Ingroup favoritism and intergroup cooperation under indirect reciprocity based on group reputation

Naoki Masuda [a,b,*]

[a] Department of Mathematical Informatics, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-8656, Japan
[b] PRESTO, Japan Science and Technology Agency, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan

HIGHLIGHTS

► I study a model of ingroup favoritism based on indirect reciprocity.
► Reputation values are assigned to groups as well as to individuals.
► I reveal the conditions for ingroup favoritism and perfect cooperation.

ARTICLE INFO

ABSTRACT

Indirect reciprocity in which players cooperate with unacquainted other players having good reputations is a mechanism for cooperation in relatively large populations subjected to social dilemma situations. When the population has group structure, as is often found in social networks, players in experiments are considered to show behavior that deviates from existing theoretical models of indirect reciprocity. First, players often show ingroup favoritism (i.e., cooperation only within the group) rather than full cooperation (i.e., cooperation within and across groups), even though the latter is Pareto efficient. Second, in general, humans approximate outgroup members' personal characteristics, presumably including the reputation used for indirect reciprocity, by a single value attached to the group. Humans use such a stereotypic approximation, a phenomenon known as outgroup homogeneity in social psychology. I propose a model of indirect reciprocity in populations with group structure to examine the possibility of ingroup favoritism and full cooperation. In accordance with outgroup homogeneity, I assume that players approximate outgroup members' personal reputations by a single reputation value attached to the group. I show that ingroup favoritism and full cooperation are stable under different social norms (i.e., rules for assigning reputations) such that they do not coexist in a single model. If players are forced to consistently use the same social norm for assessing different types of interactions (i.e., ingroup versus outgroup interactions), only full cooperation survives. The discovered mechanism is distinct from any form of group selection. The results also suggest potential methods for reducing ingroup bias to shift the equilibrium from ingroup favoritism to full cooperation.

1. Introduction

Humans and other animals often show cooperation in social dilemma situations, in which defection apparently seems more lucrative than cooperation. A main mechanism governing cooperation in such situations is direct reciprocity, in which the same pairs of players repeatedly interact to realize mutual cooperation (Trivers, 1971; Axelrod, 1984; Nowak, 2006a). In fact, individuals who do not repeatedly interact also cooperate with others. In this

situation, reputation-based indirect reciprocity, also known as downstream reciprocity, is a viable mechanism for cooperation (Nowak and Sigmund, 1998a; Leimar and Hammerstein, 2001; Ohtsuki and Iwasa, 2004, 2007; Nowak and Sigmund, 2005; Brandt and Sigmund, 2005, 2006). In this mechanism, which I refer to as indirect reciprocity for simplicity, individuals carry their own reputation scores, which represent an evaluation of their past actions toward others. Individuals are motivated to cooperate to gain good reputations so that they are helped by others in the future or to reward (punish) good (bad) others. Indirect reciprocity facilitates cooperation in a larger population than in the case of direct reciprocity because unacquainted players can cooperate with each other. Although evidence of indirect reciprocity is relatively scarce for nonhumans (but see

* Correspondence address: Department of Mathematical Informatics, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-8656, Japan.
Tel.: +81 3 5841 6931; fax: +81 3 5841 6931.
E-mail address: masuda@mist.i.u-tokyo.ac.jp

Bshary and Grutter, 2006), it is widely accepted as explanation for cooperation in humans (Nowak and Sigmund, 2005).

Humans, in particular, belong to groups identified by traits, such as age, ethnicity, and culture. Individuals presumably interact more frequently with ingroup than outgroup members. Group structure has been a main topic of research in social psychology and sociology for many decades (Brown, 2000; Dovidio et al., 2005) and in network science (Fortunato, 2010). Experimental evidence suggests that, when the population of players has group structure, two phenomena that are not captured by existing models of indirect reciprocity take place.

First, in group-structured populations, humans (Sedikides et al., 1998; Brewer, 1999; Hewstone et al., 2002; Dovidio et al., 2005; Efferson et al., 2008) and even insect larvae (Lize et al., 2006) show various forms of ingroup favoritism. In social dilemma games, individuals behave more cooperatively toward ingroup than outgroup members (e.g., De Cremer and van Vugt, 1999; Goette et al., 2006; Fowler and Kam, 2007; Rand et al., 2009; Yamagishi et al., 1998, 1999; Yamagishi and Mifune, 2008). Ingroup favoritism in social dilemma situations may occur as a result of indirect reciprocity confined in the group (Yamagishi et al., 1998, 1999; Yamagishi and Mifune, 2008). In contrast, ingroup favoritism in social dilemma games is not Pareto efficient because individuals would receive larger payoffs if they also cooperated across groups. Under what conditions are ingroup favoritism and intergroup cooperation sustained by indirect reciprocity? Can they be bistable?

Ingroup favoritism, which has also been analyzed in the context of tag-based cooperation, the green beard effect, and the armpit effect, has been considered to be a theoretical challenge (e.g., Antal et al., 2009). Nevertheless, recent research has revealed their mechanisms, including the loose coupling of altruistic trait and tag in inheritance (Jansen and van Baalen, 2006), a relatively fast mutation that simultaneously changes strategy and tag (Traulsen and Nowak, 2007; Traulsen, 2008), a tag's relatively fast mutation as compared to the strategy's mutation (Antal et al., 2009) conflicts between groups (Choi and Bowles, 2007; García and van den Bergh, 2011), partial knowledge of others' strategies (Masuda and Ohtsuki, 2007), and gene-culture coevolution (Ihara, 2011). However, indirect reciprocity accounts for ingroup favoritism, as is relevant to previous experiments (Yamagishi et al., 1998, 1999; Yamagishi and Mifune, 2008) is lacking.

Second, in a population with group structure, individuals tend to approximate outgroup individuals' characteristics by a single value attached to the group. This type of stereotype is known as outgroup homogeneity in social psychology (Jones et al., 1981; Ostrom and Sedikides, 1992; Sedikides et al., 1998; Brown, 2000), and it posits that outgroup members tend to be regarded to resemble each other more than they actually do. It is also reasonable from the viewpoint of cognitive burden of remembering each individual's properties that humans generally resort to outgroup homogeneity. Therefore, in indirect reciprocity games in group structured populations, it seems to be natural to assume outgroup homogeneity. In other words, individuals may not care about or have access to personal reputations of those in different groups and approximate an outgroup individual's reputation by a group reputation.

Some previous models analyzed the situations in which players do not have access to individuals' reputations. This is simply because it may be difficult for an individual in a large population to separately keep track of other people's reputations even if gossiping helps dissemination of information. This case of incomplete information has been theoretically modeled by introducing the probability that an individual sees others' reputations in each interaction (Nowak and Sigmund, 1998b, 1998a; Brandt

and Sigmund, 2005, 2006; Suzuki and Toquenaga, 2005; Nakamura and Masuda, 2011). However, these studies do not have to do with the approximation of individuals' personal reputations by group reputations.

By analyzing a model of an indirect reciprocity game based on group reputation, I provide an indirect reciprocity account for ingroup favoritism for the first time. In addition, through an exhaustive search, I identify all the different types of stable homogeneous populations that yield full cooperation (intragroup and intergroup cooperation) or ingroup favoritism.

2. Methods

2.1. Model

2.1.1. Population structure and the donation game

I assume that the population is composed of infinitely many groups each of which is of infinite size. Each player belongs to one group.

Players are involved in a series of the donation game, which is essentially a type of prisoner's dilemma game. In each round, a donor and recipient are selected from the population in a completely random manner. Each player is equally likely to be selected as donor or recipient. The donor may refer to the recipient's reputation and select one of the two actions, cooperation (C) or defection (D). If the donor cooperates, the donor pays cost $c > 0$, and the recipient receives benefit $b(>c)$. If the donor defects, the payoffs to the donor and recipient are equal to 0. Because the roles are asymmetric in a single game, the present game differs from the one-shot or standard iterated versions of the prisoner's dilemma game. This game is widely used for studying mechanisms for cooperation including indirect reciprocity (Nowak and Sigmund, 2005; Nowak, 2006a, 2006b).

Rounds are repeated a sufficient number of times with different pairs of donors and recipients. Because the population is infinite, no pair of players meets more than once, thereby avoiding the possibility of direct reciprocity (e.g., Nowak and Sigmund, 1998a; Ohtsuki and Iwasa, 2004). The payoff to each player is defined as the average payoff per round.

The groups to which the donor and recipient belong are denoted by g_d and g_r, respectively. The simultaneously selected donor and recipient belong to the same group with probability r^{in} (i.e., $g_d = g_r$; Fig. 1A) and different groups with probability $r^{out} \equiv 1 - r^{in}$ (i.e., $g_d \neq g_r$; Fig. 1B).
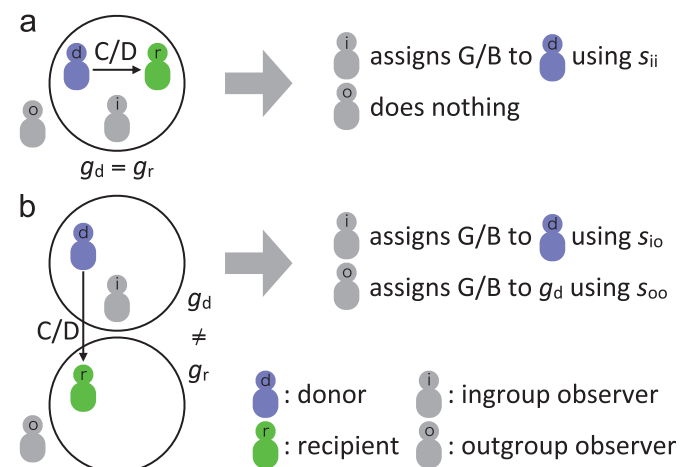

Fig. 1. Schematic representation of ingroup and outgroup observers. In A, the donor's group g_d and the recipient's group g_r are identical. This event occurs with probability r^{in}. In B, $g_d \neq g_r$. This event occurs with probability $r^{out} = 1 - r^{in}$.

2.1.2. Social norms

At the end of each round, observers assign binary reputations, good (G) or bad (B), to the donor and donor's group (g_d) according to a given social norm. I consider up to the so-called second-order social norms with which the observers assign G or B as a function of the donor's action and the reputation (i.e., G or B) of the recipient or recipient's group (g_r). Representative second-order social norms are shown in Fig. 2. Under image scoring ("scoring" in Fig. 2), an observer regards a donor's action C or D to be G or B, respectively, regardless of the recipient's reputation. In the absence of a group-structured population, scoring does not realize cooperation based on indirect reciprocity unless certain specific conditions are met (Nowak and Sigmund, 1998a; Brandt and Sigmund, 2005, 2006; Leimar and Hammerstein, 2001; Ohtsuki and Iwasa, 2004). Simple standing ("standing" in Fig. 2), and stern judging ("judging" in Fig. 2; also known as Kandori) enable full cooperation (Leimar and Hammerstein, 2001; Ohtsuki and Iwasa, 2004). Shunning also enables full cooperation if the players' reputations are initially C and the number of rounds is finite (Ohtsuki and Iwasa, 2007) or if the players' reputations are partially invisible (Nakamura and Masuda, 2011).

In the presence of group structure, four possible locations of the observer are schematically shown in Fig. 1. I call the observer belonging to g_d an "ingroup" observer. Otherwise, the observer is called an "outgroup" observer.

The observers can adopt different social norms for the four cases, as summarized in Fig. 1. When the donor and recipient belong to the same group (Fig. 1A), the ingroup observer uses the norm denoted by s_{ii} to update the donor's personal reputation. In this situation, the outgroup observer does not update the donor's or g_d's reputation (but see Appendix A). When the donor and recipient belong to different groups (Fig. 1B), the ingroup observer uses the norm denoted by s_{io} to update the donor's personal reputation. In this situation, the outgroup observer uses the norm denoted by s_{oo} to update g_d's reputation. These four cases are explained in more detail in Section 2.1.4.

The distinction between s_{ii} and s_{io} allows the ingroup observer to use a double standard for assessing donors. For example, a donor defecting against an ingroup G recipient may be regarded to be B, whereas a defection against an outgroup G recipient may be regarded as G. Such different assessments would not be allowed if s_{ii} and s_{io} are not distinguished.

I call s_{ii}, s_{io}, and s_{oo} subnorms. All the players are assumed to share the subnorms. The typical norms shown in Fig. 2 can be used as subnorms. A subnorm is specified by assigning G or B to each combination of the donor's action (i.e., C or D) and recipient's reputation (i.e., G or B). Therefore, there are $2^4 = 16$ subnorms. An entire social norm of a population consists of a combination of the three subnorms, and there are $16^3 = 4096$ social norms.

Fig. 2. Typical second-order social norms. The rows outside the boxes represent the donor's actions (C or D), and the columns represent the recipient's reputations (G or B). The entries inside the boxes represent the reputations that the observer assigns to the donor in each case.

2.1.3. Action rule

The action rule refers to the mapping from the recipient's reputation (i.e., G or B) to the donor's action (i.e., C or D). The AllC and AllD donors cooperate and defect, respectively, regardless of the recipient's reputation. A discriminator (Disc) donor cooperates or defects when the recipient's reputation is G or B, respectively. An anti-discriminator (AntiDisc) donor cooperates or defects when the recipient's reputation is B or G, respectively.

The donor is allowed to use different action rules toward ingroup and outgroup recipients. For example, a donor who adopts AllC and AllD toward ingroup and outgroup recipients, respectively, implements reputation-independent ingroup favoritism. There are $4 \times 4 = 16$ action rules. A donor refers to the recipient's personal reputation when $g_d = g_r$ (Fig. 1A) and to g_r's group reputation when $g_d \neq g_r$ (Fig. 1B).

2.1.4. Reputation updates

In each round, the ingroup and outgroup observers update the donor's and g_d's reputations, respectively.

If $g_d = g_r$, the donor is assumed to recognize the recipient's personal reputation (Fig. 1A). An ingroup observer in this situation updates the donor's personal reputation on the basis of the donor's action, the recipient's personal reputation, and subnorm s_{ii}. An outgroup observer in this situation is assumed not to update g_d's reputation because such an observer does not know the recipient's personal reputation, although the donor does. Then, the outgroup observer may want to refrain from evaluating the donor because the donor and the observer use different information about the recipient. I also analyzed a variant of the model in which the outgroup observer updates g_d's reputation in this situation. The results are roughly the same as those obtained for the original model (Appendix A).

If $g_d \neq g_r$, the donor is assumed to recognize g_r's reputation, but not the recipient's personal reputation (Fig. 1B). An ingroup observer in this situation updates the donor's personal reputation on the basis of the donor's action, g_r's reputation, and subnorm s_{io}. Both the donor and observer refer to g_r's reputation and not to the recipient's personal reputation. An outgroup observer in this situation updates g_d's reputation based on the donor's action, g_r's reputation, and subnorm s_{oo}.

An outgroup observer knows the recipient's personal reputation if the observer and recipient are in the same group. However, the observer is assumed to ignore this information for two reasons. First, it is evident for the observer that the donor does not have access to the recipient's personal reputation. To explain the second reason, let us consider an outgroup observer who belongs to g_r in a certain round. Assume that this observer assigns a new reputation to g_d according to a subnorm different from one used when the observer does not belong to g_r. The same observer does not belong to g_r when the observer updates the g_d's group reputation next time. This is because the probability that the observer belongs to g_r is infinitesimally small because of the assumption of infinite groups. Therefore, the subnorm used when the observer belongs to g_r is rarely used and immaterial in the present model.

Finally, observers commit reputation assessment error. With probability ϵ, ingroup and outgroup observers independently assign the reputation opposite to the intended one to the donor and g_d, respectively. I introduce this error because G and B players must coexist in the population to distinguish the payoff values for different pairs of action rule and social norm (action–norm pair); such a distinction is necessary for the stability analysis in the following discussion. For simplicity, I neglect other types of error.

2.1.5. Mutant types

To examine the stability of an action rule under a given social norm, I consider two types of mutants.

The first is a single mutant which invades a group. There are $16-1=15$ types of single mutants. A single mutant does not affect the action rule, norm, or reputation of the group that the mutant belongs to because of the assumption of infinite group size.

The second type is a group mutant. A homogeneous group composed of mutants may make the mutant type stronger than the resident type. For example, a group composed of players who cooperate with ingroup recipients and defect against outgroup recipients may invade a fully cooperative population if any intergroup interaction (i.e., C or D) is regarded to be G under s_{oo}. By definition, a group mutant is a homogeneous group of mutants that is different from the resident players in either the action rule or social norm. I consider two varieties of group mutants, as described in Section 3.

2.2. Analysis methods

2.2.1. Reputation scores in the equilibrium

Consider a homogeneous resident population in which all players share an action–norm pair. I will examine the stability of this population against invasion by single and group mutants. For this purpose, I calculate the fraction of players with a G reputation, probability of cooperation, and payoff after infinitely many rounds.

Denote by p^* and p_g^* the equilibrium probabilities that the player's and group's reputations are G, respectively. The self-consistent equation for p^* is given by

$$p^* = r^{\text{in}}[p^*\Phi_G^{\text{in}}(\sigma^{\text{in}})+(1-p^*)\Phi_B^{\text{in}}(\sigma^{\text{in}})]+r^{\text{out}}[p_g^*\Phi_G^{\text{in}}(\sigma^{\text{out}})+(1-p_g^*)\Phi_B^{\text{in}}(\sigma^{\text{out}})], \tag{1}$$

where σ^{in} and σ^{out} are the action rules (i.e., AllC, Disc, AntiDisc, or AllD) that the donor adopts toward ingroup and outgroup recipients, respectively. $\Phi_G^{\text{in}}(\sigma^{\text{in}})$ and $\Phi_B^{\text{in}}(\sigma^{\text{in}})$ are the probabilities that the ingroup observer, based on s_{ii}, assigns reputation G to a donor who has played with a G or B ingroup recipient (i.e., $g_d = g_r$), respectively (Fig. 1A). Similarly $\Phi_G^{\text{in}}(\sigma^{\text{out}})$ and $\Phi_B^{\text{in}}(\sigma^{\text{out}})$ apply when the recipient is in a different group (i.e., $g_d \neq g_r$) and the observer uses s_{io} (Fig. 1B). It should be noted that $\Phi_G^{\text{in}}(\sigma^{\text{in}})$ and $\Phi_G^{\text{in}}(\sigma^{\text{out}})$, for example, may differ from each other even if $\sigma^{\text{in}} = \sigma^{\text{out}}$. Owing to the reputation assignment error, $\Phi_G^{\text{in}}(\sigma^{\text{in}})$, $\Phi_B^{\text{in}}(\sigma^{\text{in}})$, $\Phi_G^{\text{in}}(\sigma^{\text{out}})$, $\Phi_B^{\text{in}}(\sigma^{\text{out}}) \in \{\epsilon, 1-\epsilon\}$ holds true. For example, if the donor is Disc toward ingroup recipients and subnorm s_{ii} is scoring, $\Phi_G^{\text{in}}(\sigma^{\text{in}}) = 1-\epsilon$ and $\Phi_B^{\text{in}}(\sigma^{\text{in}}) = \epsilon$.

The self-consistent equation for p_g^* is given by

$$p_g^* = r^{\text{in}}p_g^*+r^{\text{out}}[p_g^*\Phi_G^{\text{out}}(\sigma^{\text{out}})+(1-p_g^*)\Phi_B^{\text{out}}(\sigma^{\text{out}})], \tag{2}$$

where $\Phi_G^{\text{out}}(\sigma^{\text{out}}) \in \{\epsilon, 1-\epsilon\}$ and $\Phi_B^{\text{out}}(\sigma^{\text{out}}) \in \{\epsilon, 1-\epsilon\}$ are the probabilities that the outgroup observer, based on s_{oo}, assigns reputation G to the donor's group when the donor has played with a G or B outgroup recipient (i.e., $g_d \neq g_r$), respectively (Fig. 1B). The first term on the right-hand side of Eq. (2) corresponds to the fact that g_d's reputation is not updated in the situation illustrated in Fig. 1A. Eqs. (1) and (2) lead to

$$p^* = \frac{r^{\text{in}}\Phi_B^{\text{in}}(\sigma^{\text{in}})+r^{\text{out}}[p_g^*\Phi_G^{\text{in}}(\sigma^{\text{out}})+(1-p_g^*)\Phi_B^{\text{in}}(\sigma^{\text{out}})]}{1-r^{\text{in}}\Phi_G^{\text{in}}(\sigma^{\text{in}})+r^{\text{in}}\Phi_B^{\text{in}}(\sigma^{\text{in}})} \tag{3}$$

and

$$p_g^* = \frac{r^{\text{out}}\Phi_B^{\text{out}}(\sigma^{\text{out}})}{1-r^{\text{in}}-r^{\text{out}}\Phi_G^{\text{out}}(\sigma^{\text{out}})+r^{\text{out}}\Phi_B^{\text{out}}(\sigma^{\text{out}})}. \tag{4}$$

2.2.2. Stability against invasion by single mutants

To examine the stability of the action rule (σ^{in}, σ^{out}) against invasion by single mutants under a given social norm, I consider a single mutant with action rule ($\sigma^{\text{in}'}$, $\sigma^{\text{out}'}$). Because the group is assumed to be infinitely large, a single mutant does not change the reputation of the invaded group. The equilibrium probability p'^* that a mutant receives personal reputation G is given by

$$p'^* = r^{\text{in}}[p^*\Phi_G^{\text{in}}(\sigma^{\text{in}'})+(1-p^*)\Phi_B^{\text{in}}(\sigma^{\text{in}'})]+r^{\text{out}}[p_g^*\Phi_G^{\text{in}}(\sigma^{\text{out}'})+(1-p_g^*)\Phi_B^{\text{in}}(\sigma^{\text{out}'})]. \tag{5}$$

When the probability that the donor and g_d have a G reputation is equal to p and p_g, respectively, the resident donor cooperates with probability

$$r^{\text{in}}\Psi(\sigma^{\text{in}},p)+r^{\text{out}}\Psi(\sigma^{\text{out}},p_g), \tag{6}$$

where

$$\Psi(\tilde{\sigma},\tilde{p}) = \tilde{p}\zeta_G(\tilde{\sigma})+(1-\tilde{p})\zeta_B(\tilde{\sigma}) \quad (\tilde{p}=p,p_g) \tag{7}$$

is the probability that a donor with action rule $\tilde{\sigma} \in \{\text{AllC,Disc, AntiDisc,AllD}\}$ cooperates when the recipient's personal or group reputation is G with probability $\tilde{p}$. $\zeta_G(\tilde{\sigma})$ and $\zeta_B(\tilde{\sigma})$ ($\tilde{\sigma} = \sigma^{\text{in}}$ or σ^{out}) are the probabilities that a $\tilde{\sigma}$ donor cooperates with a G and B recipient, respectively. AllC, Disc, AntiDisc, and AllD correspond to $(\zeta_G(\tilde{\sigma}),\zeta_B(\tilde{\sigma})) = (1,1),(1,0),(0,1)$, and $(0,0)$, respectively.

The payoff to a resident (σ^{in}, σ^{out})-player is given by

$$\pi = -c[r^{\text{in}}\Psi(\sigma^{\text{in}},p^*)+r^{\text{out}}\Psi(\sigma^{\text{out}},p_g^*)]+b[r^{\text{in}}\Psi(\sigma^{\text{in}},p^*)+r^{\text{out}}\Psi(\sigma^{\text{out}},p_g^*)]. \tag{8}$$

The payoff to a ($\sigma^{\text{in}'}$, $\sigma^{\text{out}'}$)–mutant invading the homogeneous population of the resident action–norm pair is given by

$$\pi' = -c[r^{\text{in}}\Psi(\sigma^{\text{in}'},p^*)+r^{\text{out}}\Psi(\sigma^{\text{out}'},p_g^*)]+b[r^{\text{in}}\Psi(\sigma^{\text{in}},p'^*)+r^{\text{out}}\Psi(\sigma^{\text{out}},p_g^*)]. \tag{9}$$

If $\pi > \pi'$ for any mutant, the pair of the action rule (σ^{in}, σ^{out}) and social norm (s_{ii}, s_{io}, s_{oo}) is stable against invasion by single mutants.

2.2.3. Stability against invasion by group mutants

For a mutant group composed of players sharing an action–norm pair, let $p_g'^*$ denote the equilibrium probability that the mutant group has group reputation G. I obtain

$$p'^* = r^{\text{in}}[p'^*\Phi_G^{\text{in}'}(\sigma^{\text{in}'})+(1-p'^*)\Phi_B^{\text{in}'}(\sigma^{\text{in}'})]$$
$$+r^{\text{out}}[p_g^*\Phi_G^{\text{in}'}(\sigma^{\text{out}'})+(1-p_g^*)\Phi_B^{\text{in}'}(\sigma^{\text{out}'})] \tag{10}$$

and

$$p_g'^* = r^{\text{in}}p_g'^*+r^{\text{out}}[p_g^*\Phi_G^{\text{out}}(\sigma^{\text{out}'})+(1-p_g^*)\Phi_B^{\text{out}}(\sigma^{\text{out}'})], \tag{11}$$

where $\Phi_G^{\text{in}'}(\sigma^{\text{in}'})$ or $\Phi_B^{\text{in}'}(\sigma^{\text{in}'})$ is the probability that an ingroup observer assigns reputation G to a mutant donor who has played with a G or B ingroup recipient, respectively. Even if $\sigma^{\text{in}'}$ and σ^{in} are the same, $\Phi_G^{\text{in}'}(\sigma^{\text{in}'})$ will be generally different from $\Phi_G^{\text{in}}(\sigma^{\text{in}})$ because the ingroup observer in the mutant group may use a subnorm s_{ii} that is different from one used in the resident population. Parallel definitions apply to $\Phi_G^{\text{in}'}(\sigma^{\text{out}'})$ and $\Phi_B^{\text{in}'}(\sigma^{\text{out}'})$. Eqs. (10) and (11) yield

$$p'^* = \frac{r^{\text{in}}\Phi_B^{\text{in}'}(\sigma^{\text{in}'})+r^{\text{out}}[p_g^*\Phi_G^{\text{in}'}(\sigma^{\text{out}'})+(1-p_g^*)\Phi_B^{\text{in}'}(\sigma^{\text{out}'})]}{1-r^{\text{in}}\Phi_G^{\text{in}'}(\sigma^{\text{in}'})+r^{\text{in}}\Phi_B^{\text{in}'}(\sigma^{\text{in}'})} \tag{12}$$

and

$$p_g'^* = p_g^*\Phi_G^{\text{out}}(\sigma^{\text{out}'})+(1-p_g^*)\Phi_B^{\text{out}}(\sigma^{\text{out}'}), \tag{13}$$

respectively.

The payoff to a mutant player in the mutant group is given by

$$\pi_{g'} = -c[r^{\text{in}}\Psi(\sigma^{\text{in}'},p'^*)+r^{\text{out}}\Psi(\sigma^{\text{out}'},p_g^*)]+b[r^{\text{in}}\Psi(\sigma^{\text{in}'},p'^*)+r^{\text{out}}\Psi(\sigma^{\text{out}},p_g'^*)]. \tag{14}$$

If $\pi > \pi_g{}'$ holds true for any group mutant player, the resident population is stable against invasion by group mutants.

3. Results

3.1. Action–norm pairs stable against invasion by single mutants

There are 16 action rules and $16^3 = 4096$ social norms, which leads to $16 \times 4096 = 65\,536$ action–norm pairs. Because of the symmetry with respect to the swapping of G and B, I neglect action–norm pairs in which the action rule (i.e., AllC, Disc, AntiDisc, or AllD) toward ingroup recipients is $\sigma^{in} = $ AntiDisc without loss of generality. Such an action–norm pair can be converted to $\sigma^{in} = $ Disc by swapping G and B in the action rule and social norm. The model is also invariant if G and B group reputations are completely swapped in the action rule toward outgroup recipients σ^{out} and subnorms s_{io} and s_{oo}. Therefore, I can also neglect the action–norm pairs with $\sigma^{out} = $ AntiDisc without loss of generality. This symmetry consideration leaves $65\,536/4 = 16\,384$ action–norm pairs (Fig. 3).

I exhaustively examined the stability of all $16 \times 4096 = 65\,536$ action–norm pairs. A similar exhaustive search was first conducted in (Ohtsuki and Iwasa, 2004) for an indirect reciprocity game without group structure in the population. In the following, π (Eq. (8)) mentions the player's payoff in the resident population in the limit of no reputation assignment error, i.e., $\epsilon \to 0$.

I first describe action rules that are stable against invasion by single mutants under a given social norm. I identified them using Eqs. (1)–(9). Under any given social norm, action rule $(\sigma^{in}, \sigma^{out}) = $ (AllD, AllD) is stable and yields $\pi = 0$. Other action–norm pairs also yield $\pi = 0$, but there are 588 stable action–norm pairs with $\pi > 0$ (Fig. 3). For a given social norm, at most one action rule that yields a positive payoff is stable. For all 588 solutions, the condition for stability against invasion by single mutants (i.e., $\pi > \pi'$, where π and π' are given by Eqs. (8) and (9), respectively) is given by

$$br^{in} > c. \tag{15}$$

Eq. (15) implies that cooperation is likely when the benefit-to-cost ratio is large, which is a standard result for different mechanisms of cooperation in social dilemma games (Nowak, 2006b). Cooperation is also likely when intragroup interaction is relatively more frequent than intergroup interaction (i.e., large r^{in}).

3.2. Stability against invasion by group mutants

The stability of these 588 action–norm pairs against invasion by group mutants was also examined based on Eqs. (10)–(14). Properly setting the variety of group mutants is not a trivial issue. At most, $65\,536 - 1 = 65\,535$ types of group mutants that differ from the resident population in either action rule or social norm are possible. However, an arbitrarily selected homogeneous mutant group may be fragile to invasion by different single mutants into the mutant group. Although I do not model evolutionary dynamics, evolution would not allow the emergence and maintenance of such weak mutant groups. With this in mind, I consider two group mutation scenarios.

3.2.1. Scenario 1

Single mutants may invade the resident population when Eq. (15) is violated. In this scenario 1, the mutants are assumed to differ from the resident population in the action rule, but not the social norm, for simplicity. There are $16 - 1 = 15$ such mutants, and some of them, including $(\sigma^{in}, \sigma^{out}) = $ (AllD, AllD), can invade the resident population when $1 < b/c < 1/r^{in}$. Such mutant action rules may spread to occupy a single group when Eq. (15) is violated. I consider the stability of the resident population against the homogeneous groups of mutants that invade the resident population as single mutants when $1 < b/c < 1/r^{in}$.

Among the 588 action–norm pairs that yield $\pi > 0$, 440 pairs are stable against group mutation. Among these 440 pairs, I focus on those yielding perfect intragroup cooperation, i.e., those yielding $\lim_{\epsilon \to 0} \Psi(\sigma^{in}, p^*) = 1$, where Ψ and p^* are given in Section 2.2. For the other stable pairs, see Appendix B. This criterion is satisfied by 270 pairs (Fig. 3). For all 270 pairs, every player obtains personal reputation G (i.e., $\lim_{\epsilon \to 0} p^* = 1$), and the donor cooperates with ingroup recipients because the recipients have reputation G (i.e., $\sigma^{in} = $ Disc).

In all 270 pairs, s_{ii} is either standing (GBGG in shorthand notation), judging (GBBG), or shunning (GBBB) (refer to Fig. 2 for definitions of these norms). In the shorthand notation, the first, second, third, and fourth letters (either G or B) indicate the donor's or g_d's new reputation when the donor cooperates with a G recipient, the donor defects against a G recipient, the donor cooperates with a B recipient, and the donor defects against a B recipient, respectively. Standing, judging, and shunning in s_{ii} are exchangeable for any fixed combination of $\sigma^{in} = $ Disc, σ^{out}, s_{io}, and s_{oo}. Therefore, there are $270/3 = 90$ combinations of σ^{out}, s_{io}, and s_{oo}, which are summarized in Table 1. An asterisk indicates an entry that can be either G or B. For example, GB*G indicates standing (GBGG) or judging (GBBG). The probability of cooperation toward outgroup recipients, payoff (π; Eq. (8)), and the probability that a group has a G reputation (p_g^*; Eq. (2)) are also shown in Table 1. The stable action–norm pairs can be classified into three categories.

- *Full cooperation*: Donors behave as Disc toward outgroup recipients, i.e., $\sigma^{out} = $ Disc and cooperate with both ingroup and outgroup recipients with probability 1. Accordingly, $\pi = b - c$ and $p_g^* = 1$.
  In this case, indirect reciprocity among different groups as well as that within single groups is realized. Action rule $\sigma^{in} = \sigma^{out} = $ Disc is stable if s_{io} is either standing (GBGG), judging (GBBG), or shunning (GBBB) and s_{oo} is either standing or judging. The condition for stability against group mutation is the mildest one (i.e., $b > c$) for each action–norm pair.
  Under full cooperation, s_{io} and s_{io} must be the one that stabilizes cooperation in the standard indirect reciprocity game without a group-structured population (Ohtsuki and
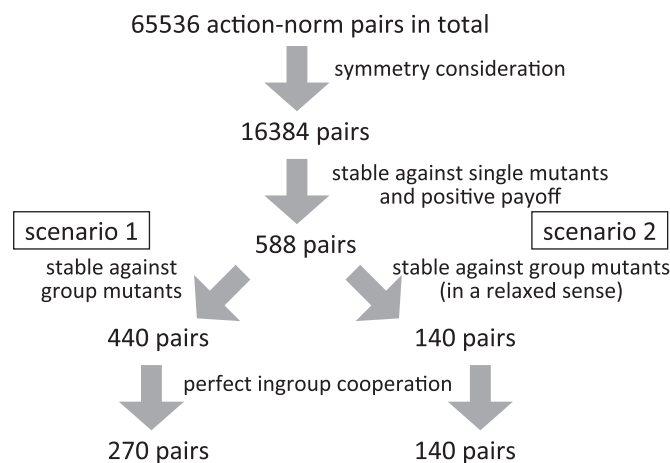

Fig. 3. Procedure for obtaining the stable action–norm pairs with perfect ingroup cooperation shown in Tables 1 and 3.

Table 1
Stable action–norm pairs with perfect ingroup cooperation under scenario 1. The probability of cooperation with outgroup recipients, π, and p_g^* are the values in the limit $\epsilon \to 0$. s_{ii} = GBGG (standing), GBBG (judging), or GBBB (shunning). Action–norm pairs only different in s_{ii} were distinguished when counting the number of stable action–norm pairs. An asterisk indicates that both G and B apply.

State	Prob. C to outgroup	π	σ^{out}	p_g^*	Social norm $(s_{io} - - s_{oo})$	No pairs
Full cooperation	1	$b-c$	Disc	1	GB*G–GB*G GBBB–GB*G	18
Partial ingroup favoritism	$\frac{1}{2}$	$\frac{(b-c)(1+r^{in})}{2}$	Disc	$\frac{1}{2}$	GB*G–GB*B	12
Perfect ingroup favoritism	0	$(b-c)r^{in}$	AllD	1	*GBB–*G*G *GBG–*G*G *GGG–*G*G	72
				$\frac{1}{2}$	*G*G–*G*B *G*G–*B*G	96
				0	BB*G–*B*B BG*G–*B*B GG*G–*B*B	72

Table 2
Conditions for stability of partial ingroup favoritism against group mutation under scenario 1. The condition on r^{in} is required for the three out of 12 social norms to prevent the invasion by group mutants that defect against ingroup recipients and cooperate with outgroup recipients.

Conditions	Social norm (s_{ii})	Social norm ($s_{io} - s_{oo}$)	No. pairs
$b > c$	GBGG, GBBG, or GBBB	GBBG–GBBB GBGG–GBBB GBBG–GBGB	9
$b > c$ and $r^{in} > \sqrt{2}-1$	GBGG	GBGG–GBGB	1
$b > c$ and $r^{in} > 1/2$	GBBG or GBBB	GBGG–GBGB	2

Iwasa, 2004; Nowak and Sigmund, 2005; Ohtsuki and Iwasa, 2007). The ingroup observer monitors donors' actions toward outgroup recipients through the use of s_{io} = standing, judging, or shunning, even though ingroup players are not directly harmed if donors defect against outgroup recipients. The ingroup observer does so because donors' defection against outgroup recipients would negatively affect the group's reputation.

- Partial ingroup favoritism: Donors adopt σ^{out} = Disc and cooperate with ingroup recipients with probability 1 and outgroup recipients with probability 1/2. Accordingly, $\pi = (b-c)(1+r^{in})/2$ and $p_g^* = 1/2$.

  In this case, action rule $\sigma^{in} = \sigma^{out}$ = Disc is stable if s_{io} is either standing (GBGG) or judging (GBBG), and s_{oo} is either scoring (GBGB) or shunning (GBBB). The condition for stability against group mutation is shown in Table 2.

- *Perfect ingroup favoritism*: Donors adopt σ^{out} = AllD and always cooperate with ingroup recipients and never with outgroup recipients regardless of the recipient's group reputation. Accordingly, $\pi = (b-c)r^{in}$.

  Table 1 suggests that action rule $(\sigma^{in}, \sigma^{out})$ = (Disc,AllD) can be stable for any subnorm s_{oo}. This is true because the group reputation, whose update rule is given by s_{oo}, is irrelevant in the current situation; the donor anyway defects against outgroup recipients. Nevertheless, s_{oo} determines s_{io} that is consistent with ingroup cooperation through the probability of a G group reputation p_g^*.

  When s_{oo} = *G*G, the outgroup observer evaluates defection against outgroup recipients to be G (Fig. 1B). Therefore, $p_g^* = 1$.

In this case, s_{io} = *GBB, *GBG, and *GGG stabilize perfect ingroup favoritism. Under any of these s_{io}, the ingroup observer assigns G to a donor that defects against a recipient in a G outgroup because the second entry of s_{io} is equal to G in each case. Therefore, $p^* = 1$, and full ingroup cooperation is stable.

When s_{oo} = *G*B or *B*G, the outgroup observer evaluates defection against outgroup recipients to be G with probability 1/2. Therefore, $p_g^* = 1/2$. In this case, s_{io} = *G*G stabilizes perfect ingroup favoritism. Under such an s_{io}, the ingroup observer assigns G to a donor that defects against a recipient in a G outgroup because the second and fourth entries of s_{io} are equal to G.

When s_{oo} = *B*B, the outgroup observer evaluates defection against outgroup recipients to be B. Therefore, $p_g^* = 0$. In this case, s_{io} = BB*G, BG*G, and GG*G stabilize perfect ingroup favoritism. Under such an s_{io}, the ingroup observer assigns G to a donor that defects against a recipient in a G outgroup because the fourth entry of s_{io} is equal to G.

In all the cases, the stability against invasion by group mutants requires $b > c$.

3.2.2. Scenario 2

In scenario 2 of group mutation, it is hypothesized that a group of mutants immigrates from a different population that is stable against invasion by single mutants. Such a group mutant may appear owing to the encounter of different stable cultures (i.e., action–norm pairs). The pairs that are stable against invasion by single mutants and yield zero payoff, such as the population of AllD players, must be also included in the group mutant list. It should be noted that a mutant group may have a different social norm from that for the resident population.

Among the 588 action–norm pairs that are stable against single mutation, no pair is stable against group mutation. However, 140 pairs are stable against group mutation for any $b > c$ in a relaxed sense that the resident player's payoff is not smaller than the group mutant's payoff, i.e., $\pi \geq \pi_g'$ (Fig. 3). The homogeneous population of each pair is neutrally invaded by some group mutants, i.e., $\pi = \pi_g'$. Therefore, I examine the evolutionary stability (e.g., Nowak, 2006a) against group mutation. In other words, for the group mutants yielding $\pi = \pi_g'$, I require $\pi > \pi_g'$ when the resident players are replaced by group mutants.

All 140 action–norm pairs are evolutionarily stable except that each pair is still neutrally invaded by their cousins. For example, four action–norm pairs specified by $\sigma^{in} = \sigma^{out}$ = Disc, s_{ii} = GB*G, s_{io} = GB*G, s_{oo} = GBGG neutrally invade each other. These pairs yield the same payoff $\pi = b-c$ and are evolutionarily stable against invasion by the other group mutants. Therefore, I conclude that the four pairs collectively form a set of stable solutions. Other sets of stable solutions consist of four or eight neutrally invadable action–norm pairs that yield the same payoff and differ only in s_{ii} and s_{io}.

All 140 pairs realize perfect intragroup cooperation such that the players have G personal reputations and σ^{in} = Disc (Fig. 3). Subnorm s_{ii} = GBGG (i.e., standing) or GBBG (i.e., judging) is exchangeable for any fixed combination of σ^{in} = Disc, σ^{out}, s_{io}, and s_{oo}. Therefore, there are $140/2 = 70$ possible combinations of σ^{out}, s_{io}, and s_{oo}, which are listed in Table 3. The 140 pairs are a subset of the 270 pairs stable under scenario 1. The stable sets of action–norm pairs can be classified into three categories. (1) Full cooperation occurs if all the subnorms are standing or judging. As already mentioned as an example, under s_{oo} = GBGG, the four action–norm pairs $(\sigma^{in}, \sigma^{out}, s_{ii}, s_{io})$ = (Disc,Disc,GBGG,GBGG), (Disc, Disc, GBGG, GBBG), (Disc, Disc, GBBG, GBGG), and (Disc, Disc, GBBG, GBBG) can neutrally invade each other. Similarly, if

Table 3
Stable action–norm pairs with perfect ingroup cooperation under scenario 2. s_{ii} = GBGG (standing) or GBBG (judging). Different action–norm pairs in the same row are neutrally invadable to each other. An asterisk indicates either G or B.

State	Prob. C to outgroup	π	σ^{out}	p_g^*	Social norm (s_{io}——s_{oo})	No pairs
Full cooperation	1	$b-c$	Disc	1	GB*G–GBGG GB*G–GBBG	8
Partial ingroup favoritism	$\frac{1}{2}$	$\frac{(b-c)(1+r^{in})}{2}$	Disc	$\frac{1}{2}$	GB*G–GBBB	4
				1	*G*G–BGBG *G*G–GGBG *G*G–BGGG *G*G–GGGG	32
Perfect ingroup favoritism	0	$(b-c)r^{in}$	AllD	$\frac{1}{2}$	*G*G–BGBB *G*G–GGBB *G*G–BGGB *G*G–GGGB *G*G–BBBG *G*G–GBBG *G*G–BBGG *G*G–GBGG	64
				0	*G*G–BBBB *G*G–GBBB *G*G–BBGB *G*G–GBGB	32

s_{oo} = GBBG, the same four action–norm pairs constitute a set realizing stable full cooperation. These two sets of four pairs are evolutionarily stable against invasion by each other. In total, there are eight pairs that realize full cooperation. (2) Partial ingroup favoritism occurs for a set of four action–norm pairs. (3) Perfect ingroup favoritism occurs under the same subnorms s_{oo} as those for scenario 1. For a fixed s_{oo}, the same eight action–norm pairs $(\sigma^{in}, \sigma^{out}, s_{ii}, s_{io}) = (\text{Disc}, \text{AllD}, \text{GB}*\text{G}, *\text{G}*\text{G})$ yield the same payoff $\pi = (b-c)r^{in}$, can neutrally invade each other, and are evolutionarily stable against the other group mutants.

3.3. When observers use simpler social norms

In fact, players may not differentiate between the three subnorms. Players may use a common norm for assessing ingroup donors irrespective of the location of recipients. Table 1 indicates that, if $s_{ii} = s_{io}$ is imposed for the resident population, but not for mutants, perfect ingroup favoritism is excluded. Under scenario 1, full cooperation is stable when $s_{ii} = s_{io}$ = standing, judging, or shunning and s_{oo} = standing or judging. Partial ingroup favoritism is stable when $s_{ii} = s_{io}$ = standing or judging and s_{oo} = scoring or shunning. Under scenario 2, full cooperation is stable when $s_{ii} = s_{io}$ = standing or judging and s_{oo} = standing or judging. Partial ingroup favoritism is stable when $s_{ii} = s_{io}$ = standing or judging and s_{oo} = shunning.

Alternatively, players may use a common norm for assessing donors playing with outgroup recipients irrespective of the location of donors. If $s_{ii} \neq s_{io}$ is allowed and $s_{io} = s_{oo}$ is imposed, partial ingroup favoritism is excluded. Under scenario 1, full cooperation is stable when s_{ii} = standing, judging, or shunning and $s_{io} = s_{oo}$ = standing or judging. Perfect ingroup favoritism is stable when s_{ii} = standing, judging, or shunning and $s_{io} = s_{oo}$ = *G*G. The results under scenario 2 differ from those under scenario 1 only in that s_{ii} = shunning is disallowed.

Finally, if all the three subnorms are forced to be equal, only full cooperation is stable, and the norm is standing or judging. This holds true for both scenarios 1 and 2.

4. Discussion

4.1. Summary of the results

I identified the pairs of action rule and social norm that are stable against invasion by single and group mutants in the game of group-structured indirect reciprocity. Full cooperation (i.e., cooperation within and across groups) based on personal and group reputations, partial ingroup favoritism, and perfect ingroup favoritism are stable under different social norms. Perfect ingroup favoritism is attained only when the donor defects against outgroup recipients regardless of their reputation (i.e., σ^{out} = AllD). Perfect ingroup favoritism does not occur with the combination of a donor that is ready to cooperate with G outgroup recipients (i.e., σ^{out} = Disc) and a B group reputation. The mechanism for ingroup favoritism revealed in this study is distinct from those proposed previously (see Section 1).

The major condition for either full cooperation, partial ingroup favoritism, and perfect ingroup favoritism, depending on the assumed social norm, is given by $br^{in} > c$. In only 3 out of 270 social norms in scenario 1, an additional condition for r^{in} is imposed (Section 3.2.1). In general, different mechanisms of cooperation can be understood in an unified manner such that cooperation occurs if and only if b/c is larger than a threshold value (Nowak, 2006b). For example, b/c must be larger than the inverse of the relatedness parameter r and the inverse of the discount factor in kin selection and direct reciprocity, respectively. The present result also fits this view; r^{in} corresponds to r in the case of kin selection.

I assumed that players approximate personal reputations of individuals in other groups by group reputations (i.e., outgroup homogeneity). Adoption of outgroup homogeneity may be evolutionarily beneficial for players owing to the reduction in the cognitive burden of recognizing others' personal reputations. Instead, the players pay potential costs of not being able to know the personal reputations of individuals in other groups. To explore evolutionary origins of group reputation, one has to examine competition between players using the group reputation and players not using it. It would also be necessary to introduce a parameter representing the cost of obtaining personal reputations of outgroup individuals. Such an analysis is warranted for future work.

All the players are assumed to use the same social norm. This assumption may be justified for well-mixed populations but less so for populations with group structure because group structure implies relatively little intergroup communication. It seems to be more natural to assume that subnorms s_{ii} and s_{io}, which are used to evaluate actions of ingroup donors, depend on groups. Under scenario 2 (Section 3.2.2), any stable action–norm pair is neutrally invaded by its cousins who are different in s_{ii} and s_{io}. This result implies that different groups can use different norms. For example, for all the solutions shown in Table 3, some groups can use $s_{ii} = $ GBGG (i.e., standing), while other groups in the same population can use $s_{io} = $ GBBG (i.e., judging). To better understand the possibility of heterogeneous social norms, analyzing a population composed of a small number of groups, probably by different methods, would be helpful.

4.2. Cooperation based on group reputation is distinct from group selection

Indirect reciprocity based on group reputation is distinct from any type of group selection. This is true for both full cooperation and ingroup favoritism. There are two dominant variants of group selection that serve as mechanisms for cooperation in social dilemma games (West et al., 2007, 2008).

The first type is group competition, in which selection pressure acts on groups such that a group with a large mean payoff would replace one with a small mean payoff. Models with group competition induce ingroup favoritism (Choi and Bowles, 2007; García and van den Bergh, 2011), altruistic punishment (Boyd et al., 2003), and evolution of the judging social norm in the standard game of indirect reciprocity whereby players interact within each group (Pacheco et al., 2006; Chalub et al., 2006). In contrast, the present study is not concerned with evolutionary dynamics including group competition. The group mutant is assumed to statically compare the payoff to the resident group with that to the mutant group.

The second type of group selection requires assortative reproduction in the sense that the offspring have a higher probability of belonging to specific groups than to other groups depending on the offspring's genotype. It is mathematically identical with kin selection (West et al., 2007, 2008). This variant of group selection is also irrelevant to the present model, which is not concerned with the reproduction process.

The analysis in this study is purely static. I avoided examining evolutionary dynamics for two reasons. First, the discovered mechanism for cooperation may be confused with group selection in the presence of evolutionary dynamics. Second, the model becomes needlessly complicated. Introducing evolutionary dynamics implies that one specifies a rule for reproduction. Offspring may be assumed to belong to the parent's group or to migrate to another group. It may then be necessary to consider the treatment of, for example, the heterogeneous group size. Because evolutionary dynamics are neglected, the present model explains neither emergence of full cooperation and ingroup favoritism nor the likelihood of different solutions, which is a main limitation of the present study.

I stress that the concept of group mutants is introduced to sift the set of stable action–norm pairs. Unless group competition is assumed, the concept of group mutants does not particularly promote cooperation in evolutionary dynamics.

4.3. Group competition can enable full cooperation and ingroup favoritism even if $br^{in} > c$ is violated

Under a proper social norm, full cooperation or ingroup favoritism is stable if $br^{in} > c$ (i.e., Eq. (15) is satisfied) in most cases. With

probability r^{in}, the donor, recipient, and observer are engaged in the standard (i.e., no group structure) indirect reciprocity game limited to a single group (Fig. 1A). In the standard indirect reciprocity game under incomplete information, $bq > c$ is quite often the condition for cooperation, where q is the probability that the recipient's reputation is observed. This holds true when q indicates the observation probability for the donor (Nowak and Sigmund, 1998a,b; Brandt and Sigmund, 2005, 2006; Suzuki and Toquenaga, 2005) or that for both the donor and observer (Nakamura and Masuda, 2011). Because r^{in} is also equal to the probability that the donor sees the recipient's personal reputation, r^{in} resembles q. In fact, replacing r^{in} by q in Eq. (15) yields $bq > c$.

If a player is capable of recognizing the personal reputation of a fixed number of others, the maximum population size for which indirect reciprocity is possible in the standard indirect reciprocity game scales as $1/q$. The consistency between Eq. (15) and $bq > c$ implies that the concept of group reputation does not increase the maximum population size for which indirect reciprocity occurs. However, under group competition (Section 4.2), full cooperation and ingroup favoritism can be stable even if the restriction imposed by Eq. (15) is removed.

To explain this point, assume that the population is subjected to evolutionary dynamics such that players with relatively large payoffs would bear more offspring in the same group and group competition occurs. The rate of group competition is denoted by $1/t_{gc}$, where t_{gc} is the mean time interval between successive group competition events. Emergence of a single mutant occurs with rate $1/t_m$. Selection and reproduction of single players occur with rate $1/t_s$.

If Eq. (15) is violated, single mutants emerge in time $\propto t_m$. Then, some types of mutants, including the AllD mutant, spread in the invaded group in time $\propto t_s$ under scenario 1 of group mutation. The invaded group presumably possesses a smaller group-averaged payoff than other resident groups because the resident population is stable against invasion by group mutants as long as $b > c$, in all but three of 270 action–norm pairs (Table 2). If $1/t_{gc} \gg 1/t_m$, such an invaded group is likely to be eradicated by group competition because group competition occurs much faster than the emergence of single mutants. In this case, full cooperation or ingroup favoritism, depending on the given social norm, can be maintained in the absence of Eq. (15). This discussion does not involve timescale t_s.

Group competition is needed to remove Eq. (15). If Eq. (15) is imposed, cooperation occurs without group competition.

4.4. Relationship to previous behavioral experiments

In this section, I discuss possible linkages between the present model and the previous experiments examining indirect reciprocity and third-party punishments.

Yamagishi and colleagues conducted a series of laboratory experiments to show that ingroup favoritism is induced by a group heuristic (Yamagishi et al., 1998, 1999; Yamagishi and Mifune, 2008). With a group heuristic, donors cooperate with ingroup recipients because the donors expect repayment from other ingroup players. Donors do not use the information about others' reputations in these experiments. In contrast, players use personal reputations of ingroup members in the present model. Nevertheless, the previous experiments and the current model do not contradict each other.

In another laboratory experiment, Mifune et al. showed that presentation of eye-like painting promotes donor's cooperation toward ingroup recipients in the dictator game (Mifune et al., 2010). For expository purposes, I define serious subnorm to be either standing, judging, or shunning. If the eye-like painting approximates an ingroup observer obeying a serious subnorm,

this experimental result is consistent with the present theory because ingroup cooperation is theoretically stable when the ingroup observer adopts a serious subnorm. Because the painting does not increase the cooperation toward outgroup recipients (Mifune et al., 2010), it may not turn s_{io} to a serious subnorm for some psychological reason. Humans may use double standards, i.e., $s_{ii} \neq s_{io}$, which favor ingroup favoritism in my model.

Other behavioral experiments have addressed the relationship between third-party altruistic punishments and ingroup favoritism (Bernhard et al., 2006; Shinada et al., 2004). In precise terms, third-party punishments and reputation-based indirect reciprocity are distinct mechanisms for cooperation (Sigmund et al., 2001; Ohtsuki et al., 2009). Nevertheless, below I discuss possible linkages between these experiments and my model.

In indigenous communities in Papua New Guinea (Bernhard et al., 2006), the amount of punishment is larger if the punisher belongs to the donor's group than to a different group (compare ABC and AB cases in their Fig. 1). Their results suggest that the ingroup observer may use a serious subnorm and the outgroup observer may not. Furthermore, given that the punisher is in the donor's group, the amount of punishment is larger if the donor and recipient belong to the same group (Fig. 1A, if the punisher is identified with the ingroup observer) than if they belong to different groups (Fig. 1B; compare the ABC and AC cases in Fig. 1 of Bernhard et al., 2006). In this situation, the ingroup observer may use a serious subnorm s_{ii} when the donor plays with ingroup recipients (Fig. 1A) and use a nonserious subnorm s_{io} when the donor plays with outgroup recipients (Fig. 1B). My model reproduces ingroup favoritism under these conditions.

However, my model and others are not concerned with a main finding in Bernhard et al. (2006) that the amount of punishment is larger when the punisher and recipient belong to the same group. For the reasons stated in Section 2.1.4, I did not assume that observers make their judgments differently when they belong to the recipient's group g_r and to a different group. To theoretically explain the main finding in Bernhard et al. (2006), one should explicitly analyze the case of a finite number of groups.

In different laboratory experiments, the amount of punishment is larger for an ingroup donor's defection than an outgroup donor's defection (Shinada et al., 2004). My results are consistent with their results in that, for ingroup favoritism, the donor's action must be seriously evaluated by the ingroup observer using s_{ii} and not seriously by the outgroup observer using s_{oo}.

4.5. Reduction of ingroup favoritism

Although ingroup favoritism seems to be a canonical behavior of humans, reduction of ingroup bias would induce intergroup cooperation and is socially preferable (Yamagishi et al., 1998). Full cooperation is Pareto efficient, whereas ingroup favoritism is not. Various psychological and sociological mechanisms for reducing the ingroup bias, such as guilt, "auto-motive" control, retraining, empathy, and decategorization have been proposed (Hewstone et al., 2002; Dovidio et al., 2005; Sedikides et al., 1998).

My results provide theory-based possibilities of reducing ingroup bias. First, if the social norm is fixed, conversion from ingroup favoritism to full cooperation is theoretically impossible because full cooperation and ingroup favoritism do not coexist under a given social norm. Therefore, advising players to change their behavior toward outgroup recipients from AllD to Disc is not recommended unless the social norm is also altered. Conversion from ingroup favoritism to full cooperation requires a change in the social norm such that players as observers seriously assess ingroup donors' actions toward outgroup recipients (with s_{io}) and outgroup–outgroup interaction (with s_{oo}). In particular, if s_{io} is a

serious subnorm, perfect ingroup favoritism with no intergroup cooperation disappears (Section 3.3).

Second, if the three subnorms are the same, the perfect and partial ingroup favoritism is eradicated. The coincidence of only two subnorms is insufficient to induce full cooperation (Section 3.3). The subnorms $s_{ii} = s_{io} = s_{oo}$ that exclude the ingroup bias and realize full cooperation are standing or judging. Therefore, without speaking of serious subnorms, forcing players to use the same subnorms consistently in assessing donors in different situations may be also effective in inducing full cooperation.

Ingroup favoritism has been mostly an experimental question except for some recent theoretical studies. This study is a first step toward understanding and even manipulating the dichotomy between full cooperation and ingroup favoritism in the context of indirect reciprocity.

Acknowledgments

Appendix A. A variant of the model with different reputation dynamics

In this section, I analyze a variant of the model in which outgroup observers update the group reputation of donors involved in ingroup interaction (i.e., $g_d = g_r$).

A.1. Reputation dynamics

I assume that the outgroup observer uses the donor's action, the recipient's personal reputation, and s_{oo}, to update g_d's (not the donor's personal) reputation.

The equivalent of Eq. (2) under this reputation update rule is given by

$$p_g^* = r^{in}[p^* \Phi_G^{out}(\sigma^{in}) + (1-p^*)\Phi_B^{out}(\sigma^{in})] + r^{out}[p_g^* \Phi_G^{out}(\sigma^{out}) + (1-p_g^*)\Phi_B^{out}(\sigma^{out})]. \tag{16}$$

I obtain p^* and p_g^* by solving the set of linear equations (1) and (16). Eqs. (5)–(10), and (12) are unchanged. As compared to the case of the original reputation update rule (original case for short), Eq. (11) is replaced by

$$p_g'^* = r^{in}[p'^* \Phi_G^{out}(\sigma^{in'}) + (1-p'^*)\Phi_B^{out}(\sigma^{in'})] + r^{out}[p_g^* \Phi_G^{out}(\sigma^{out'}) + (1-p_g^*)\Phi_B^{out}(\sigma^{out'})]. \tag{17}$$

The equivalent of Eq. (13) is obtained by substituting Eq. (12) in Eq. (17).

Because of the symmetry with respect to G and B, I exclude action rules having $\sigma^{in} =$ AntiDisc from the exhaustive search, as I did in the original case (Section 3.1). It should be noted that one cannot eliminate action–norm pairs with $\sigma^{out} =$ AntiDisc on the basis of symmetry consideration, which is different from the original case. This is because a player's personal and group reputations are interrelated through the behavior of the outgroup observer when $g_d = g_r$.

A.2. Results

Under the modified reputation update rule, there are 725 action–norm pairs that are stable against invasion by single mutants and yield $\pi > 0$.

Table 4

Stable action–norm pairs with a positive probability of cooperation that are not included in Table 1. s_{ii} = GBGG (standing), GBBG (judging), or GBBB (shunning). An asterisk indicates either G or B. The sixth and seventh rows in the table are not aggregated because the stability condition is different between these cases (Appendix B).

Prob. C to ingroup	Prob. C to outgroup	π	σ^{out}	p^*	p_g^*	Social norm ($s_{ii}-s_{io}-s_{oo}$)	No. pairs
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{b-c}{2}$	Disc	$\frac{1}{2}$	$\frac{1}{2}$	GBBB-GBBB-GB*B	2
$\frac{1}{2}$	0	$\frac{(b-c)r^{in}}{2}$	AllD	$\frac{1}{2}$	$\frac{1}{2}$	GBBB-*GBB-*G*B GBBB-BB*G-*G*B GBBB-*GBB-*B*G GBBB-BB*G-*B*G	32
$\frac{1+r^{in}}{2}$	$\frac{1}{2}$	$\frac{(b-c)(1+(r^{in})^2)}{2}$	Disc	$\frac{1+r^{in}}{2}$	$\frac{1}{2}$	GB*G-GBBB-GBGB GB*G-GBBB-GBBB	4
$\frac{1+r^{in}}{2}$	0	$\frac{(b-c)r^{in}(1+r^{in})}{2}$	AllD	$\frac{1+r^{in}}{2}$	$\frac{1}{2}$	GB*G-*GBB-*G*B GB*G-BB*G-*G*B GB*G-*GBB-*B*G GB*G-BB*G-*B*G	64
r^{in}	0	$(b-c)(r^{in})^2$	AllD	r^{in}	1	GB*G-BBBB-*G*G GB*G-BB*G-*G*G	68
					$\frac{1}{2}$	GB*G-BBBB-*G*B GB*G-BBBB-*B*G	
			Disc AllD		0	GB*G-GBBB-BB*B GB*G-BBBB-*B*B GB*G-*GBB-*B*B	

Under scenario 1, 507 out of the 725 pairs are stable against group mutation, and 324 out of the 507 pairs yield perfect ingroup cooperation. The 324 action–norm pairs are classified as follows. First, 68 pairs yield full cooperation with either $(\sigma^{in}, \sigma^{out})$ = (Disc,Disc) or (Disc, AntiDisc). Second, 14 pairs yield partial ingroup favoritism with $(\sigma^{in}, \sigma^{out})$ = (Disc,AntiDisc). Third, 236 pairs yield perfect ingroup favoritism with $(\sigma^{in}, \sigma^{out})$ = (Disc,AllD). Fourth, six pairs yield perfect ingroup favoritism with $(\sigma^{in}, \sigma^{out})$ = (Disc,AntiDisc).

As in the original case, σ^{in} = Disc, and s_{ii} is either standing, judging, or shunning for these pairs. In contrast to the original case, $(\sigma^{in}, \sigma^{out})$ = (Disc,AntiDisc) can be stable, yield perfect ingroup cooperation, and even yield outgroup cooperation, under some social norms. In such a situation, the values of the personal and group reputations (i.e., G and B) have opposite meanings. In other words, a G but not B personal reputation elicits intragroup cooperation, while a B but not G group reputation elicits intergroup cooperation. Therefore, action rule $(\sigma^{in}, \sigma^{out})$ = (Disc,AntiDisc) in this situation can be regarded as a relative of $(\sigma^{in}, \sigma^{out})$ = (Disc,Disc) in the situation in which the values of the personal and group reputations have the same meaning. On this basis, I consider that the present results are similar to those obtained for the original case (Table 1). In particular, only full cooperation is stable under standing or judging if s_{ii}, s_{io}, and s_{oo} are assumed to be the same.

Under scenario 2, 144 out of 725 pairs are stable against group mutation, and all of them yield perfect ingroup cooperation. The 140 pairs that survive in the original case (Section 3.2.2) also survive under the modified reputation update rule. The action rule in the additional four ($=144-140$) pairs is $(\sigma^{in}, \sigma^{out})$ = (Disc,AntiDisc). Another difference from the original case is that the action–norm pairs that yield partial ingroup favoritism in Table 3 realize full cooperation in the present case. Otherwise, the results are the same as those in the original case. In summary, 16 pairs realize full cooperation, and 128 pairs realize perfect ingroup favoritism. As is the case for scenario 1, only full cooperation is stable with standing or judging if the three subnorms are assumed to be the same.

Appendix B. The rest of the stable action–norm pairs under scenario 1

Under scenario 1 in the original case, 270 out of 440 stable action–norm pairs with a positive payoff realize perfect intragroup cooperation (Section 3.2.1). The other 170 stable action–norm pairs yielding $\pi > 0$ are summarized in Table 4. For all the stable action–norm pairs shown, σ^{in} = Disc. Table 4 indicates that outgroup favoritism does not occur.

There are 18 rows in Table 4. For the two action–norm pairs shown in the first row, the stability condition is given by $br^{in} > c$ and $r^{in} < 1/2$. For the two action–norm pairs shown in the sixth row, the stability condition is given by $br^{in} > c$ and $r^{in} > \sqrt{2}-1$. For the four action–norm pairs shown in the 16th row, the stability condition is given by $b/c > (1+r^{in})/r^{in}$. For all the other action–norm pairs, the stability condition is given by $br^{in} > c$.

References

Antal, T., Ohtsuki, H., Wakeley, J., Taylor, P.D., Nowak, M.A., 2009. Evolution of cooperation by phenotypic similarity. Proc. Nat. Acad. Sci. USA 106, 8597–8600.

Axelrod, R., 1984. Evolution of cooperation. Basic Books, NY.

Bernhard, H., Fischbacher, U., Fehr, E., 2006. Parochial altruism in humans. Nature 442, 912–915.

Boyd, R., Gintis, H., Bowles, S., Richerson, P.J., 2003. The evolution of altruistic punishment. Proc. Nat. Acad. Sci. USA 100, 3531–3535.

Brandt, H., Sigmund, K., 2005. Indirect reciprocity, image scoring, and moral hazard. Proc. Nat. Acad. Sci. USA 102, 2666–2670.

Brandt, H., Sigmund, K., 2006. The good, the bad and the discriminator—errors in direct and indirect reciprocity. J. Theor. Biol. 239 (2), 183–194.

Brewer, M.B., 1999. The psychology of prejudice: ingroup love and outgroup hate? J. Soc. Issues 55, 429–444.

Brown, R., 2000. Group Processes, second ed. Blackwell Publishing, Malden, MA.

Bshary, R., Grutter, A.S., 2006. Image scoring and cooperation in a cleaner fish mutualism. Nature 441 (7096), 975–978.

Chalub, F.A.C.C., Santos, F.C., Pacheco, J.M., 2006. The evolution of norms. J. Theor. Biol. 241, 233–240.

Choi, J.K., Bowles, S., 2007. The coevolution of parochial altruism and war. Science 318, 636–640.

De Cremer, D., van Vugt, M., 1999. Social identification effects in social dilemmas: a transformation of motives. Eur. J. Soc. Psychol. 29, 871–893.

Dovidio, J.F., Glick, P., Rudman, L.A., (Eds.), 2005. On the Nature of Prejudice. Blackwell Publishing, Malden, MA.

Efferson, C., Lalive, R., Fehr, E., 2008. The coevolution of cultural groups and ingroup favoritism. Science 321, 1844–1849.

Fortunato, S., 2010. Community detection in graphs. Phys. Rep. 486, 75–174.

Fowler, J.H., Kam, C.D., 2007. Beyond the self: social identity, altruism, and political participation. J. Politics 69, 813–827.

García, J., van den Bergh, J.C.J.M., 2011. Evolution of parochial altruism by multilevel selection. Evol. Human Behav. 32, 277–287.

Goette, L., Huffman, D., Meier, S., 2006. The impact of group membership on cooperation and norm enforcement: evidence using random assignment to real social groups. Am. Econ. Rev. 96, 212–216.

Hewstone, M., Rubin, M., Willis, H., 2002. Intergroup bias. Annu. Rev. Psychol. 53, 575–604.

Ihara, Y., 2011. Evolution of culture-dependent discriminate sociality: a gene-culture coevolutionary model. Phil. Trans. R. Soc. London B 366, 889–900.

Jansen, V.A.A., van Baalen, M., 2006. Altruism through beard chromodynamics. Nature 440, 663–666.

Jones, E.E., Wood, G.C., Quattrone, G.A., 1981. Perceived variability of personal characteristics in in-groups and out-groups: the role of knowledge and evaluation. Pers. Soc. Psychol. Bull. 7, 523–528.

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

Lize, A., Carval, D., Cortesero, A.M., Fournet, S., Poinsot, D., 2006. Kin discrimination and altruism in the larvae of a solitary insect. Proc. R. Soc. B 273, 2381–2386.

Masuda, N., Ohtsuki, H., 2007. Tag-based indirect reciprocity by incomplete social information. Proc. R. Soc. B 274, 689–695.

Mifune, N., Hashimoto, H., Yamagishi, T., 2010. Altruism toward in-group members as a reputation mechanism. Evol. Human Behav. 31, 109–117.

Nakamura, M., Masuda, N., 2011. Indirect reciprocity under incomplete observation. PLoS Comput. Biol. 7, e1002113.

Nowak, M.A., 2006a. Evolutionary Dynamics. The Belknap Press of Harvard University Press, MA.

Nowak, M.A., 2006b. Five rules for the evolution of cooperation. Science 314, 1560–1563.

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

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

Nowak, M.A., Sigmund, K., 2005. Evolution of indirect reciprocity. Nature 437, 1291–1298.

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., 2007. Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. J. Theor. Biol. 244, 518–531.

Ohtsuki, H., Iwasa, Y., Nowak, M.A., 2009. Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. Nature 457, 79–82.

Ostrom, T.M., Sedikides, C., 1992. Out-group homogeneity effects in natural and minimal groups. Psychol. Bull. 112, 536–552.

Pacheco, J.M., Santos, F.C., Chalub, F.A.C.C., 2006. Stern-judging: a simple, successful norm which promotes cooperation under indirect reciprocity. PLoS Comput. Biol. 2, 1634–1638.

Rand, D.G., Pfeiffer, T., Dreber, A., Sheketoff, R.W., Wernerfelt, N.C., Benkler, Y., 2009. Dynamic remodeling of in-group bias during the 2008 presidential election. Proc. Nat. Acad. Sci. USA 106, 6187–6191.

Sedikides, G., Schopler, J., Insko C.A., (Eds.), 1998. Intergroup Cognition and Intergroup Behavior. Psychology Press, New York.

Shinada, M., Yamagishi, T., Ohmura, Y., 2004. False friends are worse than bitter enemies: "Altruistic" punishment of in-group members. Evol. Human Behav. 25, 379–393.

Sigmund, K., Hauert, C., Nowak, M.A., 2001. Reward and punishment. Proc. Nat. Acad. Sci. USA 98, 10757–10762.

Suzuki, Y., Toquenaga, Y., 2005. Effects of information and group structure on evolution of altruism: analysis of two-score model by covariance and contextual analyses. J. Theor. Biol. 232, 191–201.

Traulsen, A., 2008. Mechanisms for similarity based cooperation. Eur. Phys. J. B 63, 363–371.

Traulsen, A., Nowak, M.A., 2007. Chromodynamics of cooperation in finite populations. PLoS ONE 2, e270.

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

West, S.A., Griffin, A.S., Gardner, A., 2007. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. J. Evol. Biol. 20, 415–432.

West, S.A., Griffin, A.S., Gardner, A., 2008. Social semantics: how useful has group selection been? J. Evol. Biol. 21, 374–385.

Yamagishi, T., Jin, N., Kiyonari, T., 1999. Bounded generalized reciprocity—ingroup boasting and ingroup favoritism. In: Advances in Group Processes, vol. 16, pp. 161–197.

Yamagishi, T., Jin, N., Miller, A.S., 1998. In-group bias and culture of collectivism. Asian J. Soc. Psychol. 1, 315–328.

Yamagishi, T., Mifune, N., 2008. Does shared group membership promote altruism? Rationality Soc. 20, 5–30.