Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 231 (2004) 107–120
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

How should we define goodness?—reputation dynamics in indirect reciprocity

Hisashi Ohtsuki*, Yoh Iwasa

Department of Biology, Faculty of Sciences, Kyushu University, Hakozaki 6-10-1, Fukuoka 812-8581, Japan

Received 4 February 2004; received in revised form 8 June 2004; accepted 11 June 2004

Abstract

Theory of indirect reciprocity is important in explaining cooperation between humans. Since a partner of a social interaction often changes, an individual should assess his partner by using social information such as reputation and make decisions whether to help him or not. To those who have 'good' social reputation does a player give aid as reciprocation, whereas he has to refuse to help those who have 'bad' reputation. Otherwise benefits of altruism is easily exploited by them. Little has been known, however, about the definition of 'goodness' in reputation. What kind of actions *are* and *should be* regarded as good and what kind of actions bad? And what sort of goodness enables sustaining exchange of altruism? We herein challenge this question with an evolutionary perspective. We generalize social reputation as 'Honor-score' (*H*-score) and examine the conditions under which individuals in a group stably maintain cooperative relationships based on indirect reciprocity. We examine the condition for evolutionarily stable strategies (ESSs) over 4096 possible cases exhaustively. Mathematical analysis reveals that only eight cases called '*leading eight*' are crucial to the evolution of indirect reciprocity. Each in the *leading eight* shares two common characteristics: (i) cooperation with *good* persons is regarded as good while defection against them is regarded as bad, and (ii) defection against *bad* persons should be regarded as a good behavior because it works as sanction. Our results give one solution to the definition of goodness from an evolutionary viewpoint. In addition, we believe that the formalism of reputation dynamics gives general insights into the way social information is generated, handled, and transmitted in animal societies.

Keywords: Indirect reciprocity; Reputation dynamics; Goodness; Leading eight; ESS

1. Introduction

Humans are distinct in their social ability to get along with others. It is true that eusocial animals such as honeybees or wasps are prominently social, but their well-organized societies are based upon close kinship among individuals. Humans, on the other hand, exhibit remarkable cooperative tendency towards unrelated individuals.

Indirect reciprocity (Alexander, 1987; Fehr and Fischbacher, 2003, Nowak and Sigmund, 1998b) gives one clear-cut explanation of how this preference for cooperation has evolved. When an individual behaves altruistically towards another, he suffers the cost of the help, such as time, energy, or risks, hence his action does not seem to pay for the moment. However, if a third person who knows of his good deed recompenses him with cooperation, the cost he paid will be cancelled, and consequently he may result in getting positive net benefit. Indirect reciprocity differs from direct reciprocity or reciprocal altruism (Trivers, 1971) in that the donor of the help receives the return not from the beneficiary himself but from another individual. Due to this nature of indirect reciprocity, the availability of social information on each individual is necessary in order to sustain cooperation. Without such information individuals cannot discriminate cooperative persons from cheaters or social parasites, who enjoy benefit of

*Corresponding author. Tel.: +81-92-642-2641; fax: +81-92-642-2645.
E-mail address: ohtsuki@bio-math10.biology.kyushu-u.ac.jp (H. Ohtsuki).

0022-5193/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2004.06.005

altruism with paying nothing. Since nobody assures that those who are cooperated never fail to offer help to another, there must be a way to know who is worth cooperating and who is not. Therefore social information such as reputation or rumor must have greatly contributed to the evolution of indirect reciprocity by realizing discriminate altruism. Verbal communication is a sophisticated way to transfer social information (Enquist and Leimar, 1993; Nakamaru and Kawata, 2004; Pollock and Dugatkin, 1992).

Several previous studies have revealed the characteristics of indirect reciprocity. Nowak and Sigmund (1998b) is a milestone on this issue. They introduced 'image score' as an index which measures goodness or cooperativeness of each individual and examined what behavioral rule may evolve. Image score is of an integer value, and it increases by one unit every time one cooperates, whereas it decreases by one when one refuses to cooperate. Hence, a player's image score is an appropriate portrait of his preference for cooperation. Using computer simulation, Nowak and Sigmund (1998b) showed that 'discriminator' strategy is an outcome of the evolution. This strategy prescribes cooperation only with an individual with nonnegative image score. Since nonnegative image score suggests that the focal person tends to cooperate rather than to defect, the success of discriminators shows that not unconditional altruism but conditional cooperation is necessary for the evolution of indirect reciprocity. They also considered strategies which take their own image score into account in making decisions. For example, 'And' strategy cooperates if his score is low *and* the opponent's score is high. 'Or' strategy, on the other hand, cooperates if his score is low *or* the opponent's score is high. Nowak and Sigmund (1998b) found that such strategies also promote cooperation.

Leimar and Hammerstein (2001) critically reexamined Nowak and Sigmund's study. They pointed out that a discriminator receives no benefit by caring about the image score of others. When an individual cooperates, he gains good reputation irrespective of his opponent's reputation. Thus there is no need for a player to consider other's reputation. What a player is really interested in is his own reputation because it directly affects the future benefit of him. Leimar and Hammerstein (2001) took Δq strategy for an example to demonstrate that strategies which smartly calculate the influence of giving behavior on other's decisions defeat discriminators. They claimed that the scope for discriminators are very narrow. Instead they examined an alternative called 'standing strategy', which Sugden (1986) had previously proposed. A player who uses this strategy adopts a criterion different from image score in assessing others: this criterion is called 'standing'. Standing of each individual is either good or bad and is supposed to be good at the initial state. An individual in bad standing can regain his good standing only by cooperating with others, which is the same rule as image score. However, a person in good standing falls into bad if and only if he fails to cooperate with an opponent in *good* standing. Even if he refuses to help an individual in *bad* standing, he does not lose his good standing. This is because the refusal is interpreted as punishment against a selfish individual (for studies on punishment, see Brandt and Sigmund (2003), Fehr and Gächter (2000), Fehr and Rockenbach (2003), and Henrich and Boyd (2001)). It is true that withdrawal of cooperation from a person with bad reputation might be thought as bad behavior if people pay attention only to the action itself. However, if they consider the motivation of the action; noting that the purpose of the defection is to impose sanctions on cheaters, the withdrawal can be labeled as good.

Difference between discriminator and standing strategy lies also in the way of using social reputation. A discriminator, who adopts image score as a reputation-assignment rule, considers only the reputation of his opponent in making his decisions and will cooperate with those who have a good image score. A standing strategist, on the other hand, uses his own reputation as well in order to do better in a group. He always monitors his own status within the population and sometimes tries to enhance his standing so as to further increase the possibility that he receives cooperation from others. Hence he offers help if he is in bad standing *or* if his opponent is in good standing, which is similar to the 'Or' strategy noted above.

In short, discriminator and standing strategy differs in the notion of goodness. Ohtsuki (2004) demonstrated that discriminate altruism cannot evolve, if everyone in a group adopts the image-score criterion. However, if everyone adopts the standing criterion, discriminate altruism can be maintained (Panchanathan and Boyd, 2003). We can also imagine many other variants of reputation-assignment rules. For example, Sugden (1986) considered standing of a different kind (we call it 'strict-standing'). When a population uses strict-standing, an individual in bad standing can regain his good standing only by helping a good person; helping a person in bad standing is not enough because he may be regarded as a traitor. A reputation rule of the similar property was briefly mentioned in Panchanathan and Boyd (2003), who suggested that it have a different evolutionary outcome. The question now rises what reputation-assignment rule is plausible and what realizes discriminate altruism. As far as we know, no study has systematically examined evolutionary basis of those rules so far. Here, we will formulate such reputation-assignment rules as 'reputation dynamics' and challenge the question. How should we define goodness? We examine evolutionary origin of the notion of goodness that allows the sustaining exchange of altruism. Our exhaustive ESS analysis gives an answer.

2. The model

2.1. Indirect reciprocity game

Consider an infinitely large population in which two players randomly form a pair and engage in one-shot 'indirect reciprocity game'. In this game, two players simultaneously determine whether to help the opponent (=Cooperation, denoted by C) or not (=Defection, D), considering social reputation of both self and the opponent. Help costs the donor c while the recipient of the help receives the benefit $b(>c)$. Neither gains if help is not offered. Mutual help yields $b-c(>0)$ to both, whereas unilateral help brings the best outcome to the recipient and the worst to the donor. Hence, this game represents social dilemma. After one-round of the game, each player changes his partner and will interact with a different opponent. We assume that the entire period of social interactions is sufficiently long, and that players engage in the indirect reciprocity game many times with many different opponents. Since the population is assumed to be infinitely large, a player does not meet the same opponent again. Therefore, direct reciprocity is excluded from the model.

2.2. H-score representation of social reputation

If a player wants to assess another with respect to his cooperative tendency, it is useful to know the whole history of his actions, such as how many times and with whom he has cooperated. It is, however, difficult for players to monitor and memorize all the events in a large population due to costs, time, or capacity. In order to avoid such difficulties, individuals instead use a simpler form of information storage, such as 'he is nice' or 'he is untrustworthy'. Thanks to the great fluidity of language, reputation or rumor (Nakamaru and Kawata, 2004) can contribute to achieving such a purpose.

Here, we consider social reputation of the simplest kind; that is, one's social reputation is binary, like '*good* or *bad*' or '*nice* or *wicked*'. In order to make the following mathematical analyses easy, we introduce 'Honor-score' (H-score) as an index of one's social reputation. H-score is a binary digit (either '1' or '0') and represents its owner's social reputation. In the following, we are to regard H-score of 1 as good reputation and 0 as bad for simplicity. However, if we switch H-score of 1 and 0 and regard 1 as bad and 0 as good, this does not affect the basic property of our results at all. It is only for making interpretation of our results easy that we relate H-score of 1 to good reputation. H-score classifies players in the population into two groups with different social reputation. Note that H-score realizes theoretical generalization of such criteria of social reputation as image score (binary version) (Nowak and Sigmund, 1998a) or standing (Sugden, 1986).

2.3. Reputation dynamics

How do players in a population assign H-score to each other? For example, when a player cooperates with a person whose social reputation is bad, he may be considered good if an observer pays attention to his cooperative action itself. However, if an observer notes the motivation of his action, i.e. if the observer thinks that his cooperation is based on his vicious motivation to give aid to asocial conspirators, he may be regarded as bad. Assignment of reputation comprises a matter of definition of goodness as this. Here we model the notion of goodness in a population as 'reputation dynamics'. We assume that everyone in a population shares the same reputation dynamics, like public opinion in a society.

Reputation dynamics is represented by $d=(d_{ijX})_{i,j=1,0;\ X=C,D}$. Each component is either 1 or 0. A society with reputation dynamics d assigns H-score d_{ijX} to a focal player E_1 if (i) H-score of E_1 is i, (ii) H-score of E_1's opponent, say E_2, is j, and (iii) E_1's action towards E_2 is X ($=C$ or D). For example, reputation dynamics $(d_{11C}, d_{11D}, d_{10C}, d_{10D}, d_{01C}, d_{01D}, d_{00C}, d_{00D})=(1, 0, 1, 0, 1, 0, 1, 0)$ assigns H-score of 1 only to those who cooperate. Hence this is image score. We abbreviate image score as $d=\mathsf{IMAGE}$. Reputation dynamics $(d_{11C}, d_{11D}, d_{10C}, d_{10D}, d_{01C}, d_{01D}, d_{00C}, d_{00D})=(1, 0, 1, 1, 1, 0, 1, 0)$ corresponds to standing. The abbreviation of this is given by $d=\mathsf{STAND}$. There can be eight different situations (i, j, X), so it follows that we have $2^8=256$ different reputation dynamics within our

Table 1
A few examples of reputation dynamics

d_{11C}	d_{11D}	d_{10C}	d_{10D}	d_{01C}	d_{01D}	d_{00C}	d_{00D}	Name	Abbreviation
1	0	1	0	1	0	1	0	Image score	IMAGE
1	0	1	1	1	0	1	0	Standing	STAND
1	0	1	1	1	0	0	0	Strict-standing	S-STAND
1	0	0	1	1	0	0	0	Judging	JUDGE

formalism (see also Brandt and Sigmund (2004)). Table 1 lists up several examples ofreputation dynamics.

2.4. Behavioral strategies

Under a given reputation dynamics, how should each player handle social information? We assume that each player has his own reputation usage as his 'behavioral strategy'. A player's behavioral strategy prescribes how to behave towards others, considering social reputation of both self and his opponent. It can be a conditional strategy in general. It is represented by a vector $p = (p_{ij})_{i,j=1,0}$. Four components p_{ij} are either C (=Co-operation) or D (=Defection). A player with behavioral strategy p chooses the action p_{ij} ($= C$ *or* D) when his H-score is i and that of his present opponent is j. For example, behavioral strategy $(p_{11}, p_{10}, p_{01}, p_{00}) = (C, D, C, D)$ prescribes cooperation if and only if his opponent's H-score is 1. We call it 'Co-strategy' and abbreviate it as $p = \mathsf{CO}$, because it pays attention only to the reputation of its co-player. In contrast, behavioral strategy $(p_{11}, p_{10}, p_{01}, p_{00}) = (D, D, C, C)$ considers only H-score of the player himself and gives help only when his own H-score is 0. We call it 'Self-strategy' ($p = \mathsf{SELF}$). A simple rule $(p_{11}, p_{10}, p_{01}, p_{00}) = (C, C, C, C)$ always prescribes cooperation, which we call 'All$C$-strategy' ($p = \mathrm{All}C$), while another rule $(p_{11}, p_{10}, p_{01}, p_{00}) = (D, D, D, D)$ is named 'All$D$-strategy' ($p = \mathrm{All}D$). There are four different cases in (i, j), so we have $2^4 = 16$ different behavioral strategies in total (see also Brandt and Sigmund (2004)). For several examples of behavioral strategies, seeTable 2.

When one reputation dynamics is fixed in the population, there are $2^4 = 16$ different behavioral strategies. A combination of reputation dynamics and a behavioral strategy, (d, p), is called 'ESS pair' when p is an evolutionarily stable strategy among those 16 possible behavioral strategies under the reputation dynamics d. In the following, we search for ESS pairs exhaustively. Since there are $2^8 = 256$ kinds of reputation dynamics, we need to investigate $2^8 \times 2^4 = 2^{12} = 4096$ pairs in total. Novelty of our framework is that it includes not only several famous strategies proposed so far but also all the possible cases.

Table 2
Several examples of behavioral strategies and their abbreviations

p_{11}	p_{10}	p_{01}	p_{00}	Name	Abbreviation
C	D	C	D	Co-strategy	CO
D	D	C	C	Self-strategy	SELF
D	D	C	D	And-strategy	AND
C	D	C	C	Or-strategy	OR
C	C	C	C	All$C$-strategy	All$C$
D	D	D	D	All$D$-strategy	All$D$

For example, discriminators (Nowak and Sigmund, 1998a) use image score and help only those who have good reputation, so they corresponds to $(d, p) = (\mathsf{IMAGE}, \mathsf{CO})$. In contrast, a standing strategist (Leimar and Hammerstein, 2001) adopts standing criterion and cooperates if standing of the opponent is good or if standing of self is bad. Hence, standing strategy corresponds to $(d, p) = (\mathsf{STAND}, \mathsf{OR})$.

2.5. Mirror symmetry

Recall that we introduced H-score only to classify individuals into two social states. This property causes 'mirror symmetry' in H-scores. If we switch 1 and 0 in H-score in all the places in the model, predictions derived from the model remain unchanged at all. Almost all pairs (d, p) have their mirror images (d', p'), which are equivalent to original ones provided that H-score of 1 and 0 are switched. For further details of the mirror symmetry, see Appendix A. Considering the equivalence, the number of effective pairs we need to study is reduced from 4096 to 2080.

3. Method

3.1. Indirect observation model

Consider the situation where player A meets and interacts with player B. Most other players, say player C, in a large population do not observe what A does. Only one player, say player D, who luckily observes the

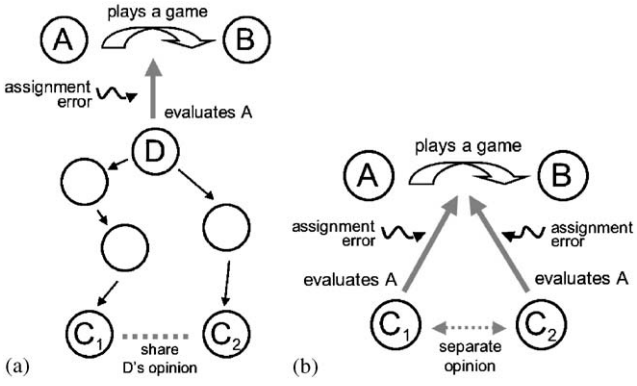

Fig. 1. (a) *Indirect observation model*: Even if player C wants to know the H-score of player A, he cannot observe what A does in the interaction with B due to the large population size. Instead C listens to D's opinion, who observed the interaction, either directly or indirectly. Therefore, if there are two individuals who are in the same situation as player C, say C_1 and C_2, they always have the same opinion on player A. (b) *Direct observation model*: Player C directly observes what A does in interaction with B, and assigns H-score to A by himself. If there are two individuals who are in the same situation as player C, say C_1 and C_2, they may have different opinion on A.

interaction, knows A's behavior. Hence other players in the population would listen to player D's evaluation about player A, either directly from D himself or indirectly by rumor. Fig. 1(a) sketches this situation. We assume that (i) D never tells a lie and (ii) C always believes what D says. As a consequence everyone in the population has the same opinion on a focal player, and hence H-score of each player can be treated as his internal state. We call this 'indirect observation model'. In contrast, Leimar and Hammerstein (2001) postulated a different model in which each player observes every social interaction in the population by himself. In order to clarify the difference, we call Leimar and Hammerstein's model 'direct observation model' (Fig. 1(b)). Later in this paper, we will discuss the difference between those two models. In the following, we study indirect observation model.

3.2. Roles of small errors

In examining evolutionary stability of strategies, we introduce two kinds of errors into the model; 'execution error' and 'assignment error'. Importance of error or noise in the analysis of indirect reciprocity has been stressed (Fishman, 2003; Leimar and Hammerstein, 2001; Lotem et al., 1999; Panchanathan and Boyd, 2003; Sherratt and Roberts, 2001; Sugden, 1986). If no one commits an error at all, difference between several strategies is lost, which results in many strategies being neutral with respect to payoffs. For example, if a discriminate altruist, who offers help only to good persons, invades a population of indiscriminate altruists, then only mutual help will occur and it will be repeated forever in the population. This will prevent us from comparing two strategies. If error is included, however, there will be some players who fail to cooperate and have bad reputation. Difference in behavior towards those accidental cheaters reveals the difference of strategies itself, and we can clearly distinguish these two.

The first kind of error, execution error, occurs when a player is to cooperate with others. He may sometimes fail to achieve cooperation against his will due to this error. We set the probability that a player commits execution error as $\mu_e(>0)$. Note that we do not consider the opposite situation in which a player mistakenly cooperates in spite that he wants to defect. The error of the second type, assignment error, happens when an observer of an interaction is to assign reputation to a focal person. He may misperceive an action or misuse reputation dynamics in assigning reputation. The probability that this error occurs is given by $\mu_a(>0)$. We incorporate both cases: (i) a player mistakenly assigns H-score of 1 when he should assign 0, and (ii) a player mistakenly assigns H-score of 0 when he should assign 1. Note that Leimar and Hammerstein (2001) considered the error caused only by perception of actions and called it 'perception error'. In contrast, assignment error includes causes other than perception of actions, because the mistake of assignment can occur even if perception of the action is correct.

3.3. Dynamics of H-score and ESS analysis

3.3.1. Dynamics of H-score in continuous time

We postulate that reputation dynamics d is fixed in the population. In other words, all in the population share the same criterion about goodness. Imagine that a small number of mutants with behavioral strategy p' invades a population dominated by p strategists. Let $h_t(p)$ be a proportion of wild-type individuals whose H-score is 1 among wild-type players at time t, and $h_t(p')$ be that among mutants. Both $h_t(p)$ and $h_t(p')$ change over time, and their dynamics depend on behavioral strategies (p and p') and reputation dynamics d.

First, we consider the change of $h_t(p')$ in a short time interval $[t, t+\Delta t]$. We assume that within this interval a fraction of Δt players in the population engage in the indirect reciprocity game once. There are three types in p' strategists, differing in the transition of their H-score.

(i) *Players with H-score* 1 *at time t, who engage in a game*: The fraction of these players among p' strategists is $\Delta t \cdot h_t(p')$. Each of them meets a wild-type player and plays one-shot indirect reciprocity game. With probability $h_t(p)$, H-score of his wild-type opponent is 1. In this case, he takes action $X = p'_{11}$ and he will be assigned H-score of $d_{11p'_{11}}$ by an observer. With probability $(1-h_t(p))$, H-score of his wild-type opponent is 0. In that case, he takes action $X = p'_{10}$ and will be assigned a new H-score $d_{10p'_{10}}$.

(ii) *Players with H-score* 0 *at time t, who engage in a game*: The fraction of these players among p' strategists is $\Delta t \cdot (1-h_t(p'))$. Similarly to (i), they take action $X = p'_{01}$ and become assigned H-score of $d_{01p'_{01}}$ (with probability $h_t(p)$), or take action $X = p'_{00}$ and become assigned H-score $d_{00p'_{00}}$ (with probability $1-h_t(p)$).

(iii) *Players who do not play a game*: The fraction of these players among p' strategists is $(1-\Delta t)$, a fraction $h_t(p')$ of them have H-score of 1 at time t. Their H-score remains unchanged.

From these we have the following equation:

$$\begin{aligned} h_{t+\Delta t}(p') = {} & \Delta t \cdot h_t(p')\left[h_t(p)d_{11p'_{11}} + (1-h_t(p))d_{10p'_{10}}\right] \\ & + \Delta t \cdot (1-h_t(p'))\left[h_t(p)d_{01p'_{01}}\right. \\ & \left. + (1-h_t(p))d_{00p'_{00}}\right] + (1-\Delta t)h_t(p'). \end{aligned} \quad (1)$$

Taking the limit of $\Delta t \to 0$, we have the following differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}h_t(p') = h_t(p')\Big[h_t(p)d_{11p'_{11}} + (1-h_t(p))d_{10p'_{10}}\Big] + (1-h_t(p'))\Big[h_t(p)d_{01p'_{01}} + (1-h_t(p))d_{00p'_{00}}\Big] - h_t(p'). \quad (2)$$

3.3.2. Effects of two kinds of errors

In addition, we need to consider errors of two kinds. First, due to execution error a player fails to cooperate against his intention with probability μ_e. In other words, he is compelled to choose defection with probability μ_e. Only with probability $(1-\mu_e)$ can he perform his intended action. As a result, $d_{ijp'_{ij}}$'s in Eq. (2) are rewritten as $(1-\mu_e)d_{ijp'_{ij}} + \mu_e d_{ijD}$.

Next, consider assignment error. Imagine that the true H-score to be assigned is d_{ijX}. Because of this error, however, one is assigned the wrong H-score $(1-d_{ijX})$ with probability μ_a. Hence, the expected proportion of individuals who are assigned H-score of 1 is given by $(1-\mu_a)d_{ijX} + \mu_a(1-d_{ijX}) = (1-2\mu_a)d_{ijX} + \mu_a$.

Combining those two errors, $d_{ijp'_{ij}}$'s in Eq. (2) are replaced by

$$(1-\mu_e)\{(1-2\mu_a)d_{ijp'_{ij}} + \mu_a\} + \mu_e\{(1-2\mu_a)d_{ijD} + \mu_a\} = (1-2\mu_a)\{(1-\mu_e)d_{ijp'_{ij}} + \mu_e d_{ijD}\} + \mu_a. \quad (3)$$

Thus Eq. (2) is rewritten as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}h_t(p') &= h_t(p')h_t(p)\Big[(1-2\mu_a)\{(1-\mu_e)d_{11p'_{11}} + \mu_e d_{11D}\} + \mu_a\Big] \\ &+ h_t(p')(1-h_t(p))\Big[(1-2\mu_a)\{(1-\mu_e) \times d_{10p'_{10}} + \mu_e d_{10D}\} + \mu_a\Big] \\ &+ (1-h_t(p'))h_t(p)\Big[(1-2\mu_a)\{(1-\mu_e) \times d_{01p'_{01}} + \mu_e d_{01D}\} + \mu_a\Big] \\ &+ (1-h_t(p'))(1-h_t(p))\Big[(1-2\mu_a)\{(1-\mu_e)d_{00p'_{00}} \\ &+ \mu_e d_{00D}\} + \mu_a\Big] - h_t(p'). \end{aligned} \quad (4)$$

3.3.3. Invasibility condition

Now we are interested in whether the reputation of p-strategist and p'-strategist are good or bad after the long run. Based on Eq. (4), we can prove analytically that both $h_t(p)$ and $h_t(p')$ converge to stationary values $h_*(p)$ and $h_*(p')$, respectively, when t becomes infinitely large, and that the limits are independent of their initial values $h_0(p)$ and $h_0(p')$. Hence, initial reputation has no effect on the evolutionary outcome. For mathematical details of the analysis, see Appendix B.

Thanks to these properties, we can derive the ESS condition of each behavioral strategy analytically. Let $h_*(p)$ and $h_*(p')$ the equilibrium of Eq. (4) expressed in terms of parameters p, p', d, μ_e and μ_a. When time t is sufficiently large, the probability that a focal wild-type player receives donation in one social interaction is given by

$$\theta(p,p) = \delta(p_{11})h_*(p)h_*(p) + \delta(p_{10})h_*(p)(1-h_*(p)) + \delta(p_{01})(1-h_*(p))h_*(p) + \delta(p_{00}) \times (1-h_*(p))(1-h_*(p)), \quad (5)$$

where $\delta(C)=1$ and $\delta(D)=0$. The probability that a focal wild-type player helps his opponent in one social interaction is the same as above. The probability that a focal mutant player receives donation from a wild-type player in one social interaction is

$$\theta(p,p') = \delta(p_{11})h_*(p)h_*(p') + \delta(p_{10})h_*(p)(1-h_*(p')) + \delta(p_{01})(1-h_*(p))h_*(p') + \delta(p_{00}) \times (1-h_*(p))(1-h_*(p')) \quad (6)$$

and the one that a focal mutant player helps his wild-type opponent in one social interaction is

$$\theta(p',p) = \delta(p'_{11})h_*(p')h_*(p) + \delta(p'_{10})h_*(p')(1-h_*(p)) + \delta(p'_{01})(1-h_*(p'))h_*(p) + \delta(p'_{00}) \times (1-h_*(p'))(1-h_*(p)). \quad (7)$$

Let $W(p|p)$ be the average payoff of a wild-type and $W(p'|p)$ be that of a mutant. Then we have

$$W(p|p) = b\cdot\theta(p,p) - c\cdot\theta(p,p), \quad (8)$$

$$W(p'|p) = b\cdot\theta(p,p') - c\cdot\theta(p',p). \quad (9)$$

The ESS criterion for behavioral strategy p under reputation dynamics d is

$$W(p|p) > W(p'|p) \quad (\textit{for all}\ \ p' \neq p). \quad (10)$$

Hence the procedure for searching for ESS pairs is as follows:

(1) Fix one reputation dynamics d. Take a behavioral strategy p, the stability of which we want to check.
(2) Take a mutant strategy p' which is different from p.
(3) Calculate average H-scores $h_*(p)$ and $h_*(p')$, by using the procedure in Appendix B.
(4) Calculate average payoffs $W(p|p)$ and $W(p'|p)$, by using Eqs. (5)–(9).
(5) If $W(p|p) > W(p'|p)$ holds, then p is stable against invasions by p'.
(6) Repeat (2) to (5) for each of 15 behavioral strategies $p'(\neq p)$. If p is stable against all the others, then p is an ESS under reputation dynamics d. In other words, (d, p) is an ESS pair.

We assumed that everyone in the population agrees to use common reputation dynamics d. Also assumed is that all players have the same opinion about one person.

Hence one's 'reputation' is given by the society as a whole, not by each individual. Roughly speaking, our mathematical approach seeks the best way to handle public opinion. When (d, p) is found to be an ESS pair, it means that behavioral rule p is the best under a criterion of goodness d. Of course under a different criterion d' may the best behavioral strategy p' be different.

4. Results of exhaustive search

We have exhaustively searched for ESS pairs among 4096 cases of (d, p). When (d, p) is an ESS pair, its mirror image (d', p') is also an ESS pair (see Appendix A for details). Hence, we will need to study only one of the two. This confines our scope to 2080 cases out of 4096.

Fig. 2(a) shows the distribution of the relative payoff of ESS pairs, when the benefit is $b = 2$ and the cost is $c = 1$. Since All$D$-strategy is always evolutionarily stable under any reputation dynamics, we regard the pairs which behavioral strategy is $p = $ All$D$ as trivial and do not consider them in Fig. 2(a). In contrast, ESS pairs which behavioral strategy is not $p = $ All$D$ earn positive payoffs, so we pay attention to them in the following. We find 25 non-trivial ESS pairs. Their details are shown in Table 3. Because the benefit of the help is $b = 2$ and the cost is $c = 1$, the average payoff can be $b - c = 1$ at its maximum in principle. We expect that in such an ideal group all individuals would both give and receive help without committing any errors. On the

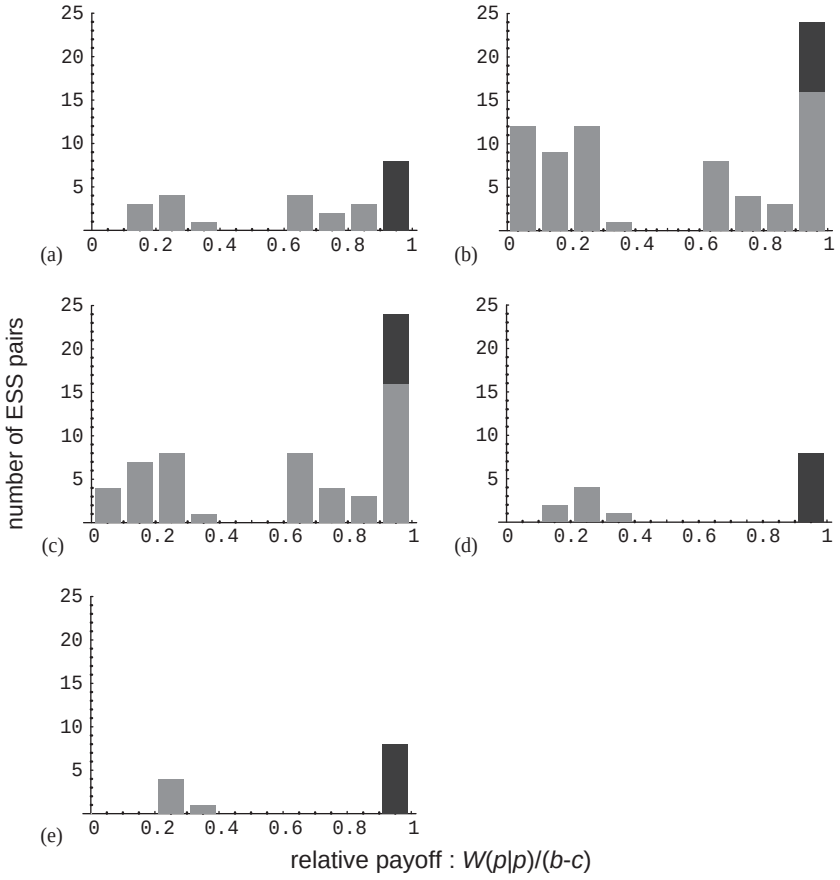

Fig. 2. Distribution of the relative payoff of ESS pairs, with different cost-benefit ratio. The horizontal axis represents the relative payoff, which is defined as the ratio of the average payoff to its theoretical maximum, $W(p|p)/(b - c)$. The vertical axis represents the number of ESS pairs. (a) Benefit of help is $b = 2$ and its cost is $c = 1$. If all in the population never failed to cooperate with one another, the relative payoff would be $W(p|p)/(b - c) = 1$ (perfectly cooperative population), which is the theoretical maximum. However, due to execition error and assignment error it is lower than 1 in general. We find 161 ESS pairs (When (d, p) is an ESS pair, then its mirror image (d', p') is also an ESS pair. In such a case we count only one of the two), behavioral strategies in 136 of which are All$D$-strategy. We pay attention to the other 25 ESS pairs, which are shown here. Roughly speaking there are two groups differing in the relative payoff, which is above 0.6 (60% of theoretical maximum) for the first group and is below 0.4 (below 40%) for the second group. Table 3 shows further details of this figure. The highest eight ESS pairs (colored in black in the figure) realize the relative payoff of more than 0.94, which is close to the theoretical maximum. We call those *leading eight* in the text. (b)–(e) Distributions of ESS pairs under various cost–benefit ratios. Cost is fixed at $c = 1$ throughout those figures whereas the benefit of help is (b) $b = 4$, (c) $b = 3$, (d) $b = 1.2$, and (e) $b = 1.1$, respectively. The number of ESS pairs except for All$D$-strategy pairs ($p = $ All$D$) decreases as the benefit declines; (b) 73 pairs, (c) 59 pairs, (d) 15 pairs, and (e) 13 pairs. The *leading eight* (in black in each figure), however, always rank the highest among them. Other parameters: $\mu_e = \mu_a = 0.02$.

Table 3
All ESS pairs that are non-trivial when $b = 2$ and $c = 1$

$(d_{11C}$	d_{11D}	d_{10C}	d_{10D}	d_{01C}	d_{01D}	d_{00C}	$d_{00D})$	$(p_{11}$	p_{10}	p_{01}	$p_{00})$	Relative payoff
(1	0	*	1	1	0	1	0)	(*C*	*D*	*C*	*C*)	0.943
(1	0	*	1	1	0	*	1)	(*C*	*D*	*C*	*D*)	0.942
(1	0	*	1	1	0	0	0)	(*C*	*D*	*C*	*D*)	0.940
………	………	………	………	………	………	………	………	………	………	………	………	………
(1	0	0	0	1	0	1	0)	(*C*	*D*	*C*	*C*)	0.838
(1	0	0	0	1	0	*	0)	(*C*	*D*	*C*	*D*)	0.809
(1	0	*	1	0	0	1	0)	(*C*	*D*	*D*	*C*)	0.705
(1	0	*	1	0	0	*	1)	(*C*	*D*	*D*	*D*)	0.680
(1	0	0	0	1	0	0	0)	(*C*	*D*	*C*	*D*)	0.331
(0	0	*	1	1	0	0	0)	(*D*	*D*	*C*	*D*)	0.244
(1	1	0	1	1	1	*	0)	(*D*	*C*	*D*	*D*)	0.232
(1	0	*	1	0	0	0	0)	(*C*	*D*	*D*	*D*)	0.170
(1	1	0	1	1	1	1	1)	(*D*	*C*	*D*	*D*)	0.101

We have 25 non-trivial ESS pairs (we omit their mirror images). The symbol '$*$' represents a wild card (i.e. either 1 or 0). Average payoff is calculated as $W(p|p)/(b-c)$.

other hand, the average payoff is zero in the group where nobody cooperates. Hence the relative payoff, which is defined as the ratio of average payoff to the maximum payoff, $W(p|p)/(b-c)$, can be used as a measure of cooperative tendency in the group. It is 1 for perfectly cooperative population and is 0 for All*D*-population. Due to errors, however, even for the most cooperative strategy the relative payoff is a little lower than 1. There are two large groups in the distribution. ESS pairs in one group have the relative payoff of around 0.2 and those in the second group are around 0.8. The most successful eight ESS pairs earn more than 0.94 (colored in black in Fig. 2(a)), next nine between 0.68 and 0.84, and the others less than 0.34. Here we pay attention to those best eight ESS pairs and call them '*leading eight*'.

Next, we examine effects of cost-benefit ratio. We change the benefit b while the cost c remains unchanged. Figs. 2(b)–(e) show the distributions of the relative payoff of ESS pairs. The parts colored in black in Fig. 2 represent the *leading eight*. Although the number of ESS pairs decreases as b declines, the *leading eight* not only keep their evolutionary stability but also always rank the highest. When the benefit is large enough (Fig. 2(b)), there are 73 non-trivial ESS pairs. As the efficiency of help declines, many ESS pairs except for the *leading eight* lose their evolutionary stability. At $b = 1.1$ (Fig. 2(e)), there remain only 13 ESS pairs, 8 of which are in the *leading eight* and they still yield large benefit whereas the other 5 less than 40% of the theoretical maximum. Compared with other ESS pairs, the *leading eight* have a remarkable robustness against the decline in the efficiency of help, provided $b > c$.

Changes in the error rates do not undermine their supremacy either. Fig. 3 show distributions of ESS pairs with various μ_e and μ_a values. The *leading eight* are evolutionarily stable regardless of error rates. Also they are always the highest in the rankings.

The *leading eight* share a few common characteristics. First, $d_{*1C} = 1$ and $d_{*1D} = 0$ (note that $*$ is a wild card) hold over the *leading eight*. The former equation states that if a player offers help to a *good* player, the donor will also be given good reputation. The latter notes that if a player refuses to help a *good* player, he will be assigned bad reputation. In short, behavior towards a *good* person critically affects the actor's reputation: good reputation to cooperation and bad reputation to defection. This is not true to behavior towards a *bad* person; sometimes cooperation towards a *bad* person yields bad reputation to the actor, or vice versa.

Second, $d_{10D} = 1$ always holds over the *leading eight*. This rule indicates that refusal of help against a *bad* individual does not undermine reputation of a good person. If we pay attention to the action of refusal itself, it may be socially bad. However, not cooperating with a bad person would be helpful for sustaining discriminate altruism in a society: if individuals were to donate even to social parasites, the cooperative relationship would be lost easily. Hence, a society as a whole should regard such refusal as a good action. This realizes the notion of sanction. Even defection itself is beneficial to the actor himself, in other words, even if the refusal saves a player the cost of cooperation, others in a society must not blame him for his defection, provided that it is aimed at a *bad* person. In contrast, defection towards a *good* player should of course be blamed, which is stated by $d_{11D} = d_{01D} = 0$.

5. Conditions for sustained cooperation

Mathematical analysis supports the conclusion from numerical search in the last section. First, we can derive the following inequality as the condition of evolutionary

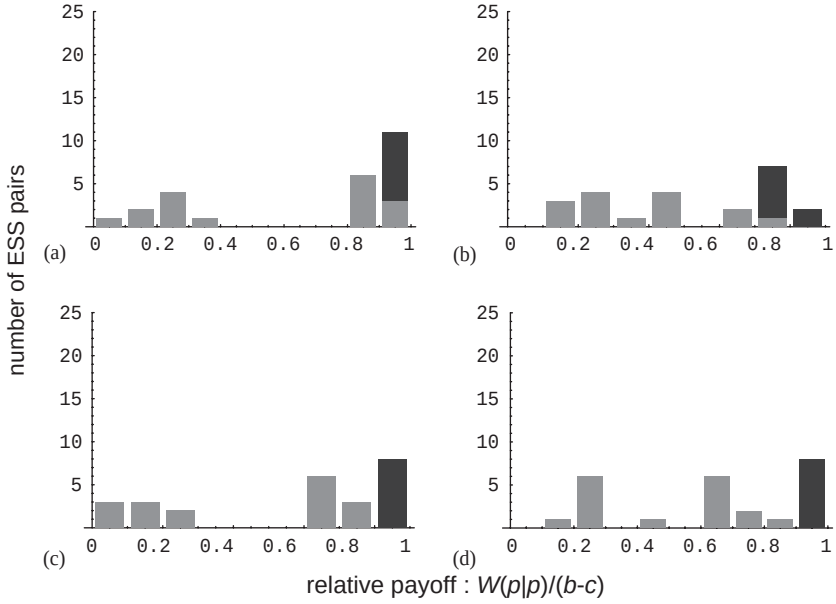

Fig. 3. Distribution of the relative payoff of ESS pairs, with different error rates, μ_e and μ_a. The horizontal axis represents the relative payoff, $W(p|p)/(b-c)$. The vertical axis represents the number of ESS pairs. We set $b=2$ and $c=1$ in these figures. The part colored in black represents the *leading eight*. (a) When error rates are too small ($\mu_e=\mu_a=0.005$). We find 25 ESS pairs, which are the same as those in Fig. 2(a). The relative payoff of the *leading eight* is above 0.985. (b) When errors frequently occur ($\mu_e=\mu_a=0.08$). We find 23 ESS pairs. Although the relative payoff of the *leading eight* is around 0.8, the *leading eight* keep their evolutionary stability and remain the highest in the ranking. (c),(d) When one of two error rates are large compared with the other ((c) $\mu_e=0.05$ and $\mu_a=0.01$, (d) $\mu_e=0.01$ and $\mu_a=0.05$). In both cases we find 25 ESS pairs, which are the same as those in Fig. 2(a). We see that the *leading eight* are still the best.

stability of each in the *leading eight* against All*D*:

$$\frac{b}{c}>1+O(\mu), \tag{11}$$

where μ is a parameter of the same order of magnitude as μ_e and μ_a (see Appendix B for derivation). This tells us that cooperation by the *leading eight* is evolutionarily stable whenever b is a little greater than c, if errors occur only infrequently. Second, the average payoff in the population where everyone adopts one of the leading eight is calculated as

$$W(p|p)=(b-c)\{1-O(\mu)\}. \tag{12}$$

The leading eight can maintain almost maximum average payoff in the society they dominate. We find that Eqs. (11) and (12) hold simultaneously only for the *leading eight*. ESS pairs other than the *leading eight* do not have this good property. Hence, we conclude that the *leading eight* are the most suitable for sustaining cooperation, especially when the yield of altruism is positive but tiny.

According to the characteristic of ESS pairs, we classify the *leading eight* into three groups, named Group I, Group II, and Group III, respectively, as in Table 4.

Two ESS pairs belong to Group I. Note that one of these, $(d,\,p)=(\mathsf{STAND},\,\mathsf{OR})$, corresponds to the standing strategy (Leimar and Hammerstein, 2001; Panchanathan and Boyd, 2003; Sugden, 1986). They are characterized by $p=\mathsf{OR}$, $d_{00C}=1$, and $d_{00D}=0$. First, $p=\mathsf{OR}$ suggests that players are interested in their own reputation. They try to enhance their reputation when they are with bad reputation. Rules $d_{00C}=1$ and $d_{00D}=0$ say that a bad individual can regain his good reputation by cooperating with cheaters, who are almost always with bad reputation. In such a society, cooperation done by a bad individual is regarded as a good action even if it is aimed at social parasites. Taylor expansion with respect to two error rates μ_e and μ_a shows that strategies in Group I are stable against invasions by All*D* if

$$\frac{b}{c}>1+(\mu_e+2\mu_a)+(\mu_e^2+4\mu_e\mu_a+6\mu_a^2)+\cdots, \tag{13}$$

which is the most strict condition among the three groups, and their average payoff is given by

$$(b-c)\{1-(2\mu_e+\mu_a)+(3\mu_e^2+6\mu_e\mu_a+\mu_a^2)+\cdots\}, \tag{14}$$

which is the largest among the three.

Four ESS pairs belong to Group II. They are characterized by $d_{00D}=1$. In a society dominated by them, defection done by a bad person against social parasites is regarded as good. It is true that those four ESS pairs are stable against a few All*D*-strategists, but they may be susceptible to a cluster of All*D*-strategists since defection between two cheaters is considered good under those rules. Taylor expansion with respect to two

Table 4
The *leading eight*

Group	$(d_{11C}$	d_{11D}	d_{10C}	d_{10D}	d_{01C}	d_{01D}	d_{00C}	$d_{00D})$	$(p_{11}$	p_{10}	p_{01}	$p_{00})$
Group I	(**1**	**0**	1	**1**	**1**	**0**	1	0)[a]	(*C*	*D*	*C*	*C*)[d]
	(**1**	**0**	0	**1**	**1**	**0**	1	0)	(*C*	*D*	*C*	*C*)[d]
Group II	(**1**	**0**	1	**1**	**1**	**0**	1	1)	(*C*	*D*	*C*	*D*)[e]
	(**1**	**0**	1	**1**	**1**	**0**	0	1)	(*C*	*D*	*C*	*D*)[e]
	(**1**	**0**	0	**1**	**1**	**0**	1	1)	(*C*	*D*	*C*	*D*)[e]
	(**1**	**0**	0	**1**	**1**	**0**	0	1)	(*C*	*D*	*C*	*D*)[e]
Group III	(**1**	**0**	1	**1**	**1**	**0**	0	0)[b]	(*C*	*D*	*C*	*D*)[e]
	(**1**	**0**	0	**1**	**1**	**0**	0	0)[c]	(*C*	*D*	*C*	*D*)[e]

[a]Standing $d = \mathsf{STAND}$.
[b]Strict-standing $d = \mathsf{S\text{-}STAND}$.
[c]Judging $d = \mathsf{JUDGE}$.
[d]Or-strategy $p = \mathsf{OR}$.
[e]Co-strategy $p = \mathsf{CO}$. We omit their mirror images.

error rates μ_e and μ_a shows that strategies in Group II are stable against invasions by All*D* if

$$\frac{b}{c} > 1 + (\mu_e + 2\mu_a) + (\mu_e^2 + 2\mu_e\mu_a + 4\mu_a^2) + \cdots, \qquad (15)$$

and their average payoff is given by

$$(b-c)\{1 - (2\mu_e + \mu_a) + (2\mu_e^2 + 4\mu_e\mu_a) + \cdots\}. \qquad (16)$$

The other two ESS pairs are in Group III. We see the strict-standing $d = \mathsf{S\text{-}STAND}$ in this group. The group is characterized by $d_{00C} = d_{00D} = 0$. It says that a bad player can never regain his good reputation through social interactions with other bad persons. The only way for him to acquire a good reputation is to meet a good person and helps him ($d_{01C} = 1$). Once one falls into bad reputation, it is hard for him to recover his previous social state. Hence, those rules in Group III are the most strict against bad players among the three groups; good social reputation is highly valuable there. Taylor expansion with respect to two error rates μ_e and μ_a shows that strategies in Group III are stable against invasions by All*D* if

$$\frac{b}{c} > 1 + (\mu_a) + (2\mu_e\mu_a + 3\mu_a^2) + \cdots, \qquad (17)$$

which is the weakest condition among the three groups, and their average payoff is given by

$$(b-c)\{1 - (2\mu_e + \mu_a) + (\mu_e^2 + 2\mu_e\mu_a - \mu_a^2) + \cdots\}, \qquad (18)$$

which is the smallest among the three.

6. Discussion

Altruism brings welfare at the sacrifice of its donors. Seen from the benefit of a group, it should be greatly encouraged. For donors, however, it is only costly. There is a conflict between group's interest and that of an individual. Since it is not a group but an individual that natural selection works on, everyone tries not to pay any costs, which leads to the situation where no one is motivated at all to participate in cooperation. If return for the help were always assured, people would be willing to pay the cost of help for they can surely acquire the benefit which exceeds the cost. But it is impossible in nature without such powerful deterrence as legal institutions.

Indirect reciprocity gives us an affirmative answer to this issue. If discriminate altruists, who carefully distinguish individuals by their social reputation, selectively choose potential receivers of their help, it becomes possible to build a society based on altruism without any external enforcement. We stress that indiscriminate altruists, who help everyone, will never achieve such a purpose because they are easily exploited by social parasites. Conditional cooperation is necessary. That is why the *leading eight* have conditional strategies such as $p = \mathsf{CO}$ and OR.

However, the above discussion misses an important question; what *goodness* is. As Fehr and Fischbacher (2003) mention, the notion of goodness in social reputation is one of the most fundamental issues in indirect reciprocity. Our intuition tells us that all cooperators are good and all defectors are bad, as recommended by the image-score criterion $d = \mathsf{IMAGE}$. Previous studies (Leimar and Hammerstein, 2001; Ohtsuki, 2004; Panchanathan and Boyd, 2003) have, however, shown that this simple criterion does not work well for maintaining general exchange of altruism: things are not so simple. As Nowak and Sigmund (1998a) mentioned in their discussion, if every defection is regarded as bad a player who happens to meet a social parasite will not want to refuse help, otherwise his reputation would be undermined. This leads to the

standing criterion (Sugden, 1986), which justifies defection against bad persons as punishment. Some have revealed that the standing is important (Leimar and Hammerstein, 2001; Panchanathan and Boyd, 2003).

Takahashi and Mashima (2003) broadened their scope to criteria other than image-score or standing. They assumed that a player's reputation is determined by (i) his action (first-order information) and (ii) his opponent's reputation (second-order information). Hence in their model there are 16 reputation dynamics (though they called those 'strategies'). Using mixed-population model and island model they showed that standing strategy is highly susceptible to misperception. They casted serious doubts on the effectiveness of the standing strategy.

Although several previous studies have examined the performance of a couple of reputation dynamics, no one has systematically studied evolutionary plausibility of the notion of goodness so far. For that reason we studied all possible reputation dynamics exhaustively. To fulfill our purpose, we had to first exclude subjectivity or bias we have as to the notion of goodness. We introduced a simple binary digit called Honor-score as a substitute for *good* or *bad*. *H*-score system makes it possible for us to study and compare a large number of reputation dynamics. We believe that this is the first paper that the notion of goodness, namely who should be regarded good in social exchange, is studied systematically as evolutionary games.

6.1. Leading eight

Our ESS analysis has shown that the *leading eight* are the most plausible candidates which have contributed to the evolution of indirect reciprocity. Even when there are errors of various kinds and when the benefit of help exceeds its cost only by a small margin, a society based on the *leading eight* can maintain social profits stably; those who help others surely receive greater return. In contrast, when the benefit b is much greater than the cost c we have a number of ESS pairs (see Fig. 2(b)). Several reputation dynamics can secure indirect reciprocity if the help yields much.

Behavioral strategies in the *leading eight* are either $p = \mathsf{OR}$ or CO. For reputation dynamics in Group I, behavioral strategy $p = \mathsf{OR}$, which considers reputation of its own, does slightly better than $p = \mathsf{CO}$. This result agrees with what Panchanathan and Boyd (2003) discussed. We have confirmed that even with assignment error (that is similar to 'perception error' (Leimar and Hammerstein, 2001)) does the behavioral strategy $p = \mathsf{OR}$ earn much benefit than $p = \mathsf{CO}$. In contrast, for reputation dynamics in Group II and III, we find that the simple behavioral strategy $p = \mathsf{CO}$ is evolutionarily stable. This result for ESSs tells us that not only the All*C*-strategy but also those which care only about the reputation of themselves cannot achieve steady social exchange. In addition, our result has shown that such queer behavioral strategies as 'cooperate only if reputation of both players agree' or 'cooperate only if reputation of both players disagree' are hopeless.

Reputation dynamics within ESS pairs in the *leading eight* bring fruitful perspective on the notion of goodness. The *leading eight* have roughly speaking two features in common. One is that a player interacting with good persons are assessed by what he does. Cooperation with good individuals should be good and defection against good ones should be bad. The second feature should we consider with much emphasis: a good player who refused to help a bad person must be labeled good. This enables players facing cheaters to refuse help without worrying about the influence of the action on their own good reputation. Defection of this kind should be regarded as righteous and retaliatory, not selfish. It is because the image-score criterion does not have this property that the society based on that fails to maintain indirect reciprocity. It is true, however, that players adopting the *leading eight* should be required high cognitive ability. We expect that indirect reciprocity has evolved owing to several capabilities of humans.

6.2. Indirect observation model vs. direct observation model

There is difference between indirect observation model studied in the present paper and direct observation model by Leimer and Hammerstein (2001) concerning the way to obtain social information of others. These two models are the same when there is no error in the assignment of *H*-score. However, the existence of assignment error drastically changes evolutionary stability of the *leading eight*.

Consider, for example, the stability of the standing strategy, $(d, p) = (\mathsf{STAND}, \mathsf{OR})$ (Leimar and Hammerstein, 2001; Panchanathan and Boyd, 2003; Sugden, 1986). For $(d, p) = (\mathsf{STAND}, \mathsf{OR})$ to be an ESS pair, it is necessary that 'Or-strategy' ($p = \mathsf{OR}$) should be stable against the invasion by All*C*-strategy ($p = \mathrm{All}C$). This holds for indirect observation model studied in the present paper. However, in direct observation model it is no more true and the standing strategy is not an ESS pair. The reason is given as follows.

Imagine a population in which Or-strategists are common and there are a few All*C*-mutants. Or-strategist and All*C*-strategist take different actions only when they meet an opponent whose *H*-score is 0; Or-strategist defects against him whereas All*C*-strategist cooperates. Note that the opponent's *H*-score of 0 is always due to errors, because he is usually cooperative. Difference in payoffs between two strategies results from errors.

In indirect observation model, H-score of a focal person is common knowledge among players in the population; all have the same opinion on him. Therefore, when a player E_1 is mistakenly assigned bad reputation (H-score $=0$) by an observer, everyone mistakenly believes that E_1's H-score is 0, though he should have assigned 1. In this case, Or-strategist outcompetes All$C$-strategist, because defection aimed at E_1 by Or-strategist is regarded as good ($d_{10D}=1$) thus he can save the cost of help ($+c$) without undermining his good reputation.

In contrast, each may have different opinion on a person in direct observation model. In this model, each player has his subjective H-score towards each person. H-score is no more shared among individuals. Therefore, even when a player mistakenly assigns (his subjective) H-score of 0 to player E_1 due to assignment error, others will not agree with his opinion. The Or-strategist's belief that E_1 is not worth cooperating is not approved by others, so he will miss the future cooperation ($-b$) in exchange for the refusal of the help ($+c$). Because $-b+c<0$ holds, Or-strategists are invaded by All$C$-mutants.

The evolutionary outcomes of two models differ in the presence of errors. Which is more plausible may depend on the situation considered. We think that indirect observation model may be more suitable than the direct observation model in human interactions. Fluidity of language must have greatly contributed to the evolution of indirect reciprocity (Nakamaru and Kawata, 2004). Thanks to the convenience and high availability of language as a source of information, each saves the cost of monitoring others and keeps aware of social information, thus it results in robust cooperative relationships. In contrast, a player in the direct observation model observes every social interaction by himself. We predict such an excessive investment to monitoring costs makes indirect reciprocity difficult to be maintained in a large group. The difference between indirect observation and direct observation suggests a merit of human language. On the other hand, direct observation model may be more suitable for a society of non-human animals where language is not available.

6.3. When reputation dynamics differ

In this paper, players have different behavioral strategies but they all have the same reputation dynamics. Instead, we may consider an evolutionary game among players which differ in both reputation dynamics and behavioral strategy. However, such a formalism is not necessarily wider in scope than the game in which only behavioral strategy differs with reputation dynamics fixed, as shown below. The strategy is then a pair of (d,p). The condition for invasibility should be calculated from the fitness of a mutant actor (d', p') in a population dominated by residents (d, p). First, we consider behavioral strategies are different ($p'\neq p$). In indirect observation model, reputation is controlled by observers. The fitness of the rare mutant depends on the opponents and the observers, which are all of dominant type (d, p). To calculate the fitness of the invader, we need the behavioral strategy of the mutant p', that of the opponent p, and the reputation dynamics of the observer d. Note that the reputation dynamics of the rare mutant d' has no effect to the mutant's fitness. As a result, the invasibility of (d', p') is exactly the same as the invasibility of a mutant p' playing with the opponent p when reputation dynamics is fixed as d. Hence, the leading eight are the only possible ESSs even when we regard (d, p) as a strategy.

This argument excludes the case in which mutant and residents have exactly the same behavioral strategy p but differ in reputation dynamics ($d'\neq d$). In such a case, the mutant is perfectly neutral. Mutant frequency is controlled only by random genetic drift. A direct computer simulation might show an apparent coexistence of multiple reputation dynamics in the population, which is in fact a neutral mixture. However, this does not imply the absence of selection on reputation dynamics in real societies. We suspect that reputation dynamics in reality are very likely to be determined by group selection—a reputation dynamics that realizes a society with a higher cooperation level is able to spread and to drive away alternatives that realize a society with a lower level of cooperation. If so, discussing the evolution of the moral judgment in a game with (d, p) as a 'strategy' can be misleading, because it implicitly assumes that the reputation dynamics should be formed by intra-group selection.

Hence, treating reputation dynamics d as a fixed rule and considering a game between different behavioral strategies under the rule d is a valid approach, which is exactly what we have done in our paper.

After we submitted our paper, we were informed that Hannelore Brandt and Karl Sigmund had developed a similar idea independently and also submitted their paper to this journal almost simultaneously (Brandt and Sigmund, 2004; published in the same issue). In the following, we briefly explain the major differences between two works. (1) Brandt and Sigmund adopted direct observation model as a manner of introducing errors. Hence, in their model different players may have different 'score' of the same person. Players do not exchange the information on the 'score' itself. In contrast our H-score is 'reputation' formed in social exchange of information among players. (2) Brandt and Sigmund developed an individual based computer simulation model. They focused 3 reputation dynamics (they called 'assessment modules') ($d=$ IMAGE, STAND, JUDGE) and 6 behavioral strategies (they called 'action modules')

($p = $ CO, SELF, AND, OR, All$C$, All$D$). Since All$C$ and All$D$ do not need reputation dynamics, they ran games among $3 \times 4 + 2 = 14$ combinations. In contrast, we have examined all 4096 possible pairs and derived ESS condition mathematically. As a result we succeeded in discovering eight successful pairs of reputation dynamics and behavioral strategies. (3) There are differences in terminology, some of which come naturally from the difference in the basic idea of the two models.

Acknowledgements

We thank Hannelore Brandt, Toshikazu Hasegawa, Rie Mashima, Martin Nowak, Karl Sigmund, and Nobuyuki Takahashi for their fruitful comments.

Appendix A. Mirror symmetry

There exist $2^{12} = 4096$ pairs (d, p) to consider in our framework. However, switching 1 and 0 in H-score causes mirror symmetry among them. If we completely switch 1 and 0, a pair represented by (d, p) becomes equivalent to the pair (d', p'), where

$$\begin{aligned} d'_{11X} &= 1 - d_{00X} \\ d'_{10X} &= 1 - d_{01X} \\ d'_{01X} &= 1 - d_{10X} \\ d'_{00X} &= 1 - d_{11X} \\ p'_{11} &= p_{00} \\ p'_{10} &= p_{01} \\ p'_{01} &= p_{10} \\ p'_{00} &= p_{11}. \end{aligned} \qquad (X = C, D) \tag{A.1}$$

In other words, (d', p') is an mirror image of (d, p). The mirror image shares the same properties as an original one. Hence, we can examine only one pair of each symmetric dyad.

A pair (d, p) is symmetric to itself when

$$\begin{aligned} d_{11X} &= 1 - d_{00X} \\ d_{10X} &= 1 - d_{01X} \\ p_{11} &= p_{00} \\ p_{10} &= p_{01} \end{aligned} \qquad (X = C, D) \tag{A.2}$$

holds. Since $2^6 = 64$ cases satisfy Eq. (A.2), the total number of effective pairs that need to be studied is now $2^6 + (2^{12} - 2^6)/2 = 2080$.

Appendix B. Analytical solution of Eq. (4)

In this section, we prove that (i) both $h_t(p)$ and $h_t(p')$ converge to stationary values $h_*(p)$ and $h_*(p')$, respectively, in the limit of $t \to \infty$ and (ii) those limits are independent of their initial ones $h_0(p)$ and $h_0(p')$.

Let us begin the proof. Define $D_{ij}(p')$ as

$$D_{ij}(p') \equiv (1 - 2\mu_a)\{(1 - \mu_e)d_{ijp'_{ij}} + \mu_e d_{ijD}\} + \mu_a \quad (i, j = 1, 0). \tag{B.1}$$

We easily see that $0 < \mu_a \leqq D_{ij}(p') \leqq 1 - \mu_a < 1$ holds. Then Eq. (4) becomes

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} h_t(p') = {} & h_t(p')h_t(p)D_{11}(p') + h_t(p')(1 - h_t(p))D_{10}(p') \\ & + (1 - h_t(p'))h_t(p)D_{01}(p') + (1 - h_t(p')) \\ & \times (1 - h_t(p))D_{00}(p') - h_t(p'). \end{aligned} \tag{B.2}$$

Substituting p for p' in Eq. (B.2) yields

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} h_t(p) &= A_2(p)\{h_t(p)\}^2 + A_1(p)h_t(p) + A_0(p) \\ &\equiv I(h_t(p)), \end{aligned} \tag{B.3}$$

where

$$A_2(p) = D_{11}(p) - D_{10}(p) - D_{01}(p) + D_{00}(p), \tag{B.4a}$$

$$A_1(p) = D_{10}(p) + D_{01}(p) - 2D_{00}(p) - 1, \tag{B.4b}$$

$$A_0(p) = D_{00}(p). \tag{B.4c}$$

Here we note the following proposition.

Proposition B.1. *Consider the dynamical system over* $[0, 1]$, *given by*

$$\frac{\mathrm{d}}{\mathrm{d}t} x = F(x) = ax^2 + bx + c \tag{B.5}$$

If $F(0) > 0$ *and* $F(1) < 0$ *hold, the limit* $x^* = \lim_{t\to\infty} x(t)$ *always uniquely exists and it is independent of the initial value* $x(t = 0)$. *To put it precisely,* x^* *is given as follows*:

$$x^* = \begin{cases} \frac{-b - \sqrt{b^2 - 4ac}}{2a} & \text{if} \quad a \neq 0, \\ -\frac{c}{b} & \text{if} \quad a = 0. \end{cases} \tag{B.6}$$

This can be proved by standard analysis of differential equations.

We note that $I(0) = D_{00}(p) \geqq \mu_a > 0$ and $I(1) = D_{11}(p) - 1 \leqq (1 - \mu_a) - 1 = -\mu_a < 0$. Using these facts and Proposition B.1, we conclude that the value $h_*(p) = \lim_{t\to\infty} h_t(p)$ exists and that it is independent of $h_0(p)$.

Let us consider $h_t(p')$ next. Eq. (B.2) is rewritten as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t} h_t(p') = -B_{1,t}(p, p')h_t(p') + B_{0,t}(p, p'), \tag{B.7}$$

where

$$B_{1,t}(p, p') = 1 - \{h_t(p)D_{11}(p') + (1 - h_t(p))D_{10}(p')\} + \{h_t(p)D_{01}(p') + (1 - h_t(p))D_{00}(p')\}, \quad \text{(B.8a)}$$

$$B_{0,t}(p, p') = h_t(p)D_{01}(p') + (1 - h_t(p))D_{00}(p'). \quad \text{(B.8b)}$$

When time t is large, $h_t(p)$ approaches $h_*(p)$, so Eq. (B.7) becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}h_t(p') = -B_1(p, p')h_t(p') + B_0(p, p'), \quad \text{(B.9)}$$

where

$$B_1(p, p') = 1 - \{h_*(p)D_{11}(p') + (1 - h_*(p))D_{10}(p')\} + \{h_*(p)D_{01}(p') + (1 - h_*(p))D_{00}(p')\}, \quad \text{(B.10a)}$$

$$B_0(p, p') = h_*(p)D_{01}(p') + (1 - h_*(p))D_{00}(p'). \quad \text{(B.10b)}$$

We see that $0 < B_0(p, p') < B_1(p, p')$ holds. Therefore, from Eq. (B.9) we have

$$\lim_{t\to\infty} h_t(p') = \frac{B_0(p, p')}{B_1(p, p')} \quad (\equiv h_*(p')). \quad \text{(B.11)}$$

Obviously, $h_*(p')$ is independent of its initial value $h_0(p')$. This ends the proof.

Since all the calculations above are analytic, the values of $h_*(p)$ and $h_*(p')$ can be expressed in terms of seven parameters d, p, p', b, c, μ_e and μ_a. Hence, we have the Taylor expansion of the payoff with respect to error rates μ_e and μ_a. Because their expressions are very complicated and messy, we do not show them here. Setting $p' = \mathrm{All}D$ and $\mu_e = \mu_a = \mu$, we have Eqs. (11) and (12). Eqs. (13) to (18) are derived in a similar manner.

References

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

Brandt, H., Hauert, C., Sigmund, K., 2003. Punishment and reputation in spatial public goods games. Proc. R. Soc. London B 270, 1099–1104.

Brandt, H., Sigmund, K., 2004. The logic of reprobation: assessment and action rules for indirect reciprocation. J. Theor. Biol., this issue.

Enquist, M., Leimar, O., 1993. The evolution of cooperation in mobile organisms. Anim. Behav. 45, 747–757.

Fehr, E., Fischbacher, U., 2003. The nature of altruism. Nature 425, 785–791.

Fehr, E., Gächter, S., 2000. Cooperation and punishment in public goods experiments. Am. Econ. Rev. 90, 980–994.

Fehr, E., Rockenbach, B., 2003. Detrimental effects of sanctions on human altruism. Nature 422, 137–140.

Fishman, M., 2003. Indirect reciprocity among imperfect individuals. J. Theor. Biol. 225, 285–292 (doi:10.1006/S0022-5193(03)00246-7).

Henrich, J., Boyd, R., 2001. Why people punish defectors; weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. J. Theor. Biol. 208, 79–89 (doi:10.1006/jtbi.2000.2202).

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

Lotem, A., Fishman, M., Stone, L., 1999. Evolution of cooperation between individuals. Nature 400, 226–227.

Nakamaru, M., Kawata, M., 2004. Evolution of rumors that discriminate lying defectors. Evol. Ecol. Res. 6, 261–283.

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

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

Ohtsuki, H., 2004. Reactive strategies in indirect reciprocity. J. Theor. Biol. 227, 299–314 (doi:10.1016/j.jtbi.2003.11.008).

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

Pollock, K., Dugatkin, L., 1992. Reciprocity and the emergence of reputation. J. Theor. Biol. 159, 25–37.

Sherratt, T., Roberts, G., 2001. The importance of phenotypic defectors in stabilizing reciprocal altruism. Behav. Ecol. 12 (3), 313–317.

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

Takahashi, N., Mashima, R., 2003. The emergence of indirect reciprocity: is the standing strategy the answer? Center for the study of cultural and ecological foundations of the mind, Hokkaido University, Japan, Working paper series No. 29.

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