Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 243 (2006) 418–436
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity

Nobuyuki Takahashi*, Rie Mashima

Department of Behavioral Science, Graduate School of Letters, Hokkaido University, Sapporo 060-0810, Japan

Received 5 April 2006; received in revised form 11 May 2006; accepted 11 May 2006
Available online 25 May 2006

Abstract

Indirect reciprocity is one mechanism that allows for unilateral resource giving among *n*-persons. Using analytical methods and computer simulations, previous studies have examined a number of strategies that make indirect reciprocity possible. In particular, previous investigations have concentrated on whether differentiating between justified and unjustified not-giving is important. However, whether or not a given strategy is ESS depends on the type of perceptual errors that are assumed. When errors are objective, regarding those who do not give to "bad" as "good" is critical. When perceptual errors are subjective, however, regarding those who give to "bad" as "bad" is critical. Since we believe that there is no guarantee that perceptual errors are shared among all individuals in a society, we argue that the latter moral principle may play a more important role in human interactions.

Keywords: Cooperation; Reputation; Indirect reciprocity; Subjective perceptual error; Unjustified giving

1. Introduction

Resource exchange is ubiquitous in all human societies. Without resource exchange, no society can hold. Among the various theories of exchange, only two have been established empirically. One is kin altruism (Hamilton, 1964), which explains unilateral giving towards kin who share the same genes, and the other is reciprocal altruism (Trivers, 1971; Axelrod, 1984), which explains resource giving in dyads over repeated interactions. Although these two explanations may be applied to certain animal species, there is yet another type of exchange that takes place only in human societies. That is, unilateral resource giving among *n*-persons. Social scientists call this type of exchange generalized-exchange, and biologists call this indirect reciprocity. Although this phenomenon has been known for decades, the existence of such exchanges has been puzzling. Recently, however, significant advances have been made in research in mathematical biology (e.g., Nowak and Sigmund, 1998a, b; Leimar and Hammerstein, 2001; Ohtsuki, 2004; Ohtsuki and Iwasa, 2004, 2006; Panchanathan and Boyd, 2003). Using computer simulations and mathematical analysis, this study extends and refines findings of these pioneering studies, and proposes that the requirements for the maintenance of indirect reciprocity are stricter than the previously suggested. The critical rule turns out to be: "the friend of my enemy is my enemy as well".

2. Previous solutions to the problem of resource giving

In the 60s, kin selection was proposed (Hamilton, 1964). Reciprocal altruism, which extends the scope of resource giving behavior to unrelated individuals, was proposed in the 70s (Trivers, 1971) and continued to be the main focus of research throughout the 80s (e.g., Axelrod, 1984). In the 1990s, theoretical and empirical research on unilateral giving outside of kin relations and without repeated interactions began. In order for such giving behavior to be adaptive, a system of indirect reciprocity is necessary

*Corresponding author. Tel.: +81 11 706 4153; fax: +81 11 706 3066.
E-mail address: ntakahas@let.hokudai.ac.jp (N. Takahashi).

0022-5193/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2006.05.014

(Alexander, 1987). Among these studies in the 90s, Nowak and Sigmund (1998a, b) played the most important role for subsequent research.

Nowak and Sigmund (1998a, b) set a general framework for research on indirect reciprocity called the "giving game". The existence of a population of individuals is assumed. On each round, each individual who is in the role of a donor is matched with a randomly selected recipient. An individual who plays the role of a donor decides whether or not to give to a recipient based on the recipient's reputation score, s, which he/she assigned to the recipient.[1] The parameter s reflects the recipient's past behavior(s). If the donor decides to give to the recipient, it will cost value c to himself. Consequently, the recipient receives a benefit of value b (with $b > c$). If the donor decides not to give to the recipient, both individuals receive zero.

Using evolutionary computer simulations and a mathematical analysis, Nowak and Sigmund (1998a, b) argued that the image scoring strategy (IS) can be evolutionarily stable and can, therefore, lead to the emergence of indirect reciprocity. The individuals who adopt IS define the score of the other individuals as follows. Initially (i.e. at the beginning of each generation), the individuals assume that everyone is "good". Afterwards, they assign "good" only to those who gave to another individual in the previous round, and assign "bad" to those who did not give to another individual. Based on this assignment rule, IS gives to "good" and does not give to "bad". Thus, IS can be considered a variation of the tit-for-tat (TFT) strategy. It gives only to previous givers. Nowak and Sigmund (1998a, b) argued that IS can maintain the system of indirect reciprocity because it does not give to previous nongivers.

Although its simplicity is attractive, later studies have revealed that IS is not evolutionarily stable (e.g., Leimar and Hammerstein, 2001; Panchanathan and Boyd, 2003). The reason is rather simple; individuals who adopt IS sometimes hurt one another. Individuals employing IS only give to others who gave to another individual in the previous round. However, not-giving to an individual who did not give to another individual in the previous round earns a reputation of "bad" from other IS players, which leads to further not-giving. On the contrary, ALLC players who give to another individual all of the time are never perceived "bad". Consequently, the expected payoff of IS is less than that of ALLC. As a result, the proportion of ALLC increases until, eventually, the population is susceptible to invasion by ALLD. At which point generalized exchange collapses entirely.

In order to overcome this weakness, Leimar and Hammerstein (2001) and Panchanathan and Boyd (2003) proposed the standing strategy (STAND).[2] STAND gives to "good" and does not give to "bad". In this sense, STAND is identical to IS. What is different is how to assign the score of the others. Following Sugden (1986)'s notion, an individual who employs STAND defines the score of the other individuals as follows. Suppose player A employs STAND. If player B gave to a recipient, player A assigns B "good" just like IS would. However, if player B did not give to a recipient, A's assignment depends on the reputation of B's recipient. If B did not give to a recipient whose score was "good," A considers B's behavior unjustified and assigns B "bad". However, if B did not give to a recipient whose score was "bad," A considers B's behavior justified and continues to assign B "good". Since STAND distinguishes justifiable from unjustifiable not-giving and, unlike IS, assigns "good" to the former, STAND players are not considered "bad" from the viewpoint of other STAND players when they punish ALLD players. Therefore, an individual who employs STAND does not lose opportunities to receive benefits from other individuals who also employ STAND. This feature makes STAND sustainable and makes indirect reciprocity possible.

Although these studies made significant contributions, none of them systematically examined all possible strategy combinations. At the time the IS strategy was proposed, things were quite simple. Since only first-order information (whether the current recipient gave last time) was considered, there could be only four strategies. However, things became more complicated once second-order information (the score of the recipient with whom the current recipient previously met) was introduced with the STAND strategy. Taking second-order information into consideration, there are a total of $2^4 = 16$ possible strategies.[3] Using evolutionary computer simulations, Takahashi and Mashima (2003) examined all 16 strategies and argued that the "strict discriminator" strategy (SDISC), represented as GBBB in Table 1, is the solution that makes generalized exchange possible. As IS and STAND do, a player who employs SDISC gives to "good" and does not give to "bad". Again, the difference lies in score assignment. A player who employs SDISC assigns "good" only to those who gave to a "good" recipient. There is a distinct contract between STAND and SDISC. STAND distinguishes justifiable from unjustifiable not-giving, while SDISC distinguishes justifiable from unjustifiable giving. Therefore, while STAND regards giving to "bad" as equal to giving to "good," SDISC punishes those who give indiscriminately.[4]

[1] In some studies, a donor determines his behavior based on both his own reputation score and his recipient's reputation score. However, we decided to omit this aspect in this paper because it is simpler and does not require high levels of cognitive capacity, and because Panchanathan and Boyd (2003) did not find any qualitative difference between RDISC which only attends to its recipient's score and CTFT which attends to its own score as well as its recipient's score.

[2] STAND corresponds to RDISC in Panchanathan and Boyd (2003).

[3] See Table 1.

[4] Precisely speaking, the standing strategy is not equal to GGBG in Table 1. GGBG regards not-giving to "bad" as "good" regardless of a donor's reputation, while the standing strategy regards not-giving to "bad" as "good" only when a donor is "good". However, in this paper we do not distinguish between these two strategies because Takahashi and Mashima (2003) did not find any significant difference between the two.

Table 1
Four genes that assign the score to potential recipients

		Second-order information: Current recipient's previous recipient's score	
		"Good"	"Bad"
First-order information: Current recipient's behavior toward the previous recipient	Gave	Gene 1: good or bad	Gene 2: good or bad
	Did not give	Gene 3: good or bad	Gene 4: good or bad

Strategies are represented by the sets of four genes. ALLC is represented as "good" "good" "good" "good", or GGGG. ALLD is represented as BBBB, IS is represented as GGBB, STAND is represented as GGBG, SDISC is represented as GBBB, and ES is represented as GBBG.

In another study that systematically examined possible strategy combinations, Ohtsuki and Iwasa (2004) treated reputation dynamics and behavioral strategies separately.[5] For reputation dynamics that assign a reputation for each individual, they considered three components: a donor's reputation ("good" or "bad"), a recipient's reputation ("good" or "bad"), and a donor's behavior (give or not to give). Thus, there are eight different situations, and there are $2^8 = 256$ possible reputation combinations. For behavioral strategy, they assumed that each individual determines his behavior by considering the reputations of both self and the recipient. Therefore, there are $2^4 = 16$ possible behavioral strategies.

Ohtsuki and Iwasa (2004) mathematically examined all combinations of reputation dynamics and behavioral strategies and showed that there are only eight combinations that are evolutionarily stable and that can achieve a high level of giving at the equilibrium. They called these combinations the "*leading eight*". They concluded that the notion of goodness should include three criteria: (1) giving to "good" individuals should be "good," (2) not-giving to "good" ones should be "bad," and (3) not-giving to "bad" ones should be "good". Thus, these three criteria are more consistent with STAND than with SDISC. In fact, SDISC was not included in the leading eight, and many strategies in the leading eight failed to maintain indirect reciprocity in Takahashi and Mashima (2003)'s simulation.

3. What made the difference?

Why did Takahashi and Mashima (2003) and Ohtsuki and Iwasa (2004) get different results? The most important factor that produced this discrepancy is, we believe, that these two studies conceptualized perceptual errors differently. Following Leimar and Hammerstein (2001), when an individual misperceives an action performed by another, it is called errors in perception. These errors can lead to misperception of the reputation score of others.[6] Takahashi and Mashima (2003) calculated the probability of misperception for each player independently, thereby making disparate opinions of the same individual possible, some regarding the individual as "good" and some as "bad", even if all players adopt the same strategy. In this sense, errors in perception are subjective. Conversely, Ohtsuki and Iwasa (2004, 2006) considered perceptual errors to be objective. In this case, the probability of misperception is calculated once for all players in the population. Either everyone misperceives a player's behavior, or everyone perceives it accurately. In other words, the notion that perceptual errors are objective requires consensus in perception among all players. Conversely, perfect consensus is highly unlikely when perceptual errors are subjective, particularly for large groups.

Whether or not a mistake is shared among all players could have a serious implication when we consider strategies that assign "good" to players who gave to "bad". For instance, consider the case of subjective errors with a STAND strategy. Suppose, at one point, player X, who adopts STAND and is regarded as "good" by other players, misperceives player A's behavior and assigns "bad" while the other players, who also adopt STAND, perceive A's behavior correctly and assign him "good". Subsequently, if X is matched with A, X will not give to A since X believes A to be "bad". Consequently, other players who adopt STAND will now assign "bad" to X since they regard A as "good". This cycle of misperception can persist, and individuals adopting STAND will hurt one another in the process. However, this cycle does not matter to ALLC players because perceptual errors are irrelevant to them. Therefore, when perceptual errors occur, the probability that ALLC players will be considered "good" by STAND players is higher than that for fellow STAND players. Consequently, ALLC increases until eventually ALLD takes over the population. However, such a cycle does not occur when perceptual errors are objective since each individual's score is shared among all players. Under this scenario, even if X does not give to A, the behavior is considered justifiable not-giving by all other STAND players. Hence ALLC players never outperform STAND players. This is the reason why the leading eight strategies could maintain generalized exchange in Ohtsuki and Iwasa (2004).

Conversely, because SDISC regards ALLC players as "bad" by definition, the expected payoff advantage

[5] Brandt and Sigmund (2004) called them assessment modules and action modules.

[6] Please note that errors in perception do not mean misperceiving the reputation score of others per se.

enjoyed by SDISC players over ALLC players remains sufficiently large that subjectivity in perceptual errors should not influence the evolutionary dynamics in a population with SDISC.[7]

If the above argument is correct, we need to reconsider how to define goodness. After Nowak and Sigmund (1998a, b), most research has agreed that STAND (or similar strategies) allows for indirect reciprocity. As Ohtsuki and Iwasa (2004) stated, the characteristics of these strategies are (1) giving to good individuals should be regarded as good, (2) not-giving to good individuals should be regarded as bad, and (3) not-giving to bad individuals should be regarded as good. Only Takahashi and Mashima (2003) disagreed with these conclusions, arguing instead that the critical characteristic sustaining indirect reciprocity is to regard giving to "bad" individuals as "bad".

The second argument, that giving to "bad" is "bad", runs parallel to the position popularly held in the social dilemma literature. In an n-person prisoner's dilemma situation, players who punish free-riders alter the incentive structure so that free-riding is no longer the dominant strategy. However, in the event that punishment is costly, other problems may emerge. For instance, what incentive is there for individual players to absorb a significant cost in order to punish free-riding? It is better to encourage other players to penalize free-riders. This second-order free-rider problem must be solved before we can address the problem of the first-order social dilemma. Otherwise, there are an infinite number of problems to consider (i.e. third-order free-rider problem, fourth-order free-rider problem, and so on). One solution is for those who punish first-order free-riders to also penalize second-order free riders (Yamagishi and Takahashi, 1994). In other words, those who do not punish free-riders should be penalized even if they cooperated in the original social dilemma. Thus, failure to punish should be regarded as "bad". If we modify the well-known maxim, "the enemy of my enemy is my friend," to fit this principle, it would be "the friend of my enemy is my enemy".

Generalized exchange also possesses a second-order dilemma. Of course, in the giving game, punishment entails not-giving, which gives an advantage to the punisher. Therefore, at first glance, there seems to be no second-order problem. However, for some strategies, not-giving may be perceived as "bad" by others pursuing the same strategy and may, therefore, impose an indirect cost on the punisher. Thus, sophisticated and tolerant strategies, such as STAND, are especially prone to subjective perceptual errors because their advantage over ALLC might be too small to withstand any decrease. Consequently, a less tolerant strategy, such as SDISC, is required to outperform ALLC in the presence of subjective perceptual errors.

The above argument is still a speculation. Takahashi and Mashima (2003)'s results were preliminary. In the next three sections, we will develop a more thorough analysis using analytical methods and computer simulations to examine whether the principle of regarding giving to "bad" as "bad" is a necessary condition for the maintenance of indirect reciprocity when perceptual errors are subjective. We will first revisit Takahashi and Mashima (2003)'s simulation, which proposed two candidate strategies that may allow for indirect reciprocity. Then, following Panchanathan and Boyd (2003), we will evaluate the evolutionary stability of these two strategies under the assumption that only one-way implementation errors can occur. Finally, since analytical solutions are not possible, we will combine new simulations with an ESS analysis to examine these two candidates under the assumption that both two-way implementation errors and two-way subjective perceptual errors can occur.

4. ESS analysis

4.1. Revisiting Takahashi and Mashima (2003)'s simulation

Although Takahashi and Mashima (2003) presented SDISC as the one (and only) strategy that makes indirect reciprocity possible, their emphasis was on whether indirect reciprocity can emerge in a population that is initially composed of equal numbers of three strategies: ALLC, ALLD, and the focal strategy. However, a given strategy is only stable if it can resist invasion by other strategies when the population is initially composed of the focal strategy only. Of course this is not a formal mathematical analysis but, qualitatively, this method can yield the same outcome. We will use a more formal mathematical analysis in subsequent sections.

Let us first briefly describe Takahashi and Mashima (2003)'s simulation design. On each round, two individuals were chosen randomly. One individual was assigned the role of giver and the other was assigned the role of recipient. Givers give to recipients if and only if the giver believes that the recipient's reputation score is "good". Otherwise, the giver does not give to the recipient. If the giver gives, he pays cost c and the recipient receives benefit b. At the beginning of each generation, individuals are assumed to have given to a "good" individual on the 0th (previous) round. Each generation consists of m rounds. At the end of a generation, cumulative profit is used to determine the relative fitness of each individual in the population. Standard replicator dynamics were used to determine the members of the next generation. Thus, the more successful an individual was in the previous generation, the more offspring he/she reproduces for the subsequent generation. Mutations occur between generations with probability μ for each of the four genes listed in Table 1; in the case of mutation, new values are determined randomly. Takahashi

[7]It is true that SDISC was not included in the leading eight in Ohtsuki and Iwasa (2004). However, this is not because SDISC performs worse than ALLC or ALLD. Actually, SDISC can make up an equilibrium, but the average giving rate among the whole population in that equilibrium is not high enough because the number of other players whom SDISC regard as "good" decreases over time.

and Mashima (2003) used the standard model for this type of simulation, in which there is only one group of players in the population (i.e. population and group are interchangeable). Stochastic universal sampling, developed by Baker (1987), was used to minimize effects associated with genetic drift. Grefenstette (2000) has suggested that sampling in this manner should reduce the effects associated with random fluctuations of genes.

In addition, the simulation could accommodate two types of errors. There was a probability α for a potential giver to perform an action different from the one prescribed by the strategy (errors in strategy execution), and there was a probability δ for an individual to misperceive an action performed by another individual (errors in perception). Perceptual errors were subjective in that they were calculated independently for each player. In other words, perceptual errors occurred for only a fraction of the group. Thus, it was possible for players adopting the same strategy to generate different reputation scores for the same recipient.

Since Takahashi and Mashima (2003) reported only a portion of their original investigation, results for all 14 possible strategies (excluding ALLC and ALLD) will be presented here. Beginning with a homogeneous population (focal strategy only), we examined whether each of the 14 strategies could resist invasion by ALLC or ALLD. If a strategy can resist invasion and maintain a high level of giving, then we can conclude that the strategy has the characteristics of an ESS. The parameters were group size ($=300$), the number of generations ($=10{,}000$), and number of rounds per generation ($m=1500$). The cost of giving (c) was always set to 1, but the benefit (b) was variable ($=2, 4, 6, 8, 10$). The rate of implementation errors was 0.025, and the rate of subjective perceptual errors was 0.025. The mutation rate (μ) was 0.0001.[8]

The results of 30 replications are shown in Table 2. It is clear that there are two strategies that can maintain indirect reciprocity: GBBG and GBBB. Please note that GGBG (STAND) sometimes failed to maintain indirect reciprocity. GBBB is SDISC, as we discussed earlier. GBBG is a new strategy that regards giving to "good" and not-giving to "bad" as "good", and regards giving to "bad" and not-giving to "good" as "bad". Let us call GBBG "extra standing" (ES). As IS, STAND, and SDISC do, a player who employs ES gives to "good" and does not give to "bad". Again, the difference lies in score assignment. The difference between GBBG (ES) and GGBG (STAND) is that ES regards giving to "bad" as "bad" while STAND regards it as "good". This strategy was included in the leading eight by Ohtsuki and Iwasa (2004), but was not reported in Takahashi and Mashima (2003) because it did not always achieve indirect reciprocity when the initial composition of the population was determined randomly. Therefore, we can infer that GBBG is less robust when compared to GBBB since the former requires greater dominance initially in order for indirect reciprocity to emerge.

Nevertheless, we will next examine the evolutionary dynamics of both these candidates: GBBB and GBBG.

4.2. ESS analysis

In this section, we will follow Panchanathan and Boyd's (2003) method to perform ESS analyses on SDISC and ES. We assume that agents interact in an infinite, unstructured population. All agents begin with a reputation score of good. In the first round of social interaction, each agent acts as a potential donor to a randomly selected potential recipient. If a donor gives to a recipient, the donor's fitness decreases by c while the recipient's fitness increases by b. It is assumed that $b>c>0$. Subsequent rounds of social interaction occur with probability w ($0\leqslant w<1$). We assume that only one-way implementation errors can occur, and the parameter α denotes the probability that a donor does not give when he/she should, according to the strategy he/she adopts. Please note that perceptual errors, analysed in the next section, are not considered here.

4.2.1. SDISC

We consider three strategies: ALLC (always gives), ALLD (never gives), and SDISC. The frequencies of these strategies are denoted by x_1, x_2, and x_3, respectively. SDISC gives only when matched with recipients whose score is "good". SDISC regards a recipient as "good" only when the recipient gave to a "good" person in the previous round.

Results of the analysis indicated that both SDISC and ALLD were ESS, whereas ALLC was not (Fig. 1).[9] Qualitatively, the performance of SDISC is similar to that of STAND reported in Panchanathan and Boyd (2003). Along the entire ALLC–SDISC edge, selection favors the SDISC strategy. This is because SDISC does not give to ALLC in the event that the latter gave to ALLD in the previous round. Compared to STAND, however, the domain of attraction for SDISC is smaller. Along the ALLD-SDISC edge, there is an unstable equilibrium point given by the equation below.

$$x_3=\frac{c}{bw(1-\alpha)}. \tag{1}$$

If the frequency of SDISC is above this threshold (1), then SDISC increases and drives out ALLD. If, instead, the frequency of SDISC is below this threshold (1), then ALLD increases and dominates the population.

[8]We could show results of the simulation without the possibility of mutation. This may be more appropriate in a pure theoretical sense as a reviewer suggested. However, we will show results with the possibility of mutation because this type of simulation was the one used in Takahashi and Mashima (2003), and because we would like to examine the evolutionary dynamics where all three strategies can be born by mutation to see if each strategy can maintain high levels of generalized exchange.

[9]The derivation of fitness functions for each strategy and the subsequent evolutionary dynamics analyses are presented in the appendix.

Table 2
Results of simulation examining whether or not the focal strategy can maintain high levels of generalized exchange despite the possibility of invasion by ALLC and ALLD

Focal strategy	b	Giving rate	Proportion of focal strategy	Proportion of ALLC	Proportion of ALLD	Focal strategy	b	Giving rate	Proportion of focal strategy	Proportion of ALLC	Proportion of ALLD
GGGB	2	0.02	0.00	0.00	1.00	GBBG	2	0.87	1.00	0.00	0.00
	4	0.03	0.00	0.00	1.00	(ES)	4	0.88	1.00	0.00	0.00
	6	0.03	0.00	0.00	1.00		6	0.88	1.00	0.00	0.00
	8	0.03	0.00	0.00	1.00		8	0.88	1.00	0.00	0.00
	10	0.02	0.00	0.00	1.00		10	0.88	0.99	0.01	0.00
GGBG	2	0.93	0.99	0.01	0.00	BGBG	2	0.02	0.00	0.00	1.00
(STAND)	4	0.93	0.92	0.08	0.00		4	0.03	0.00	0.00	1.00
	6	0.79	0.55	0.29	0.17		6	0.03	0.00	0.00	1.00
	8	0.33	0.16	0.17	0.67		8	0.03	0.00	0.00	1.00
	10	0.06	0.02	0.02	0.97		10	0.03	0.00	0.00	1.00
GBGG	2	0.03	0.00	0.00	1.00	BBGG	2	0.02	0.00	0.00	1.00
	4	0.02	0.00	0.00	1.00		4	0.02	0.00	0.00	1.00
	6	0.02	0.00	0.00	1.00		6	0.02	0.00	0.00	1.00
	8	0.02	0.00	0.00	1.00		8	0.03	0.00	0.00	1.00
	10	0.02	0.00	0.00	1.00		10	0.03	0.00	0.00	1.00
BGGG	2	0.03	0.00	0.00	1.00	GBBB	2	0.80	1.00	0.00	0.00
	4	0.02	0.00	0.00	1.00	(SDISC)	4	0.81	1.00	0.01	0.00
	6	0.03	0.00	0.00	1.00		6	0.80	0.99	0.01	0.00
	8	0.03	0.00	0.00	1.00		8	0.83	0.91	0.09	0.00
	10	0.02	0.00	0.00	1.00		10	0.87	0.52	0.44	0.03
GGBB	2	0.29	0.27	0.03	0.70	BGBB	2	0.02	0.00	0.00	1.00
(IS)	4	0.02	0.00	0.00	1.00		4	0.03	0.00	0.00	1.00
	6	0.03	0.00	0.00	1.00		6	0.02	0.00	0.00	1.00
	8	0.02	0.00	0.00	1.00		8	0.03	0.01	0.00	0.99
	10	0.02	0.00	0.00	1.00		10	0.03	0.00	0.00	1.00
GBGB	2	0.03	0.00	0.00	1.00	BBGB	2	0.02	0.59	0.00	0.41
	4	0.03	0.00	0.00	1.00		4	0.03	0.44	0.00	0.56
	6	0.02	0.00	0.00	1.00		6	0.03	0.72	0.00	0.28
	8	0.02	0.00	0.00	1.00		8	0.03	0.61	0.00	0.39
	10	0.02	0.00	0.00	1.00		10	0.02	0.60	0.00	0.40
BGGB	2	0.03	0.00	0.00	1.00	BBBG	2	0.02	0.00	0.00	1.00
	4	0.03	0.00	0.00	1.00		4	0.03	0.00	0.00	1.00
	6	0.03	0.00	0.00	1.00		6	0.02	0.00	0.00	1.00
	8	0.03	0.00	0.00	1.00		8	0.02	0.00	0.00	1.00
	10	0.03	0.00	0.00	1.00		10	0.03	0.00	0.00	1.00

The parameters are group size (= 300) and the number of generations (= 10,000). The other parameters are as follows: $m = 1500$, $c = 1$, $\alpha = 0.025$, $\delta = 0.025$, $\mu = 0.0001$. The averages of 30 replications are shown in each row of Table 2.

4.2.2. *Extra standing (ES)*

As in the previous section, we consider three strategies: ALLC, ALLD, and ES. The frequencies of these strategies are denoted by x_1, x_2, and x_4, respectively. ES gives only when matched with recipients whose score is "good". ES regards a recipient as "good" when the recipient gave to a "good" person or did not give to a "bad" person in the previous round.

We found that both ES and ALLD were ESS, whereas, again, ALLC was not (Fig. 2). Qualitatively, the performance of ES is similar to that of STAND and SDISC. Along the entire ALLC-ES edge, selection favors the ES strategy. This is because ES does not give to ALLC in the event that the latter gave to "bad" in the previous round. In terms of the domain of attraction for ES, it is larger than that for SDISC but smaller than that for STAND. Along the ALLD-ES edge, there is an unstable equilibrium point given by the equation below. However, this equation cannot be easily solved for x_4. We show the approximate solution for the unstable equilibrium,

$$x_4 = \frac{c}{bw} + \left\{\frac{c(b-c+bw)(c-3cw+2bw^2)}{bw(b-2c+bw)(2bw-c)}\right\}\alpha. \tag{2}$$

If the frequency of ES is greater than threshold (2), then ES increases and drives out ALLD. If, instead, the

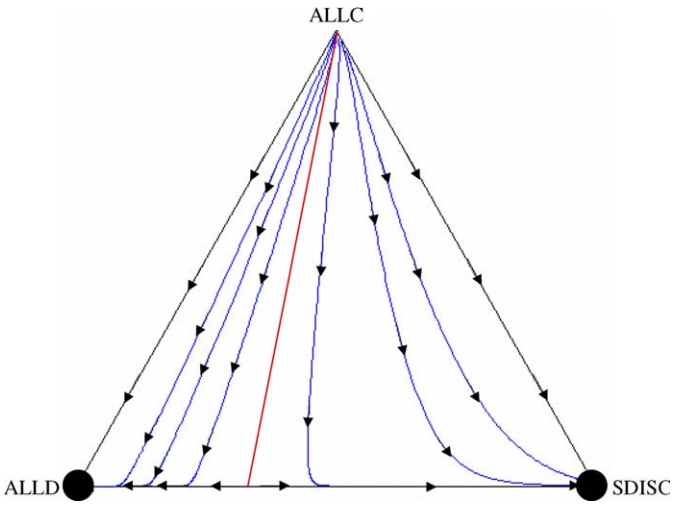

Fig. 1. Evolutionary dynamics of generalized exchange with the SDISC strategy. ALLD and SDISC are both an ESS. Model parameters for this figure are as follows: $b = 0.01$, $c = 0.003$, $w = 0.95$, $\alpha = 0.05$. We used the same parameters as Panchanathan and Boyd (2003). The thick line represents the neutral line separating the phase space into two regions. If the composition of agents resides in the right region, the SDISC equilibrium will be realized. If, however, the composition resides in the left region, the ALLD equilibrium will be realized. Given the above parameter settings, the frequency of SDISC (x_3) along this line is approximately equal to 33.241% (from Eq. (1)).

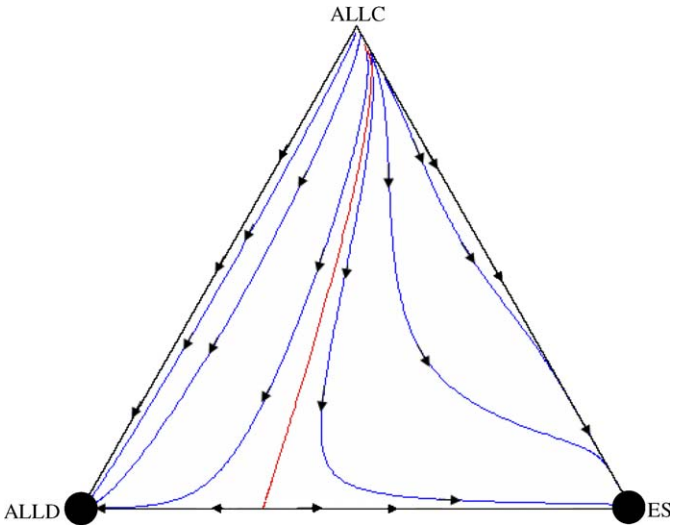

Fig. 2. Evolutionary dynamics of generalized exchange with the ES strategy. ALLD and ES are both an ESS. Model parameters for this figure are as follows: $b = 0.01$, $c = 0.003$, $w = 0.95$, $\alpha = 0.05$. The thick line represents the neutral line separating the phase space into two regions. If the composition of agents resides in the right region, the ES equilibrium will be realized. If, however, the composition resides in the left region, the ALLD equilibrium will be realized. Given the above parameter settings, the frequency of ES (x_4) along this line is approximately equal to 33.09% (from Eq. (2)).

frequency of ES is below this threshold (2), then ALLD increases and dominates the population.

The results from Sections 4.2.1 and 4.2.2 indicate that the performance of SDISC and ES are similar to that of STAND. All three strategies are an ESS. However, the above mathematical analysis considered implementation errors only. The focus of this paper is to determine what will happen in the event of subjective perceptual errors. In the next section, we will try to answer this question. However, on our way to finding the answer, we found that using the same analytical techniques was not possible with the inclusion of subjective perceptual errors because it is too complicated. Thus, we supplemented the analytic techniques with a simulation.

5. Two-way implementation errors and two-way subjective perceptual errors

In this section, we examine the effect of perceptual errors (i.e., falsely perceiving giving as not-giving, and vice versa) on the emergence of indirect reciprocity. In the event that perception is subjective, the probability of making an error is calculated independently for each player in the population. Conversely, if perception is objective, the probability of attributing the wrong label is calculated once for all players. Thus, whereas objective perceptual errors are shared among all individuals adopting the same strategy, subjective perceptual errors often are not, and players will disagree as to what the appropriate behavior should be. Consequently, subjectivity in perception may cause future exchanges between players sharing the same strategy to be uncooperative. We agree with Ohtsuki and Iwasa (2004), who pointed out that the evolutionary outcome of a model depends on whether it employs subjective or objective perceptual errors in its design. Ohtsuki and Iwasa (2004) employed objective perceptual errors and their investigation resulted in the leading eight strategies. However, they did not evaluate these strategies under the condition that perceptual errors were subjective. The main purpose of this section was to determine whether STAND, SDISC and ES remain ESS when perceptual errors are subjective.

For both the SDISC and ES strategies, we were able to derive fitness functions using the methods presented in Appendices A and B, once the proportion of "good" players present in the population after the nth round was known (see Appendix C). However, it turns out that we could not derive general expressions to denote the proportion of "good" individuals among the population and for each strategy in each round. The simultaneous differential equations for the recursive process are too complicated, and each individual has his subjective opinions about the reputation scores of other players. To resolve this issue, we designed a simulation to estimate, given specific parameter values, the frequency of "good" individuals for each strategy in each round, which was then entered into the respective fitness function. This allowed us to calculate each strategy's fitness in each round, and its cumulative fitness after all rounds given specific parameter values. By combining this analytic method with a simulation, we were able to determine the evolutionary stability of each strategy given subjectivity in perceptual errors.

Although it is possible to examine the entire evolutionary dynamics in principle, it is impractical to construct representations such as those depicted in Figs. 1 and 2.

Let us explain the simulation in full detail. Please note that this simulation is not an evolutionary simulation. In each replication, a small number of individuals who employ a certain strategy attempted to invade a population employing a majority strategy.[10] In each round, players were matched with another randomly chosen individuals and decided whether or not to give. At the same time, we output the proportion of "good" individuals for each strategy from each individual's viewpoint. For example, one individual employing SDISC regards 990 other individuals out of 1000 as "good", while another individual employing the same SDISC regards 989 other individuals as "good". Then, we calculated the average proportion of "good" individuals for each strategy in each round. For example, in one round, if there were five individuals, SDISC1, SDISC2, SDISC3, SDISC4, and SDISC5, employing SDISC and regard 990, 991, 989, 992, 988 other individuals as "good", respectively, then the average proportion of "good" individuals for SDISC in this round is $990/1000 = 0.99$. There were 50 replications and we took the average for estimated gn (ALLC), gn (ALLD), and gn for the focal strategy (i.e. gn (STAND), gn (SDISC), or gn (ES)). The other parameters were group size (= 1000), number of rounds (= 50), cost ($c = 1$), benefit ($b = 2, 4, 6, 8, 10$), the rate of implementation error ($\alpha = 0.00, 0.01, 0.02, 0.03, 0.04, 0.05$), and the rate of subjective perceptual error ($\delta = 0.00, 0.01, 0.02, 0.03, 0.04, 0.05$). All together, there were $5 \times 6^2 = 180$ conditions for each ESS examination, with 50 observations for each condition.

Estimated gn (ALLC), gn (ALLD), and gn for the focal strategy (i.e. gn (STAND), gn (SDISC), or gn (ES)) were input into the fitness function for each round to calculate the total fitness for each strategy. Since we were interested in whether an invading strategy could outperform a majority strategy, the dependent variable was the fitness differential (= invading strategy's total fitness—majority strategy's total fitness). Therefore, the greater the value, the greater the success observed for the invading strategy.

The results for STAND (with $b = 4$) are shown in Figs. 3–6.[11] In Fig. 3, we can see that STAND is more likely to be invaded by ALLC as δ increases. Therefore, given the possibility of subjective perceptual errors, STAND is not an ESS. Fig. 4 indicates that STAND is robust to invasion by ALLD. Fig. 5 indicates that STAND can invade ALLC, and Fig. 6 indicates that STAND cannot invade ALLD. Therefore, ALLD is the only ESS in this system with ALLC, ALLD, and STAND as the alternatives.[12] This finding conflicts with those reported by Panchanathan and Boyd (2003) and Ohtsuki and Iwasa (2004), the primary difference being the subjectivity of perceptual errors.

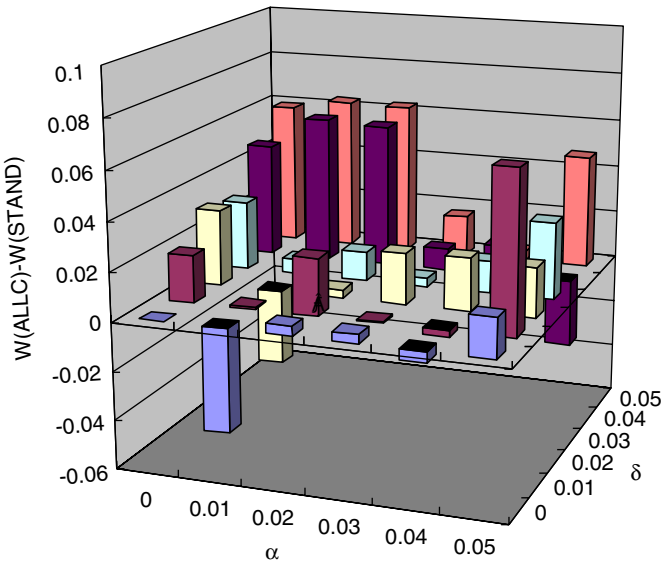

Fig. 3. Fitness differentials between the STAND strategy and ALLC when the benefit/cost ratio is 4. ALLC can invade the STAND population.

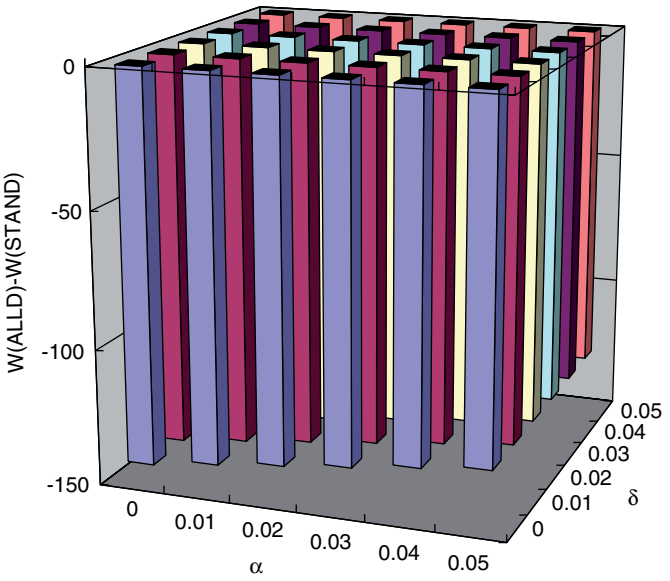

Fig. 4. Fitness differentials between the STAND strategy and ALLD when the benefit/cost ratio is 4. ALLD cannot invade the STAND population.

The results for STAND (with $b = 4$) are shown in Figs. 7 and 8 indicate that SDISC is an ESS when either ALLC or ALLD attempts to invade. Although SDISC can invade ALLC, it cannot invade ALLD (Figs. 9 and 10, respectively). Therefore, in the advent of subjective perceptual

[10]More precisely, 1% is the invading strategy, and 99% is the majority strategy.

[11]We present the results with $b = 4$ only. The results with $b = 2$, 6, 8, and 10 are available from the authors upon request. However, general patterns are similar.

[12]We omit to examine whether ALLD can invade ALLC and whether ALLC can invade ALLD since answers are too obvious.

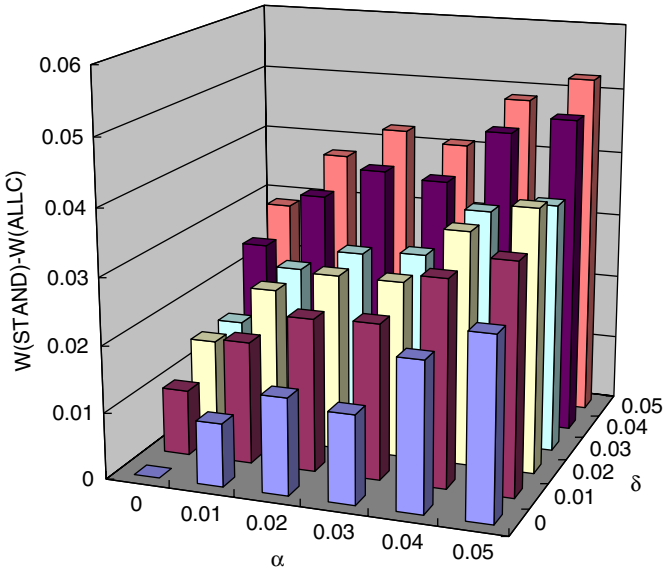

Fig. 5. Fitness differentials between the STAND strategy and ALLC when the benefit/cost ratio is 4. STAND can invade the ALLC population.

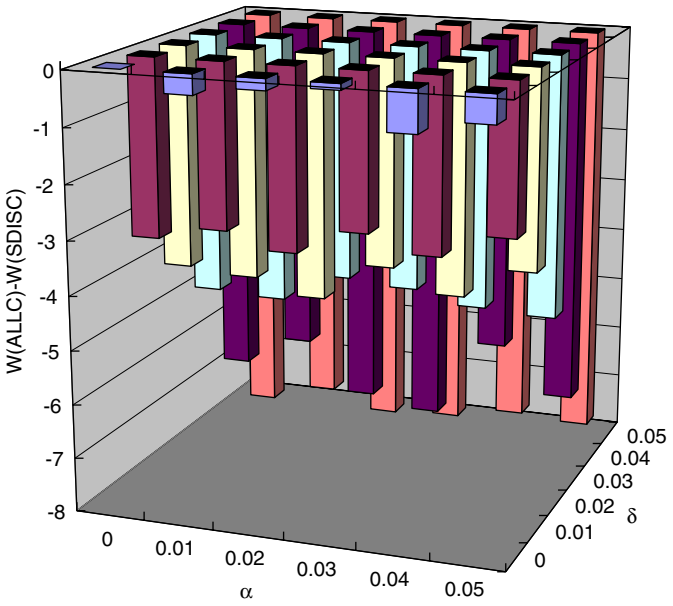

Fig. 7. Fitness differentials between the SDISC strategy and ALLC when the benefit/cost ratio is 4. ALLC cannot invade the SDISC population.

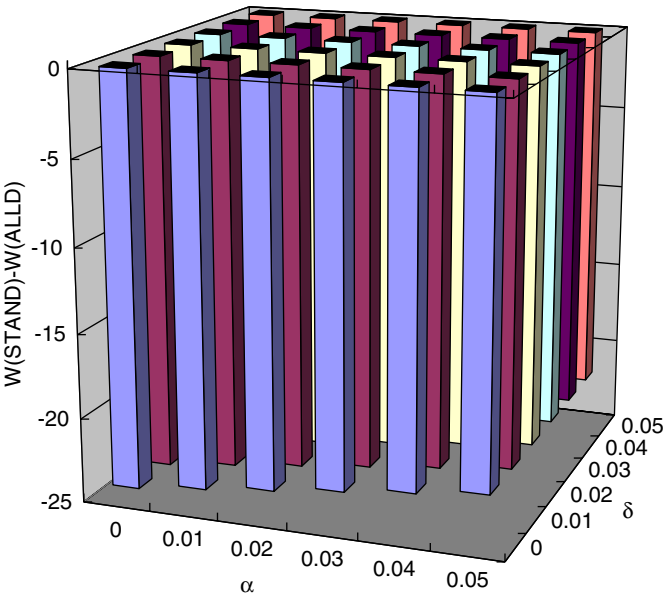

Fig. 6. Fitness differentials between the STAND strategy and ALLD when the benefit/cost ratio is 4. STAND cannot invade the ALLD population.

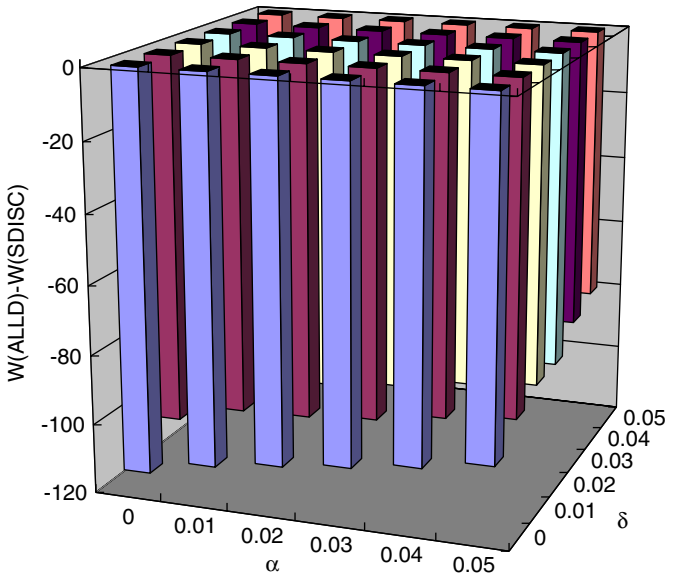

Fig. 8. Fitness differentials between the SDISC strategy and ALLD when the benefit/cost ratio is 4. ALLD cannot invade the SDISC population.

errors, SDISC remains an ESS as does ALLD. Thus, we can infer that evolutionary dynamics of indirect reciprocity with subjective perceptual errors is qualitatively similar to Fig. 1.

Results were almost identical for the Extra Standing strategy (Figs. 11–14). Fig. 11 indicates that ALLC cannot invade a polulation of ES, and Fig. 12 indicates that ALLD cannot invade a population of ES. Fig. 13 indicates that ES can invade ALLC. However, Fig. 14 indicates that ES cannot invade ALLD. Therefore, in the event that perceptional errors are subjective, both ES and ALLD remain ESS and again, we can infer that the evolutionary dynamics of indirect reciprocity with subjective perceptual errors is qualitatively similar to Fig. 2.

6. Discussion

What is the definition of goodness that makes indirect reciprocity possible among *n*-persons in generalized exchange settings? Ohtsuki and Iwasa (2004) proposed eight evolutionarily stable combinations of reputation dynamics and strategies that achieved a high level of giving in

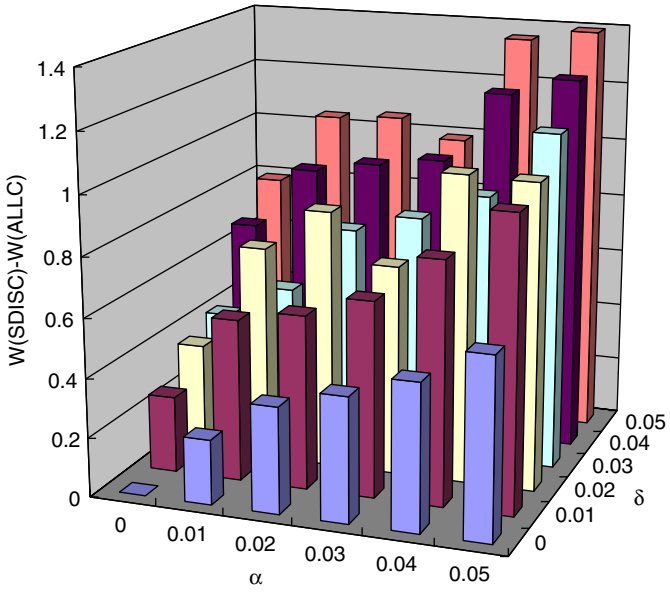

Fig. 9. Fitness differentials between the SDISC strategy and ALLC when the benefit/cost ratio is 4. SDISC can invade the ALLC population.

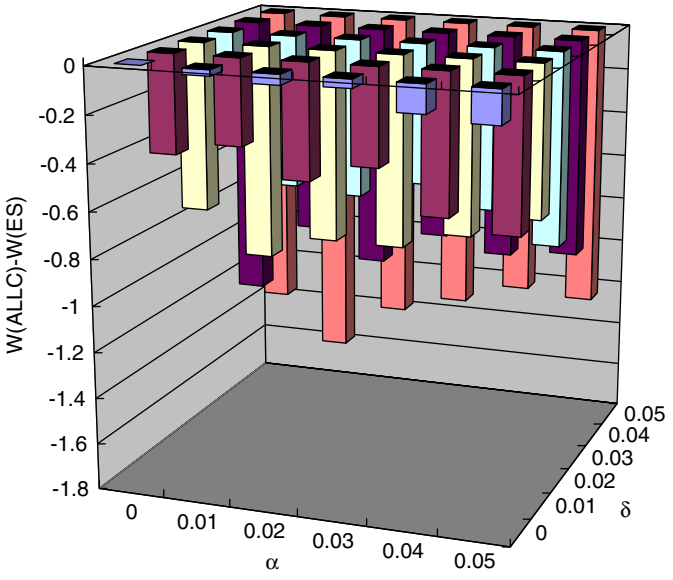

Fig. 11. Fitness differentials between the Extra Standing strategy and ALLC when the benefit/cost ratio is 4. ALLC cannot invade the ES population.

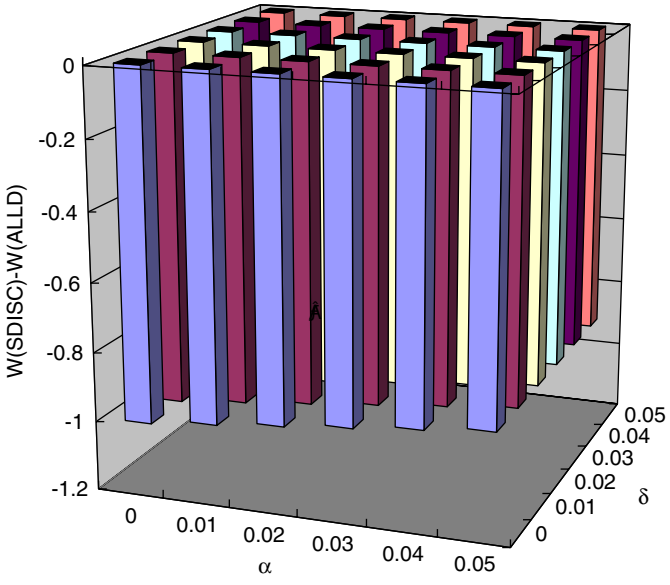

Fig. 10. Fitness differentials between the SDISC strategy and ALLD when the benefit/cost ratio is 4. SDISC cannot invade the ALLD population.

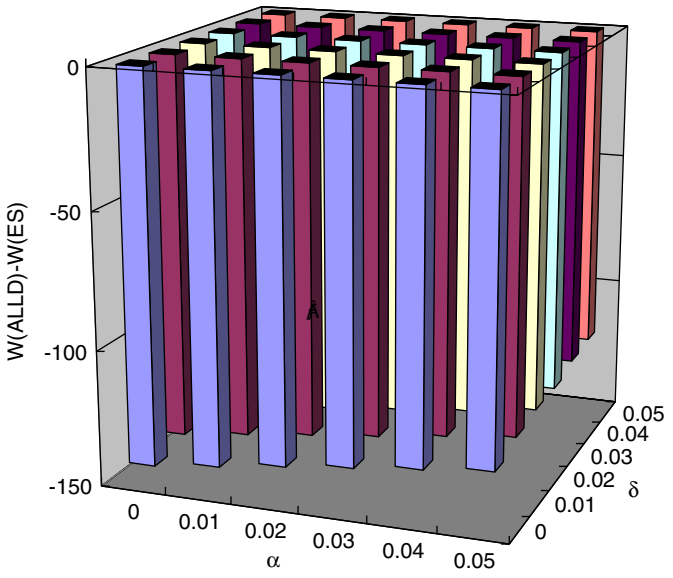

Fig. 12. Fitness differentials between the ES strategy and ALLD when the benefit/cost ratio is 4. ALLD cannot invade the ES population.

generalized exchange settings. The common characteristics of these "leading eight" were (a) giving to "good" persons should be regarded as "good", (b) not-giving to "good" persons should be regarded as "bad", and (c) not-giving to "bad" persons should be regarded as "good". However, this conclusion is limited to situations in which perceptual errors are assumed to be shared by everyone. When perceptual errors were assumed to be subjective, as they were in this study, at least one of those strategies (STAND) was not an ESS, and, moreover, two strategies not considered in the "leading eight" (ES and SDISC) were ESS. These two evolutionarily stable strategies share three common characteristics: (i) giving to "good" persons should be regarded as "good", (ii) not-giving to "good" persons should be regarded as "bad", and (iii) giving to "bad" persons should be regarded as "bad". The crucial point is the third characteristic, since it distinguishes SDISC and ES from STAND. Given these results, we offer a variation of the well-known maxim—"the enemy of my enemy is my friend", namely, "the friend of my enemy is my enemy".

Although we found two candidates, SDISC and ES, to solve the problem of indirect reciprocity, it is difficult to determine which strategy is better. As we discussed earlier,

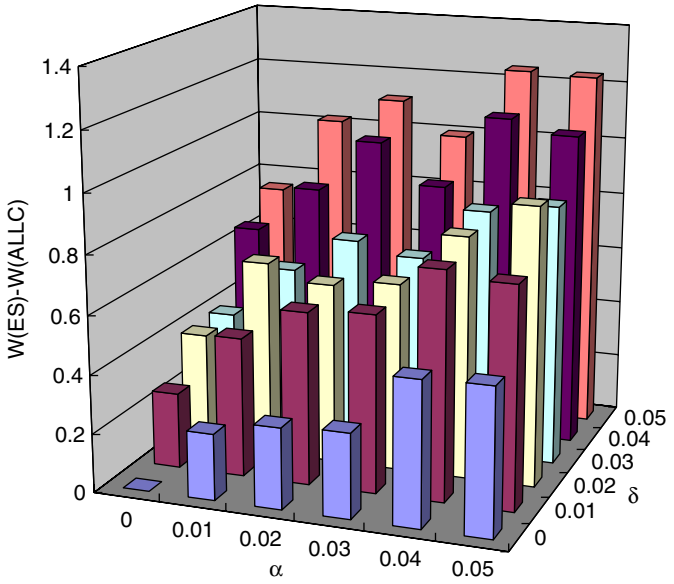

Fig. 13. Fitness differentials between the ES strategy and ALLC when the benefit/cost ratio is 4. ES can invade the ALLC population.

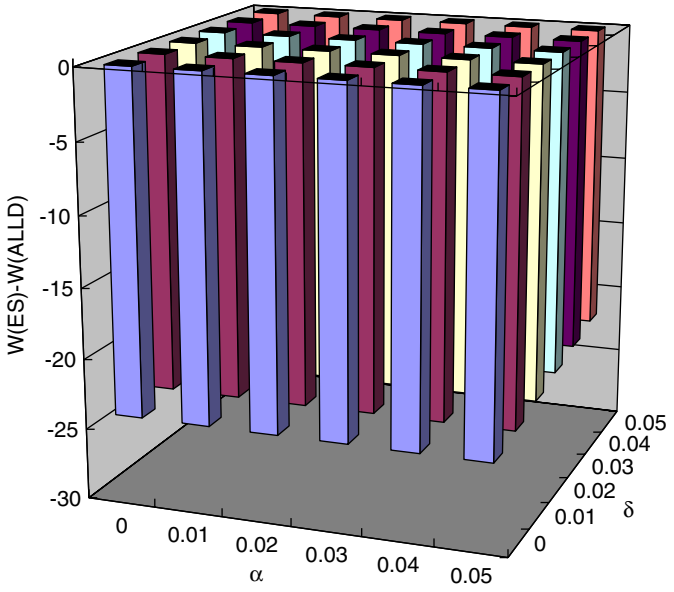

Fig. 14. Fitness differentials between the ES strategy and ALLD when the benefit/cost ratio is 4. ES cannot invade the ALLD population.

the domain of attraction for ES is larger than that for SDISC if perceptual errors are assumed to be objective. Currently, however, we cannot be sure that this remains true assuming subjectivity in perceptual error. Another point to consider is that ES was included in the original "leading eight" whereas SDISC was not. Finally, as Ohtsuki and Iwasa (2006) have pointed out, a society with ALLC, ALLD, and SDISC cannot attain a high level of generalized exchange if the number of rounds per generation is very large. This is because SDISC regards anyone matched with a "bad" individual as "bad", regardless of whether they give or do not give. Therefore, although at the beginning of each generation everyone is regarded as "good", the proportion of "good" individuals decreases over time, and eventually SDISC players do not give to anyone.[13] Conversely, a society with ALLC, ALLD, and ES can attain a high level of generalized exchange regardless of the number of rounds per generation because the proportion of "good" individuals does not decrease over time.[14,15] Considering these factors, we conclude that ES is the superior strategy. However, further research is needed to verify this inference.

Although we used virtually the same techniques as those in previous studies, we came to different conclusions regarding the evolutionary dynamics of generalized exchange. The main difference between our study and previous investigations concerned the assumption governing perceptual errors; whereas other studies have commonly assumed errors in perception to be objective (i.e. shared by everyone), we assumed that errors were subjective.[16] When perceptual errors are not shared, individuals may treat other individuals differently. Ohtsuki and Iwasa (2004) explained the difference between objective and subjective perceptual errors in terms of direct and indirect observation models. In a direct observation model, each individual observes every other individual's behavior. Therefore, only the individual who committed the perceptual error incurs the consequence of that error. In contrast, an indirect observation model stipulates that each individual learns of every other individual's behavior through the observations of a single player. In this case, everyone commits the same perceptual error because everyone receives the same inaccurate information, and the consequences of perceptual errors are shared among all individuals.

Although Ohtsuki and Iwasa (2004) conceded that the standing strategy was not an ESS in direct observation models, they failed to indicate which strategies were ESS. This study showed that both the strict discriminator strategy and the extra standing strategy are an ESS in direct observation models.

Nevertheless, Ohtsuki and Iwasa (2004)'s position has remained that the the solution to the problem of indirect reciprocity lays somewhere in the "leading eight" since, they have argued, indirect observation models are more suitable to describe human societies. However, we disagree. Indirect observation models assume that each piece of information flows uncorrupted from one individual to all other members of a group. If a TV news reporter makes a mistake during the filming of a news story, then every

[13]In Appendix A, we see this logic analytically in (A1).

[14]A simple calculation indicated that gn converges within a relatively small number of rounds. However, how soon it converges depends on the parameter values of x_1, x_2, x_4, and α.

[15]A closer look at the results of Section 5 indicated that strictly speaking the above logic holds only when there is no possibility of subjective percepttual errors. If this possibility exists, gn also continues to decrease in cases of STAND and ES. However, the rate of decrease is minimal. In the case of SDISC, the rate of decrease is significantly large. For example, when $\alpha = 0.05$ and $\delta = 0.05$, g50 for STAND is 0.9984 and g50 for ES is 0.944, while g50 for SDISC is 0.46.

[16]Brandt and Sigmund (2004) also assumed subjective perceptual errors.

viewer receives the same erroneous information. However, this example may be unusual for human societies, in which information often travels from person to person, collecting errors along the way and degrading over time, and where information is often encumbered with ambiguities that require interpretation.[17]

We agree with Ohtsuki and Iwasa (2004)'s assessment that it is too costly for individuals to observe every social interaction. Most of the time, information flows from one person to another. However, even if we assume indirect observation, we still need to consider the possibility that rumors and gossip distort information. For example, think about a telephone game. We are often amused to discover that what the first person told the second person is completely different from what the last person is told. Indeed, even before WWII, by using the method of serial reproduction,[18] Bartlett (1932), who is one of the pioneers of cognitive psychology, argued that remembering is not reproductive, but reconstructive and hence inherently unreliable. There are all kinds of biases in the information transmission process. Furthermore, if we consider the possibility that individuals may lie, the situation becomes even more complicated. The result is that each person has his/her own opinions about other people, and this is much more indicative of direct observation models. Therefore, we believe that Ohtsuki and Iwasa (2004)'s terminology is misleading. Whether or not each person directly observes behavior is not the issue. What is important is whether or not mistakes are shared among all members of a society. In this sense, we believe that perceptual errors should be treated as subjective. This does not imply that information must differ between individuals. Most of the time information and perception are shared. Our point is that 100% consensus is extremely unlikely. Even if the degree of consensus is 99%, the consequence matches that of the model with subjective perceptual errors. There are a small number of individuals who have different views about the other individuals, and this disagreement has a significant effect on the maintenance of indirect reciprocity.

7. Future directions

Although we have shown that it is a necessary condition for the emergence of generalized exchange to consider individuals who give to "bad" persons as "bad", there is still much to accomplish. First, we would like to expand our analysis into the selective-play situation. In this paper we assume a situation in which a pair of individuals is chosen randomly from a population. Using Yamagishi and Hayashi (1996)'s terminology, such a situation is considered forced-play (players are "forced" to interact with particular partners). However, since each player knows every other player's reputation (either subjective or objective), why not choose a desired recipient from the population when he meets a person whom he considers "bad"? Actually, in social science studies investigating the prisoner's dilemma, such situations are considered selective-play (i.e. players have the option to leave the current relationship and/or choose a new partner). We believe that selective-play environments are more indicative of human societies. Whether the same three principles are necessary for the maintenance of indirect reciprocity is a challenging theoretical question.

The second direction is to examine how people actually behave in generalized exchange situations. So far, compared to the number of theoretical studies, there are only a small number of empirical studies on indirect reciprocity in terms of how people actually use information and behave in generalized exchange settings (e.g., Bolton et al., 2005; Mashima and Takahashi, in press; Milinski et al., 2001; Wedekind and Milinski, 2000), and results have been mixed. First of all, it is still unclear whether people actually use second-order information given that it requires a greater level of cognitive complexity and effort. Milinski et al. (2001) found that people did not use second-order information, while Bolton et al. (2005) and Mashima and Takahashi (in press) found that they did. Second, even if people actually use second-order information, there is no satisfactory result that can resolve the debate explained above (i.e. whether people regard giving to "bad" as "bad" and/or not-giving to "bad" as "good"). Although Bolton et al. (2005) found that the giving rate was higher when second-order information was available than when it was not available, their results were not strong enough to answer whether participants distinguished justifiable giving from unjustifiable giving and/or justifiable not-giving from unjustifiable not-giving. On the contrary, Mashima and Takahashi (in press) found that people distinguished justifiable giving from unjustifiable giving, but what they conducted is not a laboratory experiment but a vignette experiment. Third, although in this paper we do not take each individual's own reputation into consideration, it may turn out to be an important factor. A laboratory experiment that can examine these questions is clearly needed. Such an experiment may shed light on the issue of whether people's behavioral strategies have an evolutionary origin or they are socially constructed.

Acknowledgement

We thank Hisashi Ohtsuki, Karthik Panchanathan, Yoh Iwasa for their valuable comments and suggestions. We also thank an anonymous reviewer for pointing out errors in equations in Appendix A of the original draft and suggesting the approximate solution for the unstable equilibrium along the ALLD-ES edge in Fig. 2 (Eq. (2)).

[17]If we assume a collective mind of a society, such as the Borg in Star Trek, then all perceptual errors are completely shared among all members of a society.

[18]One individual transmits a piece of information to a second person, and then the second person transmits it to a third person, and so on in a chain.

Appendix A. Indirect reciprocity based on the SDISC strategy with the possibility of one-way implementation errors

The frequencies for ALLC, ALLD, and SDISC are given by x_1, x_2, and x_3, respectively. The rate of one-way implementation errors (intended giving leading to accidental not-giving) is denoted by the parameter α. The parameter w measures the probability of an additional round of interaction. We denote the frequency of "good" individuals among the whole population in any particular round as g_n. We denote the frequency of good individuals in round n for ALLC as $g_n(\mathrm{ALLC})$, for ALLD as $g_n(\mathrm{ALLD})$, and for SDISC as $g_n(\mathrm{SDISC})$, respectively. In round 1, we assume all individuals are regarded as "good".

Thus, in round n, where $n>1$,

$$\begin{aligned}
&g_n(\mathrm{ALLC}) = g_{n-1}(1-\alpha),\\
&g_n(\mathrm{ALLD}) = 0,\\
&g_n(\mathrm{SDISC}) = g_{n-1}(1-\alpha),\\
&g_n = g_n(\mathrm{ALLC})x_1 + g_n(\mathrm{ALLD})x_2 + g_n(\mathrm{SDISC})x_3\\
&\quad = (x_1+x_3)g_n(\mathrm{ALLC}) = (x_1+x_3)g_n(\mathrm{SDISC}).
\end{aligned} \tag{A.1}$$

$$g_2(\mathrm{ALLC}) = g_2(\mathrm{SDISC}) = 1-a, \quad \text{and} \quad g_2(\mathrm{ALLD}) = 0.$$

Thus, where $n>2$,

$$\begin{aligned}
&g_n(\mathrm{ALLC}) = g_{n-1}(1-\alpha) = (x_1+x_3)(1-\alpha)g_{n-1}(\mathrm{ALLC}),\\
&g_n(\mathrm{ALLD}) = 0,\\
&g_n(\mathrm{SDISC}) = (x_1+x_3)(1-\alpha)g_{n-1}(\mathrm{SDISC}).
\end{aligned} \tag{A.2}$$

To compute fitness functions, we first derive round 1 payoffs.

$$\begin{aligned}
&W_1(\mathrm{ALLC}) = -c(1-\alpha) + b(1-\alpha)x_1 + b(1-\alpha)x_3,\\
&W_1(\mathrm{ALLD}) = b(1-\alpha)x_1 + b(1-\alpha)x_3,\\
&W_1(\mathrm{SDISC}) = -c(1-\alpha) + b(1-\alpha)x_1 + b(1-\alpha)x_3.
\end{aligned} \tag{A.3}$$

Then, we derive the payoffs in round n.

$$\begin{aligned}
&W_n(\mathrm{ALLC}) = -c(1-\alpha) + b(1-\alpha)x_1 + g_n(\mathrm{ALLC})b(1-\alpha)x_3,\\
&W_n(\mathrm{ALLD}) = b(1-\alpha)x_1 + g_n(\mathrm{ALLD})b(1-\alpha)x_3 = b(1-\alpha)x_1,\\
&W_n(\mathrm{SDISC}) = -cg_n(1-\alpha) + b(1-\alpha)x_1 + g_n(\mathrm{SDISC})b(1-\alpha)x_3.
\end{aligned} \tag{A.4}$$

Summing up:

$$\begin{aligned}
W(\mathrm{ALLC}) &= W_1(\mathrm{ALLC}) + wW_2(\mathrm{ALLC}) + w^2W_3(\mathrm{ALLC}) + \cdots\\
&= \frac{1}{1-w}\{-c(1-\alpha) + b(1-\alpha)x_1\} + b(1-\alpha)x_3\{g_1(\mathrm{ALLC}) + wg_2(\mathrm{ALLC}) + w^2g_3(\mathrm{ALLC})\cdots\}\\
&= \frac{1}{1-w}\{-c(1-\alpha) + b(1-\alpha)x_1\} + b(1-\alpha)x_3\frac{1+w(1-\alpha)(1-x_1-x_3)}{1-w(1-\alpha)(x_1+x_3)},
\end{aligned} \tag{A.5}$$

$$\begin{aligned}
W(\mathrm{ALLD}) &= W_1(\mathrm{ALLD}) + \left(\frac{1}{1-w} - 1\right)W_n(\mathrm{ALLD})\\
&= b(1-\alpha)x_1 + b(1-\alpha)x_3 + \frac{w}{1-w}b(1-\alpha)x_1\\
&= \frac{b(1-\alpha)}{1-w}x_1 + b(1-\alpha)x_3,
\end{aligned} \tag{A.6}$$

$$
\begin{aligned}
W(\mathrm{SDISC}) &= W_1(\mathrm{SDISC}) + wW_2(\mathrm{SDISC}) + w^2W_3(\mathrm{SDISC}) + \cdots \\
&= \frac{1}{1-w}b(1-\alpha)x_1 - c(1-\alpha)(g_1 + wg_2 + w^2g_3\cdots) \\
&\quad + b(1-\alpha)x_3\{g_1(\mathrm{ALLC}) + wg_2(\mathrm{ALLC}) + w^2g_3(\mathrm{ALLC})\cdots\} \\
&= \frac{b(1-\alpha)x_1}{1-w} - \frac{c(1-\alpha)}{1-w(1-\alpha)(x_1+x_3)} + b(1-\alpha)x_3\frac{1+w(1-\alpha)(1-x_1-x_3)}{1-w(1-\alpha)(x_1+x_3)} \\
&= \frac{b(1-\alpha)x_1}{1-w} + \frac{b(1-\alpha)x_3\{1+w(1-\alpha)(1-x_1-x_3)\} - c(1-\alpha)}{1-w(1-\alpha)(x_1+x_3)}. \qquad \text{(A.7)}
\end{aligned}
$$

We now turn to fitness comparisons.

$$
W(\mathrm{ALLD}) - W(\mathrm{ALLC}) = \frac{c(1-\alpha)}{1-w} - \frac{wb(1-\alpha)^2x_3}{1-w(1-\alpha)(x_1+x_3)}. \qquad \text{(A.8)}
$$

Thus, when there is no SDISC ($x_3 = 0$),

$$
W(\mathrm{ALLD}) - W(\mathrm{ALLC}) = \frac{c(1-\alpha)}{1-w}. \qquad \text{(A.9)}
$$

It is clear that $W(\mathrm{ALLD}) > W(\mathrm{ALLC})$ regardless of c, w, and α. Therefore, there is no equilibrium on the ALLC–ALLD edge.

$$
W(\mathrm{SDISC}) - W(\mathrm{ALLC}) = \frac{c(1-\alpha)}{1-w} - \frac{c(1-\alpha)}{1-w(1-\alpha)(x_1+x_3)}. \qquad \text{(A.10)}
$$

When $x_2 = 0$, $x_1 + x_3 = 1$. Then, we have

$$
W(\mathrm{SDISC}) - W(\mathrm{ALLC}) = \frac{wc\alpha(1-\alpha)}{(1-w)\{1-w(1-\alpha)\}}. \qquad \text{(A.11)}
$$

When $\alpha > 0$, $W(\mathrm{SDISC}) > W(\mathrm{ALLC})$ regardless of c and w. Therefore, there is no equilibrium on the ALLC–SDISC edge.

$$
W(\mathrm{SDISC}) - W(\mathrm{ALLD}) = \frac{(1-\alpha)\{bw(1-\alpha)x_3 - c\}}{1-w(1-\alpha)(x_1+x_3)}. \qquad \text{(A.12)}
$$

Since $1-\alpha > 0$ and $1-w(1-\alpha)(x_1+x_3) > 0$, the sign of (A.12) depends on the sign of (A.13).

$$
bw(1-\alpha)x_3 - c. \qquad \text{(A.13)}
$$

Thus, there is a critical line, (A.14), that determines whether $W(\mathrm{SDISC})$ is greater than $W(\mathrm{ALLD})$.

$$
x_3 = \frac{c}{bw(1-\alpha)}. \qquad \text{(A.14)}
$$

This is also the equilibrium point on the SDISC–ALLD edge.

Appendix B. Indirect reciprocity based on the Extra Standing (ES) strategy with the possibility of one-way implementation errors

The frequencies for ALLC, ALLD, and ES are given by x_1, x_2, and x_4, respectively. The rate of one-way implementation errors (intended giving leading to accidental not-giving) is denoted by the parameter α. The parameter w measures the probability of an additional round of interaction. We denote the frequency of "good" individuals among the whole population in any particular round as g_n. We denote the frequency of good individuals in round n for ALLC as $g_n(\mathrm{ALLC})$, for ALLD as $g_n(\mathrm{ALLD})$, and for ES as $g_n(\mathrm{ES})$, respectively. In round 1, we assume all individuals are regarded as "good".

Thus, in round n, where $n > 1$,

$$
\begin{aligned}
&g_n(\mathrm{ALLC}) = g_{n-1}(1-\alpha) + (1-g_{n-1})\alpha = (1-2\alpha)g_{n-1} + \alpha, \\
&g_n(\mathrm{ALLD}) = 1 - g_{n-1}, \\
&g_n(\mathrm{ES}) = g_{n-1}(1-\alpha) + 1 - g_{n-1} = 1 - g_{n-1}\alpha, \\
&g_n = g_n(\mathrm{ALLC})x_1 + g_n(\mathrm{ALLD})x_2 + g_n(\mathrm{ES})x_4 \\
&\quad = \{(1-2\alpha)x_1 - x_2 - \alpha x_4\}g_{n-1} + \alpha x_1 + x_2 + x_4. \qquad \text{(B.1)}
\end{aligned}
$$

We can solve the recursion for g_n, and we have

$$g_n = \frac{\{1+(\alpha-1)(x_1+x_4)\}\{(1-2\alpha)x_1 - x_2 - \alpha x_4\}^{n-1} + \alpha x_1 + x_2 + x_4}{1-\{x_1(1-2\alpha) - x_2 - \alpha x_4\}}$$
$$= \frac{\{1-(1-\alpha)(x_1+x_4)\}\{(1-\alpha)(2x_1+x_4)-1\}^{n-1} + 1 - (1-\alpha)x_1}{2-(1-\alpha)(2x_1+x_4)}. \tag{B.2}$$

Therefore, we can solve all recursions.

$$\begin{aligned}
g_n(\mathrm{ALLC}) &= (1-2\alpha)g_{n-1} + \alpha \\
&= (1-2\alpha)\frac{\{1+(\alpha-1)(x_1+x_4)\}\{(1-2\alpha)x_1 - x_2 - \alpha x_4\}^{n-2} + \alpha x_1 + x_2 + x_4}{1-\{x_1(1-2\alpha) - x_2 - \alpha x_4\}} + \alpha \\
&= \frac{(1-2\alpha)\{1+(\alpha-1)(x_1+x_4)\}\{(1-\alpha)(2x_1+x_4)-1\}^{n-2} + (1-\alpha)(x_1+\alpha x_4) + 1}{2-(1-\alpha)(2x_1+x_4)}, \\
g_n(\mathrm{ALLD}) &= 1 - g_{n-1} \\
&= 1 - \frac{\{1+(\alpha-1)(x_1+x_4)\}\{(1-2\alpha)x_1 - x_2 - \alpha x_4\}^{n-1} + \alpha x_1 + x_2 + x_4}{1-\{x_1(1-2\alpha) - x_2 - \alpha x_4\}} \\
&= \frac{\{(1-\alpha)(x_1+x_4)-1\}\{(1-\alpha)(2x_1+x_4)-1\}^{n-2} - (1-\alpha)(x_1+x_4) + 1}{2-(1-\alpha)(2x_1+x_4)}, \\
g_n(\mathrm{ES}) &= 1 - g_{n-1}\alpha \\
&= 1 - \alpha\frac{\{1+(\alpha-1)(x_1+x_4)\}\{(1-2\alpha)x_1 - x_2 - \alpha x_4\}^{n-1} + \alpha x_1 + x_2 + x_4}{1-\{x_1(1-2\alpha) - x_2 - \alpha x_4\}} \\
&= \frac{2-\alpha-(1-\alpha)\{(2-\alpha)x_1+x_4\} - \alpha\{1-(1-\alpha)(x_1+x_4)\}\{(1-\alpha)(2x_1+x_4)-1\}^{n-2}}{2-(1-\alpha)(2x_1+x_4)}.
\end{aligned} \tag{B.3}$$

To compute fitness functions, we first derive round 1 payoffs.

$$\begin{aligned}
W_1(\mathrm{ALLC}) &= -c(1-\alpha) + b(1-\alpha)x_1 + b(1-\alpha)x_4, \\
W_1(\mathrm{ALLD}) &= b(1-\alpha)x_1 + b(1-\alpha)x_4, \\
W_1(\mathrm{ES}) &= -c(1-\alpha) + b(1-\alpha)x_1 + b(1-\alpha)x_4.
\end{aligned} \tag{B.4}$$

Then, we derive the payoffs in round n.

$$\begin{aligned}
W_n(\mathrm{ALLC}) &= -c(1-\alpha) + b(1-\alpha)x_1 + g_n(\mathrm{ALLC})b(1-\alpha)x_4, \\
W_n(\mathrm{ALLD}) &= b(1-\alpha)x_1 + g_n(\mathrm{ALLD})b(1-\alpha)x_4, \\
W_n(\mathrm{ES}) &= -cg_n(1-\alpha) + b(1-\alpha)x_1 + g_n(\mathrm{ES})b(1-\alpha)x_4.
\end{aligned} \tag{B.5}$$

Summing up:

$$\begin{aligned}
W(\mathrm{ALLC}) &= W_1(\mathrm{ALLC}) + wW_2(\mathrm{ALLC}) + w^2W_3(\mathrm{ALLC}) + \cdots \\
&= \frac{1}{1-w}\{-c(1-\alpha) + b(1-\alpha)x_1\} + b(1-\alpha)x_4\{g_1(\mathrm{ALLC}) + wg_2(\mathrm{ALLC}) + w^2g_3(\mathrm{ALLC})\cdots\} \\
&= \frac{1-\alpha}{1-w}\left[bx_1 - c + \frac{bx_4\{1+w(1-2x_1-x_4)(1-\alpha) - w^2(1-\alpha)(1-x_1-x_4+\alpha x_4)\}}{1+w\{1-(1-\alpha)(2x_1+x_4)\}}\right],
\end{aligned} \tag{B.6}$$

$$\begin{aligned}
W(\mathrm{ALLD}) &= W_1(\mathrm{ALLD}) + wW_2(\mathrm{ALLD}) + w^2W_3(\mathrm{ALLD}) + \cdots \\
&= \frac{b(1-\alpha)}{1-w}x_1 + b(1-\alpha)x_4\{g_1(\mathrm{ALLD}) + wg_2(\mathrm{ALLD}) + w^2g_3(\mathrm{ALLD})\cdots\} \\
&= b(1-\alpha)\left[\frac{x_1}{1-w} + \frac{x_4\{w^2x_1(1-\alpha) - w(2x_1+x_4)(1-\alpha)+1\}}{(1-w)[1+w\{1-(1-\alpha)(2x_1+x_4)\}]}\right],
\end{aligned} \tag{B.7}$$

$$
\begin{aligned}
W(\mathrm{ES}) &= W_1(\mathrm{ES}) + wW_2(\mathrm{ES}) + w^2W_3(\mathrm{ES}) + \cdots \\
&= \frac{b(1-\alpha)x_1}{1-w} - c(1-\alpha)(g_1 + wg_2 + w^2g_3\cdots) \\
&\quad + b(1-\alpha)x_4\{g_1(\mathrm{ES}) + wg_2(\mathrm{ES}) + w^2g_3(\mathrm{ES})\cdots\} \\
&= (1-\alpha)\left[\frac{bx_1}{1-w} - \frac{c\{1-wx_1(1-\alpha)\}}{(1-w)\{1+w(1-(2x_1+x_4)(1-\alpha))\}}\right. \\
&\quad + bx_4\left\{1 + \frac{w\alpha\{1-(x_1+x_4)(1-\alpha)\}}{\{2-(2x_1+x_4)(1-\alpha)\}\{1+w(1-2(x_1+x_4))(1-\alpha)\}}\right. \\
&\quad \left.\left. - \frac{w\{x_1(1-\alpha)(2-\alpha) + x_4(1-\alpha) - 2 + \alpha\}}{(1-w)\{2-(2x_1+x_4)(1-\alpha)\}}\right\}\right]. \qquad \text{(B.8)}
\end{aligned}
$$

We now turn to fitness comparisons.

$$
W(\mathrm{ALLD}) - W(\mathrm{ALLC}) = \frac{1-\alpha}{1-w}\left[c - \frac{bwx_4\{1-w(1-x_4(1-\alpha))\}(1-\alpha)}{1+w\{1-(2x_1+x_4)(1-\alpha)\}}\right]. \qquad \text{(B.9)}
$$

When there is no ES ($x_4 = 0$),

$$
W(\mathrm{ALLD}) - W(\mathrm{ALLC}) = \frac{c(1-\alpha)}{1-w}. \qquad \text{(B.10)}
$$

It is clear that $W(\mathrm{ALLD}) > W(\mathrm{ALLC})$ regardless of c, w, and α. Therefore, there is no equilibrium on the ALLC–ALLD edge.

$$
\begin{aligned}
W(\mathrm{ES}) - W(\mathrm{ALLC}) &= \frac{(1-\alpha)w\{1-(x_1+x_4)(1-\alpha)\}}{(1-w)\{2-(1-\alpha)(2x_1+x_4)\}} \\
&\quad \times \frac{J}{1-w^2(2x_1+2x_4-1)\{1-(2x_1+x_4)(1-\alpha)\}(1-\alpha) + w\{2-(4x_1+3x_4)(1-\alpha)-\alpha\}},
\end{aligned}
$$

where

$$
\begin{aligned}
J &= c\{2-(2x_1+x_4)(1-\alpha)\}\{1-w(2x_1+2x_4-1)(1-\alpha)\} \\
&\quad - bx_4[-2\alpha + w\{\alpha^2 + 2\alpha - 2 + 2x_1(1+\alpha)(1-\alpha) + x_4(1-\alpha)(1+2\alpha)\} \\
&\quad + w^2\{-3\alpha^2 + 6\alpha - 2 - 2(2x_1^2 + x_4^2)(1-\alpha)^3 - 2x_1(1-\alpha) \\
&\quad \times(-\alpha^2 + 6\alpha - 3 + 3x_4(1-\alpha)^2) - x_4(1-\alpha)(-\alpha^2 + 9\alpha - 5)\}]. \qquad \text{(B.11)}
\end{aligned}
$$

When $x_2 = 0$, $x_1 + x_3 = 1$. Then, we have

$$
\begin{aligned}
W(\mathrm{ES}) - W(\mathrm{ALLC}) &= \frac{-w\alpha(1-\alpha)}{(1-w)\{1-w(1-\alpha)\}\{x_1 - 1 - (1+x_1)\alpha\}\{1+w(\alpha+x_1\alpha-x_1)\}} \\
&\quad \times [c\{1-w(1-\alpha)\}\{1-x_1(1-\alpha)+\alpha\} - b(1-x_1) \\
&\quad \times\{-2\alpha - w(1-x_1(1-\alpha) - 3\alpha + \alpha^2) + w^2\{1+\alpha(\alpha-1) \\
&\quad \times(\alpha+2) - x_1(1-\alpha(2-\alpha(2-\alpha)))\}\}]. \qquad \text{(B.12)}
\end{aligned}
$$

Since $1-w>0$, $1-w(1-\alpha)>0$, $x_1-1-(1+x_1)\alpha<0$, $1+w(\alpha+x_1\alpha-x_1)>0$, the denominator is always negative. For the numerator, $-w(1-\alpha)\alpha<0$. Therefore, the sign of $W(\mathrm{ES})-W(\mathrm{ALLC})$ depends on the sign of the following:.

$$
\begin{aligned}
&[c\{1-w(1-\alpha)\}\{1-x_1(1-\alpha)+\alpha\} - b(1-x_1) \\
&\{-2\alpha - w(1-x_1(1-\alpha) - 3\alpha + \alpha^2) + w^2\{1+\alpha(\alpha-1) \\
&(\alpha+2) - x_1(1-\alpha(2-\alpha(2-\alpha)))\}\}]. \qquad \text{(B.13)}
\end{aligned}
$$

Given that $0<x_1<1$, $0<\alpha<1$, $0<w<1$, $b>0$, and $c>0$, it turns out that (B.13) is always positive. Therefore, $W(\mathrm{ES})$ is always greater than $W(\mathrm{ALLC})$. There is no equilibrium on the ALLC–ES edge.

$$
\begin{aligned}
W(\mathrm{ES}) - W(\mathrm{ALLD}) &= \frac{-(1-\alpha)}{(1-w)\{2-(2x_1+x_4)(1-\alpha)\}} \\
&\quad \times \frac{Z}{1+w\{2-\alpha-(4x_1+3x_4)(1-\alpha)\} - w^2(1-\alpha)(2x_1+2x_4-1)\{1-(2x_1+x_4)(1-\alpha)\}},
\end{aligned}
$$

where

$$\begin{aligned}Z = {} & c\{1-wx_1(1-\alpha)\}\{2-(2x_1+x_4)(1-\alpha)\}\{1-w(2x_1+2x_4-1)(1-\alpha)\}\\ & +bwx_4[-2+2(1-\alpha)x_1+(1-\alpha^2)x_4\\ & +w\{-\alpha^2+4\alpha-2-2x_1^2(3-\alpha)(1-\alpha)^2\\ & +x_1(1-\alpha)(\alpha^2-6\alpha+8-7x_4(1-\alpha))\\ & +x_4(5-3\alpha)(1-\alpha)-x_4^2(1-\alpha)^2(2+\alpha)\}\\ & +w^2\{4x_1^3(1-\alpha)^4+2x_1^2(1-\alpha)^2(-\alpha^2+6\alpha-3+3x_4(1-\alpha)^2)\\ & +\alpha(2-\alpha+3x_4^2(1-\alpha)^2-x4(1-\alpha)(5-\alpha))\\ & -x_1(1-\alpha)(-3\alpha^2+10\alpha-2-2x_4^2(1-\alpha)^3\\ & +x_4(1-\alpha)(\alpha^2-13\alpha+5))\}]. \end{aligned} \tag{B.14}$$

When $x_1 = 0$, we have

$$W(\mathrm{ES}) - W(\mathrm{ALLD}) = \frac{-(1-\alpha)Z}{(1-w)\{2-x_4(1-\alpha)\}[1+w\{2-\alpha-3x_4(1-\alpha)\}+w^2(1-\alpha)(1-2x_4)\{1-x_4(1-\alpha)\}]},$$

where

$$\begin{aligned}Z = {} & c\{2-x_4(1-\alpha)\}\{1-w(2x_4-1)(1-\alpha)\}\\ & +bwx_4[-2+(1-\alpha^2)x_4+w\{-\alpha^2+4\alpha-2\\ & +x_4(5-3\alpha)(1-\alpha)-x_4^2(1-\alpha)^2(2+\alpha)\}\\ & +w^2\{\alpha(2-\alpha+3x_4^2(1-\alpha)^2-x4(1-\alpha)(5-\alpha))\}]. \end{aligned} \tag{B.15}$$

Given that $0<x_4<1$, $0<\alpha<1$, $0<w<1$, $b>0$, and $c>0$, it turns out that the denominator is always positive. Also, since $-(1-\alpha)<0$, Eq (B.15) is positive when the sign of Eq (B.16) is negative.

$$\begin{aligned}Z = {} & c\{2-x_4(1-\alpha)\}\{1-w(2x_4-1)(1-\alpha)\}\\ & +bwx_4[-2+(1-\alpha^2)x_4+w\{-\alpha^2+4\alpha-2+x_4(5-3\alpha)(1-\alpha)\\ & -x_4^2(1-\alpha)^2(2+\alpha)\}+w^2\{\alpha(2-\alpha+3x_4^2(1-\alpha)^2\\ & -x4(1-\alpha)(5-\alpha))\}]. \end{aligned} \tag{B.16}$$

To find the equilibrium on the ES–ALLD edge, set Eq (B.16) = 0 as shown in Eq. (B.17) and solve it for x_4.

$$\begin{aligned}& bw^2(1-\alpha)^2\{(3w-1)\alpha-2\}x_4^3+w(1-\alpha)[2c(1-\alpha)\\ & +b\{1+\alpha+w(5-3\alpha)+w^2(\alpha-5)\alpha\}]x_4^2\\ & +[c\{-1+w(-5+\alpha)\}(1-\alpha)-bw\{2+w^2(-2+\alpha)\alpha\\ & +w(2-4\alpha+\alpha^2)\}]x_4+2c+2cw(1-\alpha)=0 \end{aligned} \tag{B.17}$$

Deriving the solution is too tedious to present here. Since this is a cubic equation, there are three solutions, and two of the solutions may be an imaginary number. Given that $0<x_4<1$, $0<\alpha<1$, $0<w<1$, $b>0$, and $c>0$, only one of them is the answer that we need. Although we could write the solution, it is not practical. Therefore, we show the approximate solution. If $\alpha = 0$, (B.17) becomes

$$\begin{aligned}& -2bw^2x_4^3+w\{2c+b(1+5w)\}x_4^2-\{c(1+5w)\\ & +2bw(1+w)\}x_4+2c(1+w)=0. \end{aligned} \tag{B.18}$$

The three solutions are $x_4 = 2$, c/bw, and $(1+w)/2w$. Since $x_4 \leqslant 1$, c/bw is the solution that we need. If $\alpha>0$ but very small, we expect that the solution is close to c/bw but slightly different. More concretely, here we assume that the solution can be represented by Eq. (B.19), where s_1, s_2, s_3 are unknown constants.

$$x_4 = \frac{c}{bw}+s_1\alpha+s_2\alpha^2+s_3\alpha^3+\cdots. \tag{B.19}$$

Here, we approximate that α^2, α^3, and higher orders are all zero. Then, Eq (B.19) becomes

$$x_4 = \frac{c}{bw} + s_1\alpha. \tag{B.20}$$

We input Eqs (B.20)–(B.17) and again approximate that α^2 and higher order terms are all zero.

$$\frac{\{bw(b-2c+bw)(2bw-c)s_1 + c(b-c+bw)(c-3cw+2bw^2)\}\alpha}{b^2w} = 0. \tag{B.21}$$

Then, we have the solution for s_1 as

$$s_1 = \frac{c(b-c+bw)(c-3cw+2bw^2)}{bw(b-2c+bw)(2bw-c)}. \tag{B.22}$$

Input Eqs (B.22)–(B.20), and we have

$$x_4 = \frac{c}{bw} + \left\{\frac{c(b-c+bw)(c-3cw+2bw^2)}{bw(b-2c+bw)(2bw-c)}\right\}\alpha. \tag{B.23}$$

Appendix C. Indirect reciprocity with the possibility of subjective perceptual errors

The frequencies for ALLC, ALLD, and the focal strategy (STAND, SDISC, or ES) are given by x_1, x_2, and x_5, respectively. In this section, we consider the possibility of two-way implementation errors (intended giving leading to accidental not-giving and intended not-giving leading to accidental giving) whose rate is denoted by the parameter α. We also consider the possibility of two-way subjective perceptual errors whose rate is denoted by the parameter δ. The parameter w measures the probability of an additional round of interaction. We denote the frequency of "good" individuals among the whole population in any particular round as g_n. We denote the frequency of good individuals in round n for ALLC as g_n(ALLC), for ALLD as g_n(ALLD), and for the focal strategy as g_n(FOCAL), respectively. In round 1, we assume all individuals are regarded as "good".

To compute fitness functions, we first derive round 1 payoffs.

$$\begin{aligned}
W_1(\text{ALLC}) &= -c(1-\alpha) + b(1-\alpha)x_1 + b\alpha x_2 + b(1-\alpha)x_5,\\
W_1(\text{ALLD}) &= -c\alpha + b(1-\alpha)x_1 + b\alpha x_2 + b(1-\alpha)x_5,\\
W_1(\text{FOCAL}) &= -c(1-\alpha) + b(1-\alpha)x_1 + b\alpha x_2 + b(1-\alpha)x_5.
\end{aligned} \tag{C.1}$$

Then, we derive the payoffs in round n.

$$\begin{aligned}
W_n(\text{ALLC}) &= -c(1-\alpha) + b(1-\alpha)x_1 + b\alpha x_2\\
&\quad + b\{g_n(\text{ALLC})(1-\alpha) + (1-g_n(\text{ALLC}))\alpha\}x_3,\\
W_n(\text{ALLD}) &= -c\alpha + b(1-\alpha)x_1 + b\alpha x_2 + b\{g_n(\text{ALLD})(1-\alpha)\\
&\quad + (1-g_n(\text{ALLD}))\alpha\}x_3,\\
W_n(\text{FOCAL}) &= -c\{g_{n-1}(1-\alpha) + (1-g_{n-1})\alpha\} + b(1-\alpha)x_1 + b\alpha x_2\\
&\quad + b\{g_n(\text{FOCAL})(1-\alpha) + (1-g_n(\text{FOCAL}))\alpha\}x_3.
\end{aligned} \tag{C.2}$$

From the above assumptions, we have

$$g_1(\text{ALLC}) = g_1(\text{ALLD}) = g_1(\text{FOCAL}) = 1,$$
$$g_n(\text{ALLC})x_1 + g_n(\text{ALLD})x_2 + g_n(\text{FOCAL})x_3 = g_n.$$

Up to this point, we can derive all expressions, and they are common regardless of the focal strategy. However, we cannot derive general expressions for g_n(ALLC), g_n(ALLD), and g_n(FOCAL). Since we assume that perceptual errors are subjective, g_n(ALLC), g_n(ALLD), and g_n(FOCAL) are different among individuals.

References

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

Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, New York.

Baker, J., 1987. Reducing bias and inefficiency in the selection algorithm. In: Grefenstette, J. (Ed.), Proceedings of the Second International Conference on Genetic Algorithms. Erlbaum, Hillsdale, NJ, pp. 14–21.

Bartlett, F.C., 1932. Remembering: A Study in Experimental and Social Psychology. Cambridge, Cambridge University Press Reprint 1977.

Bolton, G.E., Katok, E., Ockenfels, A., 2005. Cooperation among strangers with limited information about reputation. J. Public. Econ. 89, 1457–1468.

Brandt, H., Sigmund, K., 2004. The logic of reprobation: action and assessment rules in indirect reciprocity. J. Theor. Biol. 231, 475–486.

Grefenstette, J., 2000. Proportional selection and sampling algorithms. In: Back, T., Fogel, D., Michalewicz, T. (Eds.), Evolutionary Computation 1: basic Algorithms and Operators. Institute of Physics Publishing, Bristol, UK, pp. 172–180.

Hamilton, W.D., 1964. The genetic theory of social behavior. I and II. J. Theor. Biol. 7, 1–52.

Leimar, O., Hammerstein, P., 2001. Evolution of cooperation through indirect reciprocity. Proc. Roy. Soc. Lond. B 268, 745–753.

Mashima, R., Takahashi, N., in press. The emergence of generalized exchange by indirect reciprocity. In: Biel, A., Eek, D., Gärling, T., Gustafsson, M. (Eds.), New Issues and Paradigms in Research on Social Dilemmas, Springer, Berlin, NY.

Milinski, M., Semmann, D., Bakker, T.C.M., Krambeck, H.J., 2001. Cooperation through indirect reciprocity:image scoring or standing strategy? Proc. Roy. Soc. Lond. Ser. B—Biol. Sci. 268 (1484), 2495–2501.

Nowak, M.A., Sigmund, K., 1998a. Evolution of indirect reciprocity by image scoring. Nature 393, 573–577.

Nowak, M., Sigmund, K., 1998b. The dynamics of indirect reciprocity. J. Theor. Biol. 194, 561–574.

Ohtsuki, H., 2004. Reactive strategies in indirect reciprocity. J. Theor. Biol. 227, 299–314.

Ohtsuki, H., Iwasa, Y., 2004. How should we define goodness?: reputation dynamics in indirect reciprocity. J. Theor. Biol. 231, 107–120.

Ohtsuki, H., Iwasa, Y., 2006. The leading eight: social norms that can maintain cooperation by indirect reciprocity. J. Theor. Biol. 239, 435–444.

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.

Sugden, R., 1986. The Economics of Rights, Co-operation and Welfare. Basil Blackwell, Oxford, UK.

Takahashi, N., Mashima, R., 2003. The emergence of indirect reciprocity: is the standing strategy the answer? COE Working Paper 29, Hokkaido University,. Hokkaido, Japan.

Trivers, R.L., 1971. The evolution of reciprocal altruism. Q. Rev. Biol. 46, 35–57.

Yamagishi, T., Hayashi, N., 1996. Selective play: social embeddedness of social dilemmas. In: Liebrand, W.B.G., Messick, D.M. (Eds.), Frontiers in Social Dilemmas Research. Springer, Berlin, pp. 363–384.

Yamagishi, T., Takahashi, N., 1994. Evolution of Norms without Meta-Norms. In: Schulz, U., Albers, W., Mueller, U. (Eds.), Social Dilemmas and Cooperation. Springer, Berlin, pp. 311–326.

Wedekind, C., Milinski, M., 2000. Cooperation through image scoring in humans. Science 288, 850–852.