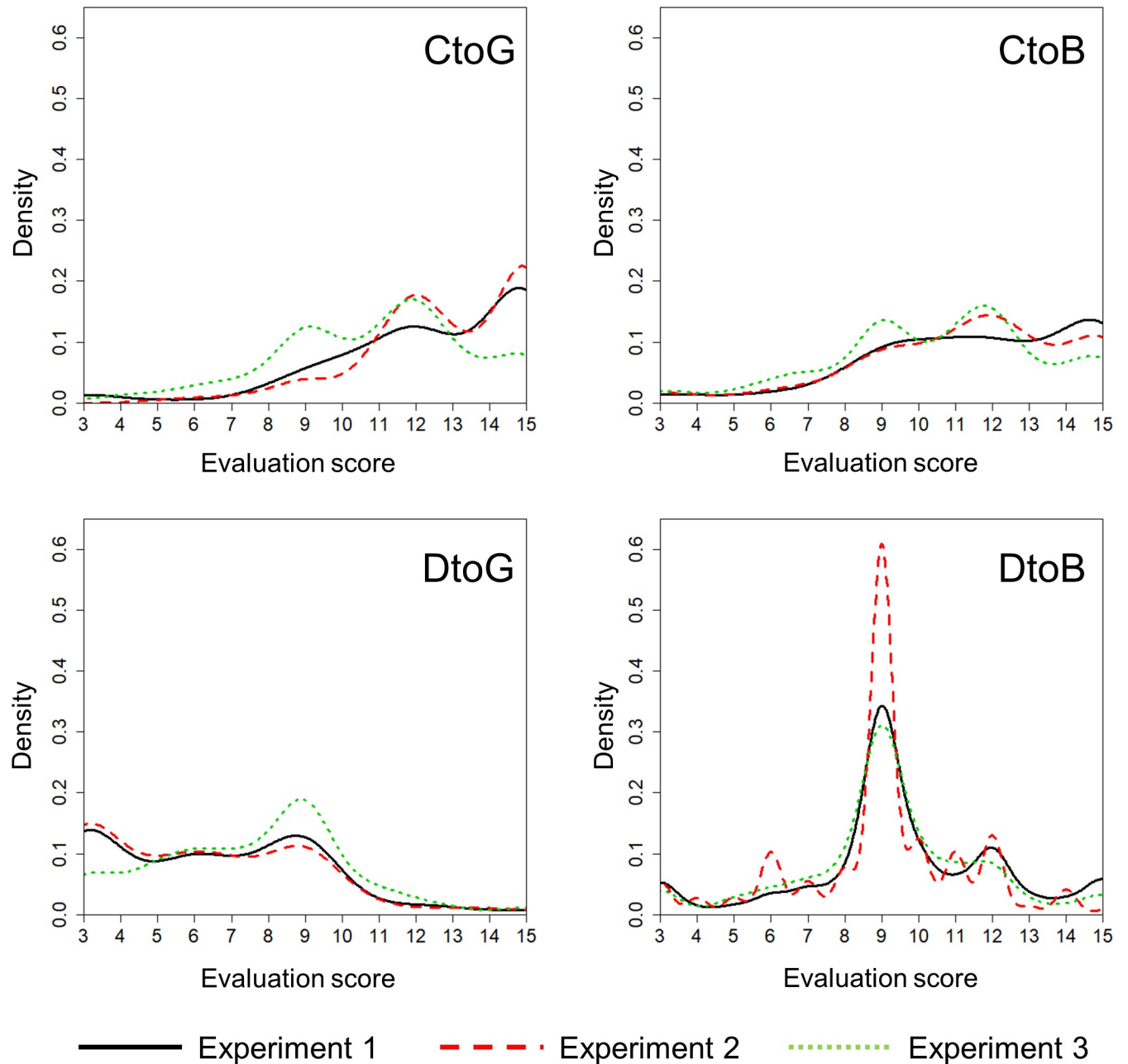
Fig 2. Distribution of evaluation for donors: The horizontal axis shows the evaluation score. The scores were calculated by simply adding together the scores of three statements on a 5-point scale, so it has minimum and maximum values of 3 and 15. The vertical axis shows the density of kernel density estimation. The solid black line, dotted red line, and dotted green line represent Experiments 1, 2, and 3, respectively.

https://doi.org/10.1371/journal.pone.0235137.g002

Discussion

For indirect reciprocity to work, people need to have norms that distinguish the good and bad of others' behavior and cooperate only with good people. Extensive theoretical and empirical studies have looked for suitable norms for the evolution of cooperation, but there is no clear answer for the norms people adopt in our society. From a theoretical prediction, to stabilize cooperation through indirect reciprocity, a norm is needed that prescribes regarding as good

Table 1. Results of quantile regression: CtoB was the baseline category. We used "sqreg" command in Stata var.15 with a bootstrap method repeated 400 times.

		restaurant worker		personal advice		economic game	
		Coef.		Coef.		Coef.	
q25	CtoG	1		3	***	0	
	DtoB	-1		0		-1	
	DtoG	-7	***	-6	***	-3	***
	Cons	10	***	9	***	9	***
q50	CtoG	1		1	*	0	
	DtoB	-3	***	-3	***	-2	***
	DtoG	-6	***	-6	***	-3	***
	Cons	12	***	12	***	11	***
q75	CtoG	1		1		0	
	DtoB	-3	***	-4	***	-2	***
	DtoG	-5	***	-5	***	-3	***
	Cons	14	***	14	***	12	***

*$p < .05$.

**$p < .01$.

***$p < .001$.

https://doi.org/10.1371/journal.pone.0235137.t001

those who do not cooperate with a bad person (justified defection) [5,6,12,14–16]. On the other hand, some empirical evidence has indicated that people use simple norms to rate cooperation as good and defection as bad [19]. We analyzed the distributions of norms in indirect reciprocity using scenario-based experiments.

Our results reveal that people do not evaluate the defection with a bad recipient as either good or bad and withheld their evaluation. This means that justified defection is neither justified nor unjustified. We also found, as expected, that cooperation was mostly considered good and defection with good recipients was considered bad.

As a result, the norms adopted by people can be summarized as shown in Table 2. Typical norms proposed by previous studies have unconditionally evaluated donor behavior as either good or bad. In fact, people held back on the evaluation to justifiably defect. The norm observed in our study, which can be described as GBGN (good, bad, good, neutral), is consistent with the "L1" rule of the "leading eight" norms [17] from the viewpoint of an evaluation rule. However, our experiments only examined the assessment rule and did not clarify behavioral rules [33]. The relationship between the evaluation rule and the behavioral rule is a future extension. It is necessary to explore why most participants adopt GBGN and under what conditions the norm is more advantageous than other norms.

Table 2. How social norms make assessments in indirect reciprocity: "G" and "B" describe a good and bad reputation, respectively. "C" and "D" denote cooperation and defection, respectively. "N" means the reputation of a donor remains neutral ("Neutral").

Recipient's reputation	G	G	B	B
Donor's action	C	D	C	D
The observed norm	G	B	G	N
Simple-standing	G	B	G	G
Image-scoring	G	B	G	B
Stern-judging	G	B	B	G
Shunning	G	B	B	B

https://doi.org/10.1371/journal.pone.0235137.t002

In a theoretical model of indirect reciprocity, a player's reputation depends on his/her behavior, the reputation of his/her recipients, and the norms of observers. Therefore, the good and bad evaluations of the recipients may not be directly associated with cooperation and defection. To unify the reputations of the recipients among all participants in our experiments, they were designated as perfect cooperators or perfect defectors. In the future study, it will be necessary to consider the higher-order information, i.e. reputation of the donor or reputation of the previous recipient of the present recipient [34]. This may allow us to understand how the reputation of the recipients emerges under conditions of repeated interaction.

This paper is a first step to reveal a neutral reputation as an evaluation of justified defection. Many extensions are needed to understand the mechanism of this phenomenon. To explore the reason for the neutral evaluation of the defection against bad recipients, it is necessary to analyze whether the neutrality in evaluation is "don't know" or "withhold evaluation." People acquire an understanding of indirect reciprocity in childhood [35], so another challenging extension is to explore when second-order information begins to be adopted during growth. We should also consider cases where the observer has had a direct reciprocal relationship with the donor or recipient in the past [36]. Extensions by theoretical approaches can also be considered. It is necessary to analyze whether GBGN is evolutionarily stable. Another extension would be exploring whether the GBGN norm survives in an environment where many norms coexist [30].

Supporting information

S1 Fig. Experiment 1. Restaurant worker scenario.
(DOCX)

S2 Fig. Experiment 2. Personal advice scenario.
(DOCX)

S3 Fig. Experiment 3. Economic game scenario.
(DOCX)

S1 Table. Experiment 1. Restaurant worker scenario.
(DOCX)

S2 Table. Experiment 2. Personal advice scenario.
(DOCX)

S3 Table. Experiment 3. Economic game scenario.
(DOCX)

S1 Text.
(DOCX)

Author Contributions

Conceptualization: Hitoshi Yamamoto.

Data curation: Hitoshi Yamamoto, Ryohei Umetani.

Formal analysis: Hitoshi Yamamoto, Ryohei Umetani.

Funding acquisition: Hitoshi Yamamoto.

Methodology: Takahisa Suzuki.

Project administration: Hitoshi Yamamoto.

Visualization: Hitoshi Yamamoto.

Writing – original draft: Hitoshi Yamamoto.

Writing – review & editing: Hitoshi Yamamoto, Takahisa Suzuki.

References

1. Sugden R. The Economics of Rights, Cooperation and Welfare. Oxford: Basil Blackwell; 1986.

2. Alexander R. The Biology of Moral Systems. Aldine de Gruyter, New York; 1987.

3. Nowak MA, Sigmund K. Evolution of indirect reciprocity by image scoring. Nature. 1998; 393:573–577. https://doi.org/10.1038/31225 PMID: 9634232

4. Nowak MA, Sigmund K. The dynamics of indirect reciprocity. J Theor Biol. 1998; 194(4):561–574. https://doi.org/10.1006/jtbi.1998.0775 PMID: 9790830

5. Leimar O, Hammerstein P. Evolution of cooperation through indirect reciprocity. Proc. R. Soc. Lond. B. 2001; 268(1468):745–753. https://doi.org/10.1098/rspb.2000.1573 PMID: 11321064

6. Wedekind C, Braithwaite VA. The long-term benefits of human generosity in indirect reciprocity. Curr. Biol. 2002; 12(12):1012–1015. https://doi.org/10.1016/s0960-9822(02)00890-4 PMID: 12123575

7. Nowak MA, Sigmund K. Evolution of indirect reciprocity. Nature. 2005; 437(7063):1291–1298. https://doi.org/10.1038/nature04131 PMID: 16251955

8. Sommerfeld RD, Krambeck HJ, Semmann D, Milinski M. Gossip as an alternative for direct observation in games of indirect reciprocity. Proc. Natl. Acad. Sci. USA. 2007; 104(44):17435–17440. https://doi.org/10.1073/pnas.0704598104 PMID: 17947384

9. Sigmund K. The calculus of selfishness. vol. 6. Princeton University Press; 2010.

10. Kandori M. Social norms and community enforcement. Rev Econ Stud. 1992; 59(1):63–80. https://doi.org/10.2307/2297925

11. Pacheco JM, Santos FC, Chalub FAC. Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity. PLoS Comput Biol. 2006; 2(12):e178. https://doi.org/10.1371/journal.pcbi.0020178 PMID: 17196034

12. Panchanathan K. A tale of two defectors: the importance of standing for evolution of indirect reciprocity. J Theor Biol. 2003; 224(1):115–126. https://doi.org/10.1016/s0022-5193(03)00154-1 PMID: 12900209

13. Takahashi N.; Mashima R. The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity. J Theor. Biol. 2006; 243, 418–436. https://doi.org/10.1016/j.jtbi.2006.05.014 PMID: 16904697

14. Brandt H. & Sigmund K. The logic of reprobation: assessment and action rules for indirect reciprocation. J Theor Biol. 2004, 231, 475–486. https://doi.org/10.1016/j.jtbi.2004.06.032 PMID: 15488525

15. Suzuki S., Akiyama E. Evolution of indirect reciprocity in groups of various sizes and comparison with direct reciprocity. J Theor Biol. 2007, 245, 539–552. https://doi.org/10.1016/j.jtbi.2006.11.002 PMID: 17182063

16. Santos FP, Pacheco JM, Santos FC. Evolution of cooperation under indirect reciprocity and arbitrary exploration rates. Sci Rep. 2016; 6:37517. https://doi.org/10.1038/srep37517 PMID: 27892509

17. Ohtsuki H, Iwasa Y. The leading eight: social norms that can maintain cooperation by indirect reciprocity. J Theor Biol. 2006; 239(4):435–44. https://doi.org/10.1016/j.jtbi.2005.08.008 PMID: 16174521

18. Ohtsuki H, Iwasa Y. Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. J Theor Biol. 2007, 244:518–31. https://doi.org/10.1016/j.jtbi.2006.08.018 PMID: 17030041

19. Milinski M, Semmann D, Bakker TCM, Krambeck HJ. Cooperation through indirect reciprocity: image scoring or standing strategy? Proc. R. Soc. Lond. B. 2001; 268(1484):2495–2501. https://doi.org/10.1098/rspb.2001.1809 PMID: 11747570

20. Wedekind C, Milinski M. Cooperation through image scoring in humans. Science. 2000; 288 (5467):850–852. https://doi.org/10.1126/science.288.5467.850 PMID: 10797005

21. Milinski M, Semmann D, Krambeck HJ. Reputation helps solve the 'tragedy of the commons'. Nature. 2002; 415(6870):424–426. https://doi.org/10.1038/415424a PMID: 11807552

22. Bolton GE, Katok E, Ockenfels A. Cooperation among strangers with limited information about reputation. J Public Econ. 2005; 89(8):1457–1468. https://doi.org/10.1016/j.jpubeco.2004.03.008

23. Ule A, Schram A, Riedl A, Cason TN. Indirect punishment and generosity toward strangers. Science. 2009; 326(5960):1701–1704. https://doi.org/10.1126/science.1178883 PMID: 20019287

24. Yoeli E, Hoffman M, Rand DG, Nowak MA. Powering up with indirect reciprocity in a large-scale field experiment. Proc. Natl. Acad. Sci. USA. 2013; 110(Supplement 2):10424–10429. https://doi.org/10.1073/pnas.1301210110 PMID: 23754399

25. Swakman V, Molleman L, Ule A, Egas M. Reputation-based cooperation: Empirical evidence for behavioral strategies. Evol Hum Behav. 2016; 37(3):230–235. https://doi.org/10.1016/j.evolhumbehav.2015.12.001

26. Okada I, Yamamoto H, Sato Y, Uchida S, Sasaki T. Experimental evidence of selective inattention in reputation-based cooperation. Sci Rep. 2018; 8(1):14813. https://doi.org/10.1038/s41598-018-33147-x PMID: 30287848

27. Sasaki T, Okada I, Nakai Y. The evolution of conditional moral assessment in indirect reciprocity. Sci Rep. 2017; 7:1–8. https://doi.org/10.1038/s41598-016-0028-x

28. Okada I, Sasaki T, Nakai Y. Tolerant indirect reciprocity can boost social welfare through solidarity with unconditional cooperators in private monitoring. Sci Rep. 2017; 7(1):9737. https://doi.org/10.1038/s41598-017-09935-2 PMID: 28852005

29. Siegel JZ, Mathys C, Rutledge RB, Crockett MJ. Beliefs about bad people are volatile. Nat Hum Behav. 2018; 2(10):750–756. https://doi.org/10.1038/s41562-018-0425-1 PMID: 31406285

30. Yamamoto H, Okada I, Uchida S, Sasaki T. A norm knockout method on indirect reciprocity to reveal indispensable norms. Sci Rep. 2017; 7:44146. https://doi.org/10.1038/srep44146 PMID: 28276485

31. Uchida S, Yamamoto H, Okada I, Sasaki T. A Theoretical Approach to Norm Ecosystems: Two Adaptive Architectures of Indirect Reciprocity Show Different Paths to the Evolution of Cooperation. Front Phys. 2018; 6:14. https://doi.org/10.3389/fphy.2018.00014

32. Hao L, Naiman D. Quantile regression. Sage, London; 2007;.

33. Ohtsuki H, Iwasa Y. How should we define goodness? -reputation dynamics in indirect reciprocity. J Theor Biol. 2004; 231(1):107–120. https://doi.org/10.1016/j.jtbi.2004.06.005 PMID: 15363933

34. Santos FP, Santos FC, Pacheco JM. Social norm complexity and past reputations in the evolution of cooperation. Nature. 2018; 555(7695):242–245. https://doi.org/10.1038/nature25763 PMID: 29516999

35. Kato-Shimizu M, Onishi K, Kanazawa T, Hinobayashi T. Preschool children's behavioral tendency toward social indirect reciprocity. PLOS ONE. 2013; 8(8). https://doi.org/10.1371/journal.pone.0070915 PMID: 23951040

36. Shiraki Y, Igarashi T. We Can't Return Evil for Good: The Comparison between Direct and Indirect Reciprocity. Letters on Evolutionary Behavioral Science. 2017; 8(1):4–7. https://doi.org/10.5178/lebs.2017.54