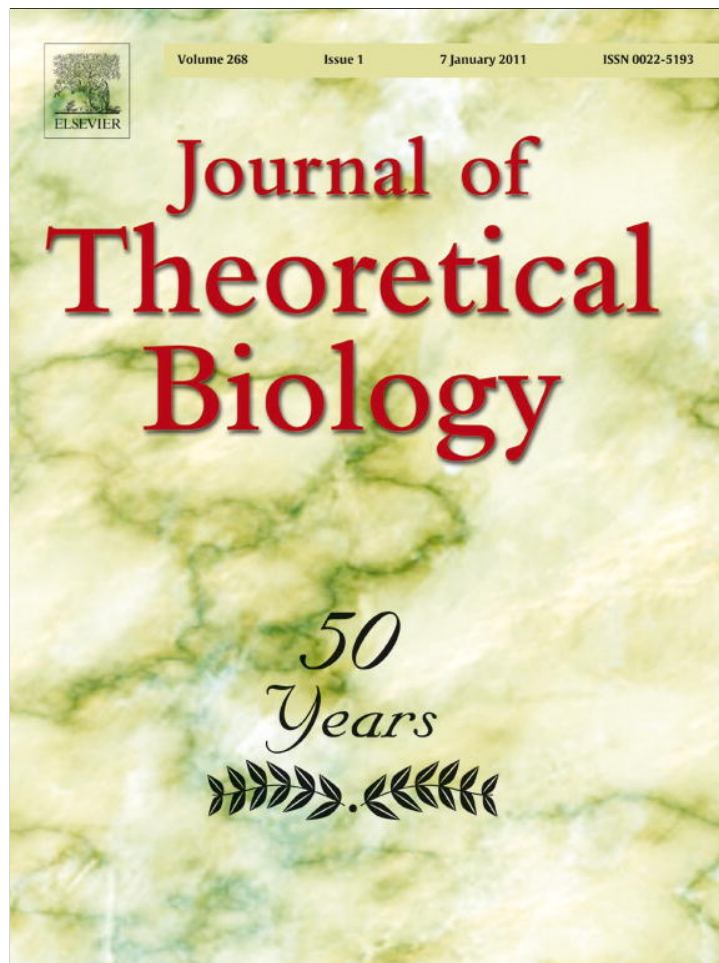


Provided for non-commercial research and education use.  
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

## Journal of Theoretical Biology

journal homepage: [www.elsevier.com/locate/yjtbi](http://www.elsevier.com/locate/yjtbi)

## Letter to Editor

## Strengthening strong reciprocity

## ARTICLE INFO

## Keywords:

Altruistic cooperation  
Altruistic punishment

Gintis (2000) presents a model of the evolutionary emergence of strong reciprocity that may in part explain a high level of sociality in human groups despite a low level of relatedness among members. A strong reciprocator is predisposed to cooperate with others and punish defectors at a cost to himself. Gintis (2000) argues that those groups that have strong reciprocators as members might have an evolutionary advantage under conditions of high social instability, where cooperation based on self-regarding reciprocity, the so-called reciprocal altruism (Trivers, 1971), will collapse. However, both the argument that leads to Eq. (8), which provides the condition for self-interested agents cooperating, and the derivation that leads to Eq. (16), which provides the condition for an increase in the fraction of strong reciprocators in the population, are inaccurate.

Moreover, several important papers have noted that Gintis (2000) assumes without justification that strong reciprocators are obligate cooperators, and have even speculated that strong reciprocity cannot be evolutionarily stable unless strong reciprocators are assumed to cooperate unconditionally (West and Gardner, 2010). In response, several researchers have dropped this assumption, studying the evolution of strong reciprocity assuming agents have fixed behaviors independent from payoffs, but with punishment and cooperation unlinked (Sigmund et al., 2001; Nakamaru and Iwasa, 2005, 2006; Nakamaru and Dieckmann, 2009). Here we show that, using the original Gintis (2000) model, except that cooperation and punishment are unlinked, so that all agents exhibit discretionary behavior by reacting optimally in deciding whether or not to cooperate, strong reciprocity can evolve under plausible conditions.

We first correct the two errors in Gintis (2000). Then we add an analysis of another two cases that have turned out to be important in the literature: (i) there are no surveillance costs, but punishment costs are incurred when agents intend to cooperate but mistakenly defect and (ii) we drop the assumption that the fraction of strong reciprocators is common knowledge. Our results strengthen the notion that strong reciprocity can lower the probability of group extinction in situations where groups are frequently threatened and thus cooperation based on reciprocal altruism collapses.

We begin by introducing the model of strong reciprocity and the notations used in Gintis (2000). In each period, a group of  $n$  agents faces a public goods game in which each member, by

sacrificing an amount  $c > 0$ , contributes an amount  $b > c$  shared equally by the other members of the group. If all members cooperate, each receives a net payoff of  $b - c > 0$ . However, the only Nash equilibrium in this game is universal defection, in which no member contributes, and thus the benefits of public goods are forgone. Suppose at the end of each period the game is continued for an additional period with probability  $\delta \in (0, 1)$ . Theorem 1 in Gintis (2000) states that cooperation can be sustained in repeated interactions if and only if  $c/b \leq \delta$ .

Empirical work in population ecology has shown that many animal species, including large mammals, are characterized by periods of growth interrupted by relatively infrequent crashes brought about by starvation, disease, and other environmental factors that may be, to varying extents, density independent (Boone and Kessler, 1999). Accordingly, we suppose that all agents cooperate in a good period and the group persists into the next period with probability  $\delta^*$ , while in a bad period, which occurs with probability  $p > 0$ , all agents defect and the group dissolves with probability one. We also suppose that in a bad period the group persists with probability  $\delta_* < \delta^*$  provided all members cooperate. Here we investigate whether strong reciprocity helps to sustain cooperation among group members in a bad period, and thus lower the probability of the group dissolving.

Suppose there are groups  $i = 1, 2, \dots, m$ , and let  $f_i$  be the fraction of strong reciprocators in group  $i$ . Each strong reciprocator can inflict a total amount of harm  $h > 0$  on non-cooperators, at a personal cost of surveillance  $c_r > 0$ . The question is: under what condition will an agent choose to cooperate in the bad period? In other words, under what condition is “cooperate” a Nash equilibrium in the bad period? To answer this, we must assume all agents cooperate and ask if one agent can gain by deviating; i.e., by defecting.

Suppose that the group continues with probability  $\delta_d < \delta_*$  if exactly one member defects, and that the probability of a defection being caught is  $1/n_i$ , where  $n_i$  is the size of group  $i$ , considering that all agents are suspects with the same probability if a defection happens. Then, the expected harm to one defector imposed by strong reciprocators is  $n_i f_i h / n_i = f_i h$ , and each strong reciprocator pays the expected cost of punishment  $c_r / n_i$ .

If all group members cooperate, the expected fitness  $\pi$  before the state of the period is revealed satisfies the recursion equation

$$\pi = b - c + [(1 - p)\delta^* + p\delta_*]\pi, \quad (1)$$

and thus we have

$$\pi = \frac{b-c}{1-\delta^*+p(\delta^*-\delta_*)} \quad (2)$$

If one member defects and all the others cooperate, the expected fitness  $\pi_d$  of the defector before the state of the period is revealed satisfies the recursion equation

$$\pi_d = b - [(1-p)c + pf_i(h + \mu c_r/n_i)] + [(1-p)\delta^* + p\delta_d]\pi_d, \quad (3)$$

where parameter  $\mu = 1$  for a strong reciprocator and  $\mu = 0$  for a self-interested agent. Hence we have

$$\pi_d = \frac{b-c+p[c-f_i(h+\mu c_r/n_i)]}{1-\delta^*+p(\delta^*-\delta_d)} \quad (4)$$

Then we can infer that for one defector in group  $i$  in the bad period the cost of defecting is  $f_i(h + \mu c_r/n_i) - \delta_d \pi_d$ , while the cost of cooperating is  $c - \delta_* \pi$ . If the former is not less than the latter, then self-interested agents choosing to cooperate in the bad period is a Nash equilibrium. Algebraic manipulation gives us that if the fraction of strong reciprocators in a group is at least

$$f_* = \frac{c - \delta_* \pi}{h} + \frac{\delta_d}{h} \cdot \frac{b-c+p\delta_*\pi}{1-(1-p)\delta_*} \quad (5)$$

complete cooperation will hold in the bad period. On the other hand, if  $f_i < f_*$  then strong reciprocators, as well as self-interested agents, will choose to defect in a bad period. That is, strong reciprocators defect under the same conditions as self-interested agents.

In the following we consider the severest situation where  $\delta_d = 0$ , which means that a single defection leads to the group dissolving with probability one. Alternative assumptions concerning  $\delta_d$  do not substantively alter our results. Then Eq. (5) is simplified as

$$f_* = \frac{c - \delta_* \pi}{h} \quad (6)$$

which is the lower bound of Eq. (5) [compared to the incorrect argument that leads to Eq. (8) in Gintis, 2000]. Of course, we have to assume that  $h > c - \delta_* \pi$ ; that is, the harm that a strong reciprocator can impose is greater than the cost of cooperating when all members cooperate. It straightforwardly follows that  $f_*$  in Eq. (6) decreases with increasing  $\delta_*$  or decreasing  $p$ , and that  $f_*$  approaches zero with increasing  $h$ .

Suppose that the fraction of strong reciprocators in a group is a common knowledge. If  $f_i < f_*$  there will be no cooperation in a bad period and the group will dissolve. The fitness  $\pi_s$  of members in such non-cooperative groups satisfies the recursion equation  $\pi_s = (1-p)(b-c + \delta^* \pi_s)$ , so

$$\pi_s = \frac{(1-p)(b-c)}{1-(1-p)\delta^*} \quad (7)$$

The relative fitness benefit from being in a cooperative group is thus

$$\pi - \pi_s = p\pi \left[ 1 + \frac{(1-p)\delta_*}{1-(1-p)\delta_*} \right] > 0. \quad (8)$$

We assume that the cooperation cost  $c$  satisfies the following inequalities:  $\delta_* \pi < c < \delta^* \pi_s$ , or equivalently,

$$\frac{\delta_*}{1-(1-p)(\delta^*-\delta_*)} < \frac{c}{b} < (1-p)\delta^*. \quad (9)$$

This implies  $\delta_* \pi < c < \delta^* \pi$ . Therefore, there is full cooperation in a good period, while all agents in any group always defect in a bad period if there are no strong reciprocators.

Let  $q_i$  be the fraction of the population in group  $i$  and  $\pi_i$  be the mean fitness of group  $i$ . Then,  $\bar{\pi} = \sum_i q_i \pi_i$  is the mean fitness of the

whole population, and  $\bar{f} = \sum_i q_i f_i$  is the mean fraction of strong reciprocators in the population. Gintis (2000) assumes that new groups in the next generation formed by the assignment of strong reciprocators and self-interested agents in proportion to their frequency in the population, and investigates the evolutionary emergence of strong reciprocity by calculating the change  $\Delta \bar{f}$  between two generations by means of Price's equation:

$$\Delta \bar{f} = \frac{1}{\bar{\pi}} E(\pi \Delta f) + \frac{1}{\bar{\pi}} \text{Cov}(\pi, f). \quad (10)$$

However, the calculations in Eqs. (13)–(15) of Gintis (2000) are not correct, so the condition for an increase in the fraction of strong reciprocators in the population provided by Eq. (16), and thus the following analysis, must be revised, even though we do not drop the assumption that strong reciprocators are obligate cooperators.

**Case A: Surveillance costs:** In a group where  $f_i = f_*$  all agents will cooperate in a bad period, because the expected harm inflicted on them would be  $f_* h$  if they defect when every strong reciprocator pays a surveillance cost  $c_r$  to entice group members to cooperate. Then in a group where  $f_i > f_*$  it is not necessary for every strong reciprocator to pay  $c_r$ . To ensure an expected loss of  $f_* h$  for any agent if he defects, it is sufficient for every strong reciprocator to pay  $f_* c_r / f_i$ . That is to say, when  $f_i = f_*$ , each strong reciprocator must pay  $c_r$ , but for  $f_i > f_*$ , each strong reciprocator pays only  $f_* c_r / f_i < c_r$ . This is in contrast to the assumption in Gintis (2000) that strong reciprocators always pay the largest cost of surveillance  $c_r$  in cooperative groups.

In a cooperative group  $i$  where  $f_i \geq f_*$ , the mean fitness of a self-interested agent is  $\pi_{is} = \pi$ . However, each strong reciprocator pays an additional surveillance cost  $f_* c_r / f_i$  in the bad period, so the mean fitness of a strong reciprocator satisfies the recursion equation

$$\pi_{ir} = b - c - p c_r f_* / f_i + [(1-p)\delta^* + p\delta_*] \pi_{ir}. \quad (11)$$

Hence we have  $\pi_{ir} = \pi - \pi_r / f_i$ , where

$$\pi_r = \frac{f_* p c_r}{1-\delta^*+p(\delta^*-\delta_*)} \quad (12)$$

Relative to a self-interested agent, the fitness loss of a strong reciprocator due to surveillance is  $\pi_r / f_i$ . Then the mean fitness of a random member in group  $i$  is  $\pi_i = f_i \pi_{ir} + (1-f_i) \pi_{is} = \pi - \pi_r$ .

We define the fraction of the population in cooperative groups  $q_f = \sum_{f_i \geq f_*} q_i$ , and the mean fraction of strong reciprocators in cooperative (resp., non-cooperative) groups  $f_c = \sum_{f_i \geq f_*} q_i f_i / q_f$  [resp.,  $f_s = \sum_{f_i < f_*} q_i f_i / (1-q_f)$ ]. If the relative fitness benefit from being in a cooperative group  $\sum_{f_i \geq f_*} q_i \pi_i / q_f - \pi_s = \pi - \pi_r - \pi_s$  is non-positive, cooperative groups could not evolve. Hence we assume  $\pi_r < \pi - \pi_s$ , or equivalently,

$$\frac{c_r}{h} < \left[ 1 + \frac{(1-p)\delta_*}{1-(1-p)\delta_*} \right] \frac{b-c}{c - \delta_* \pi}. \quad (13)$$

If  $h$  is sufficiently large relative to  $c_r$ , then this condition can be satisfied without difficulty.

Groups grow from one generation to the next in proportion to their relative fitness, so the fraction of strong reciprocators in a cooperative group  $i$  in the next generation is  $f'_i = \pi_{ir} f_i / \pi_i$ . Considering that  $f'_i = f_i$  in non-cooperative groups where  $f_i < f_*$ , we have

$$\frac{1}{\bar{\pi}} E(\pi \Delta f) = \frac{1}{\bar{\pi}} \sum_{f_i \geq f_*} q_i f_i (\pi_{ir} - \pi_i) = -\frac{\pi_r}{\bar{\pi}} q_f (1-f_c), \quad (14)$$

where the mean fitness of the population  $\bar{\pi} = q_f (\pi - \pi_r) + (1-q_f) \pi_s$ . Note that  $\bar{\pi} \geq \pi_s > 0$ . Algebraic manipulation gives us

$$\frac{1}{\bar{\pi}} \text{Cov}(\pi, f) = \frac{1}{\bar{\pi}} \sum_i q_i \pi_i (f_i - \bar{f}) = \frac{\pi - \pi_r - \pi_s}{\bar{\pi}} q_f (f_c - \bar{f}), \quad (15)$$

where the mean fraction of strong reciprocators in the population  $\bar{f} = q_f f_c + (1 - q_f) f_s$ . Substituting (14) and (15) into (10) gives us

$$\Delta \bar{f} = \frac{q_f}{\pi} [(f_c - \bar{f})(\pi - \pi_s) - (1 - \bar{f})\pi_r]. \quad (16)$$

Clearly  $q_f = 0$  and  $\bar{f} = 1$  entail  $\Delta \bar{f} = 0$ . If rare strong reciprocators that invade a population of self-interested agents are randomly distributed in groups, and if there are sufficiently many groups,  $q_f$  will be strictly positive with probability one. Eq. (6) shows that the higher the  $h$ , the smaller the population needed to ensure  $q_f > 0$ .

We consider the evolutionary stability of the equilibrium of  $\bar{f} = 1$ . When rare self-interested agents invade a population of strong reciprocators, we will have  $\bar{f} < 1$ . If  $q_f = 1$ , then  $f_c = \bar{f}$  and thus  $\Delta \bar{f} < 0$ . That is, when all groups are cooperative and not the whole population are strong reciprocators, the fraction of strong reciprocators in the population will decrease in the next generation. So the equilibrium of  $\bar{f} = 1$  is not evolutionarily stable.

If  $\bar{f} < 1$  and  $f_c = 1$ , then  $\Delta \bar{f} > 0$ , which means that the fraction of strong reciprocators in the population will increase in the next generation. However, when  $\bar{f} < 1$  the state of  $f_c = 1$  cannot last forever in all of the following generations due to the randomness of forming groups of finite size.

In the situation where  $0 < q_f < 1$ , which implies  $0 < \bar{f} < 1$ , solving  $\Delta \bar{f} > 0$  gives us the condition for an increase in the fraction of strong reciprocators in the population in the next generation

$$\frac{1 - \bar{f}}{f_c - \bar{f}} < \frac{\pi - \pi_s}{\pi_r} = \left[ 1 + \frac{(1-p)\delta_*}{1 - (1-p)\delta^*} \right] \frac{b-c}{c - \delta_* \pi} \cdot \frac{h}{c_r} \quad (17)$$

[compared to Eq. (16) and Theorem 3 in Gintis, 2000]. The left-hand side term of Eq. (17) is not less than one, so the condition given in Eq. (13) is absolutely necessary. If  $h$  is sufficiently large relative to  $c_r$ , then Eq. (17) will be satisfied for rare strong reciprocators. On the other hand, when  $q_f$  approaches one,  $f_c$  will approach  $\bar{f}$  and thus Eq. (17) will be violated. This means that there exists at least one stable equilibrium of the fraction of strong reciprocators in the population. Therefore, we reach the following conclusion.

**Proposition 1.** Suppose the discount factor is  $\delta^*$  in a good period and  $\delta_*$  ( $< \delta^*$ ) in a bad period, and bad periods occur with probability  $p > 0$ . Suppose Eq. (9) holds, so there is cooperation in the good but not the bad periods in groups in which the fraction of strong reciprocators is less than  $f_*$  given by (6). Then self-interested agents can always invade a population of strong reciprocators, and when surveillance cost  $c_r$  is sufficiently low relative to harm  $h$  on non-cooperators, a small fraction  $\bar{f}$  of strong reciprocators can always invade a population of self-interested agents. Therefore the population will reach a heterogeneous equilibrium with both self-interested and strongly reciprocating types.

Compared to Theorem 4 in Gintis (2000), the one difference in this proposition is that the condition for a small fraction of strong reciprocators invading a population of self-interested agents is that surveillance cost  $c_r$  is sufficiently low relative to harm  $h$  on non-cooperators, not that surveillance cost  $c_r$  is sufficiently low. Due to the incorrect calculations in Eqs. (13)–(15) of Gintis (2000), Theorem 4 in Gintis (2000) fails to incorporate the effect of harm  $h$  on the evolution of strong reciprocity.

**Case B: Punishment costs due to error:** In this section we investigate a model in which there are no surveillance costs, but there is a cost of punishing defectors that is proportional to the amount of punishment actually delivered.

We assume that there is a probability  $\varepsilon > 0$  that an agent, whether self-interested or strongly reciprocating, who intends to

cooperate will defect by mistake and that this is a common knowledge. No other forms of error are involved. If all group members cooperate and there is no punishment, the expected fitness  $\pi_\varepsilon$  before the state of the period is revealed satisfies the recursion equation  $\pi_\varepsilon = (1 - \varepsilon)(b - c) + [(1 - p)\delta^* + p\delta_*]\pi_\varepsilon$ , so

$$\pi_\varepsilon = \frac{(1 - \varepsilon)(b - c)}{1 - \delta^* + p(\delta^* - \delta_*)} = (1 - \varepsilon)\pi. \quad (18)$$

Accordingly, in the severest situation where  $\delta_d = 0$ , if the fraction  $f_i$  of strong reciprocators is at least

$$f_* = \frac{c - \delta_* \pi_\varepsilon}{h}, \quad (19)$$

complete cooperation will hold in the group.

In a non-cooperative group  $i$  where  $f_i < f_*$  the fitness  $\pi_{se}$  of members satisfies the recursion equation  $\pi_{se} = (1 - p)[(1 - \varepsilon)(b - c) + \delta^* \pi_{se}]$ , so

$$\pi_{se} = \frac{(1 - p)(1 - \varepsilon)(b - c)}{1 - (1 - p)\delta^*} = (1 - \varepsilon)\pi_s. \quad (20)$$

Ensuring that there is full cooperation in a good period while full defection in a bad period both in cooperative and non-cooperative groups when there is no punishment requires that cooperation cost  $c$  should satisfy the following inequalities:  $\delta_* \pi_\varepsilon < c < \delta^* \pi_{se}$ , or equivalently,

$$\frac{1 - \varepsilon}{1 - (1 - p)(\delta^* - \delta_*) - \varepsilon} < \frac{c}{b} < \frac{1 - \varepsilon}{(1 - p)\delta^* - \varepsilon}, \quad (21)$$

which, in the situation of  $\varepsilon = 0$ , reduces to Eq. (9).

In a cooperative group  $i$  where  $f_i \geq f_*$  agents also make mistakes with probability  $\varepsilon > 0$ . However, if a defection is caught in the bad period, punishment will be actually executed. Otherwise, those inadvertent defections may cause a series of intentional defections and thus the group dissolving. The same argument as in Case A leads to that punishment will cost each strong reciprocator an amount  $f_* c_r / f_i$  and inflict harm  $f_* h$  on each agent who defects by mistake.

The mean fitness of a self-interested agent  $\pi_{is}$  satisfies the recursion equation

$$\pi_{is} = (1 - \varepsilon)(b - c) - p\varepsilon f_* h + [(1 - p)\delta^* + p\delta_*]\pi_{is}. \quad (22)$$

Solving this equation gives us  $\pi_{is} = \pi_\varepsilon - \pi_{he}$ , where

$$\pi_{he} = \frac{f_* \varepsilon p h}{1 - \delta^* + p(\delta^* - \delta_*)}. \quad (23)$$

Likewise, the mean fitness of a strong reciprocator  $\pi_{ir}$  satisfies the recursion equation

$$\pi_{ir} = (1 - \varepsilon)(b - c) - p\varepsilon f_* h - p\varepsilon f_* c_r / f_i + [(1 - p)\delta^* + p\delta_*]\pi_{ir}, \quad (24)$$

so  $\pi_{ir} = \pi_\varepsilon - \pi_{he} - \pi_{re} / f_i$ , where

$$\pi_{re} = \frac{f_* \varepsilon p c_r}{1 - \delta^* + p(\delta^* - \delta_*)}. \quad (25)$$

Relative to a self-interested agent, the fitness loss of a strong reciprocator due to punishment is  $\pi_{re} / f_i$ . Therefore, the mean fitness of a random member in group  $i$  is  $\pi_i = f_i \pi_{ir} + (1 - f_i) \pi_{is} = \pi_\varepsilon - \pi_{re} - \pi_{he}$ .

If the relative fitness benefit from being in a cooperative group  $\sum_{f_i \geq f_*} q_i \pi_i / q_f - \pi_{se} = \pi_\varepsilon - \pi_{re} - \pi_{he} - \pi_{se}$  is non-positive, cooperative groups could not evolve. Hence we assume  $\pi_{re} + \pi_{he} < \pi_\varepsilon - \pi_{se}$ , or equivalently,

$$\frac{c_r}{h} < \frac{1 - \varepsilon}{\varepsilon} \left[ 1 + \frac{(1 - p)\delta_*}{1 - (1 - p)\delta^*} \right] \frac{b - c}{c - \delta_* \pi_\varepsilon} - 1. \quad (26)$$

If  $\varepsilon$  is sufficiently low, then this condition can be satisfied without difficulty.

Similar manipulation as in Case A gives us the change of the mean fraction of strong reciprocators in the population in the next generation

$$\Delta \bar{f} = \frac{q_f}{\pi_\varepsilon} [(f_c - \bar{f})(\pi_\varepsilon - \pi_{se} - \pi_{he}) - (1 - \bar{f})\pi_{re}], \quad (27)$$

where the mean fitness of the population  $\bar{\pi}_\varepsilon = q_f(\pi_\varepsilon - \pi_{re} - \pi_{he}) + (1 - q_f)\pi_{se}$ . Note that  $\bar{\pi}_\varepsilon \geq \pi_{se} > 0$ .

Clearly  $q_f = 0$  and  $\bar{f} = 1$  entail  $\Delta \bar{f} = 0$ . The same arguments as in Case A show that the equilibrium of  $q_f = 0$  is trivial, that the equilibrium of  $\bar{f} = 1$  is not evolutionarily stable, and that the state of  $f_c = 1$  cannot last forever in all of the following generations.

In the situation where  $0 < q_f < 1$ , by solving  $\Delta \bar{f} > 0$  we get the condition for an increase in the fraction of strong reciprocators in the population in the next generation

$$\frac{1 - \bar{f}}{f_c - \bar{f}} < \frac{\pi_\varepsilon - \pi_{se} - \pi_{he}}{\pi_{re}} = \left\{ \frac{1 - \varepsilon}{\varepsilon} \left[ 1 + \frac{(1 - p)\delta^*}{1 - (1 - p)\delta^*} \right] \frac{b - c}{c - \delta^* \pi_\varepsilon} - 1 \right\} \frac{h}{c_r}. \quad (28)$$

As in Case A, this also shows that the condition given in Eq. (26) is necessary. When  $q_f$  approaches one,  $f_c$  will approach  $\bar{f}$  and thus Eq. (28) will be violated. However, if  $\varepsilon$  is sufficiently low, then Eq. (28) will be satisfied for either rare or relatively large fraction of strong reciprocators. Therefore, we reach the following conclusion.

**Proposition 2.** Suppose the discount factor is  $\delta^*$  in a good period and  $\delta_*( < \delta^*)$  in a bad period, and bad periods occur with probability  $p > 0$ . Suppose Eq. (21) holds, so there is cooperation in the good but not the bad periods in groups in which the fraction of strong reciprocators is less than  $f_*$  given by (19). Then self-interested agents can always invade a population of strong reciprocators, and if probability  $\varepsilon$  for agents defecting by mistake is sufficiently low, a small fraction  $\bar{f}$  of strong reciprocators can always invade a population of self-interested agents and will reach a relatively high level of cooperation. Therefore the population will reach a heterogeneous equilibrium with both self-interested and strongly reciprocating types.

Case C: Punishment costs based on a quorum: Boyd et al. (2010) derive strong reciprocity by unlinking cooperation and punishment, and assume that strong reciprocators only punish when they form a sufficiently large fraction of the group (a “quorum”). We now show that this mechanism supports strong reciprocity as well in the framework of Gintis (2000).

We still assume as in Case B that there is a probability  $\varepsilon > 0$  that an agent who intends to cooperate will defect by mistake. As mentioned before, in the severest situation where  $\delta_d = 0$ , if the fraction  $f_i$  of strong reciprocators is at least

$$f_* = \frac{c - \delta_* \pi_\varepsilon}{h}, \quad (29)$$

complete cooperation will hold in the group.

However, we drop the assumption that the fraction of strong reciprocators in a group is a common knowledge. Instead, as in Boyd et al. (2010), we assume that in each group there is a signaling stage just before playing the repeated public goods game, in which strong reciprocators signal their intent to punish defectors. Suppose that the cost of signaling,  $c_g$ , is high enough so that it does not pay to signal and then fail to punish. All agents defect in the bad period unless there is a quorum  $f_*$  of strong reciprocators. By this way we also drop the assumption that strong reciprocators are obligate cooperators in favor of the assumption that they cooperate under exactly the same conditions as self-interested agents.

Accordingly, in a cooperative group  $i$  where  $f_i \geq f_*$ , the mean fitness of a self-interested agent  $\pi_{is} = \pi_\varepsilon - \pi_{he}$ , while the mean

fitness of a strong reciprocator  $\pi_{ir} = \pi_\varepsilon - \pi_{he} - \pi_{re} / f_i - c_g$ . Therefore, the mean fitness of a random member in group  $i$  is  $\pi_i = \pi_\varepsilon - \pi_{re} - \pi_{he} - f_i c_g$ .

In a non-cooperative group  $i$  where  $f_i < f_*$ , the mean fitness of a self-interested agent  $\pi_{is} = \pi_{se}$ , while the mean fitness of a strong reciprocator  $\pi_{ir} = \pi_{se} - c_g$ . Therefore, the mean fitness of a random member in group  $i$  is  $\pi_i = \pi_{se} - f_i c_g$ . We still assume that Eq. (26) holds. Otherwise cooperative groups could not evolve.

The mean fitness of the population  $\bar{\pi}_\varepsilon = q_f(\pi_\varepsilon - \pi_{re} - \pi_{he}) + (1 - q_f)\pi_{se} - c_g \bar{f}$ . Here we assume that the cost of signaling  $c_g$  satisfies

$$c_g < \pi_{se}, \quad (30)$$

which means that the cost of signaling is less than the mean fitness in a non-cooperative group. Then, together with Eq. (26), we have  $\bar{\pi}_\varepsilon > 0$ .

The fraction of strong reciprocators in group  $i$  in the next generation is  $f_i = \pi_{ir} f_i / \pi_i$ , so we have

$$\frac{1}{\pi_\varepsilon} E(\pi \Delta f) = \frac{1}{\pi_\varepsilon} \sum_i q_i f_i (\pi_{ir} - \pi_i) = -\frac{1}{\pi_\varepsilon} \left[ \pi_{re} q_f (1 - f_c) + c_g \sum_i q_i f_i (1 - f_i) \right] \quad (31)$$

and

$$\begin{aligned} \frac{1}{\pi_\varepsilon} \text{Cov}(\pi, f) &= \frac{1}{\pi_\varepsilon} \sum_i q_i \pi_i (f_i - \bar{f}) \\ &= \frac{1}{\pi_\varepsilon} \left[ q_f (f_c - \bar{f})(\pi_\varepsilon - \pi_{re} - \pi_{he} - \pi_{se}) - c_g \sum_i q_i f_i (f_i - \bar{f}) \right]. \end{aligned} \quad (32)$$

Then, by Price's equation (10), we get the change of the mean fraction of strong reciprocators in the population in the next generation

$$\Delta \bar{f} = \frac{1}{\pi_\varepsilon} \{ q_f [(f_c - \bar{f})(\pi_\varepsilon - \pi_{se} - \pi_{he}) - (1 - \bar{f})\pi_{re}] - c_g \bar{f} (1 - \bar{f}) \}. \quad (33)$$

Clearly  $\bar{f} = 0$  and 1, which implies  $q_f = 0$  and 1, respectively, entail  $\Delta \bar{f} = 0$ . If  $q_f = 0$  and  $\bar{f} > 0$ , or if  $q_f = 1$  and  $\bar{f} < 1$ , then  $\Delta \bar{f} < 0$ . In the situation where  $0 < q_f < 1$ , by solving  $\Delta \bar{f} > 0$  we get the condition for an increase in the fraction of strong reciprocators in the population in the next generation is that Eq. (28) holds and the fraction of the population in cooperative groups satisfies

$$q_f > \frac{c_g \bar{f} (1 - \bar{f})}{(f_c - \bar{f})(\pi_\varepsilon - \pi_{se} - \pi_{he}) - (1 - \bar{f})\pi_{re}}, \quad (34)$$

which, in the situation of  $\varepsilon = 0$ , reduces to

$$q_f > \frac{c_g \bar{f} (1 - \bar{f})}{(\pi - \pi_s)(f_c - \bar{f})}. \quad (35)$$

If probability  $\varepsilon$  for agents defecting by mistake is sufficiently low, Eq. (28) will be satisfied for either rare or relatively large fraction of strong reciprocators. If the cost of signaling  $c_g$  is sufficiently low relative to the fitness benefit from being in a cooperative group  $\pi - \pi_s$  given in Eq. (8), then the right-hand side terms of Eq. (34) and (35) are very small. Therefore, we reach the following conclusion.

**Proposition 3.** Suppose the discount factor is  $\delta^*$  in a good period and  $\delta_*( < \delta^*)$  in a bad period, and bad periods occur with probability  $p > 0$ . Suppose Eq. (21) and (30) hold. Strong reciprocators bear a cost  $c_g$  to signal their intent to punish defectors just before playing the repeated public goods game. All agents defect in the bad period unless there is a quorum  $f_*$  given by (29). Then self-interested agents can

always invade a population of strong reciprocators, and if probability  $\varepsilon$  for agents defecting by mistake is sufficiently low and the cost of signaling  $c_g$  is sufficiently low relative to the fitness benefit from being in a cooperative group  $\pi - \pi_s$ , a small fraction  $\bar{f}$  of strong reciprocators can always invade a population of self-interested agents and will reach a relatively high level of cooperation. Therefore the population will reach a heterogeneous equilibrium with both self-interested and strongly reciprocating types.

To sum up, in this paper we extend the results of Gintis (2000) by analyzing three different models without assuming that strong reciprocators are obligate cooperators. In the first model we have rectified both the incorrect argument that leads to Eq. (8) and the incorrect calculations in Eqs. (13)–(15), and thus the following analysis, of Gintis (2000). In the second model we have dispensed with the concept of surveillance cost. And in the third model we have further dropped the assumption that the fraction of strong reciprocators in a group is a common knowledge. Our analysis shows that the model originating from Gintis (2000) can support the evolutionary emergence of strong reciprocity, and that under plausible conditions rare strong reciprocators can always invade a population of self-interested agents and keep at a relatively high fraction. These results extend the research on strong reciprocity to large groups the behaviors of whose members are contingent on their payoffs (Sigmund et al., 2001; Nakamaru and Iwasa, 2005, 2006; Nakamaru and Dieckmann, 2009), and further strengthen the notion behind Gintis (2000) that strong reciprocity may lower the probability of group extinction in situations where groups are frequently threatened and thus cooperation based on reciprocal altruism collapses.

The comments and suggestions of the reviewer are gratefully acknowledged. This work was supported by the National Natural Science Foundation of China (Nos. 60974064, 60736022, and 60674047) and the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20060001013).

## References

- Boone, J.L., Kessler, K.L., 1999. More status or more children? Social status, fertility reduction, and long-term fitness. *Evol. Hum. Behav.* 20, 257–277.
- Boyd, R., Gintis, H., Bowles, S., 2010. Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* 328, 617–620.
- Gintis, H., 2000. Strong reciprocity and human sociality. *J. Theor. Biol.* 206, 169–179.
- Nakamaru, M., Dieckmann, U., 2009. Runaway selection for cooperation and strict-and-severe punishment. *J. Theor. Biol.* 257, 1–8.
- Nakamaru, M., Iwasa, Y., 2005. The evolution of altruism by costly punishment in the lattice structured population: score-dependent viability versus score-dependent fertility. *Evol. Ecol. Res.* 7, 853–870.
- Nakamaru, M., Iwasa, Y., 2006. The coevolution of altruism and punishment: role of the selfish punisher. *J. Theor. Biol.* 240, 475–488.
- Sigmund, K., Hauert, C., Nowak, M.A., 2001. Reward and punishment. *Proc. Natl. Acad. Sci. USA* 98, 10757–10762.
- Trivers, R.L., 1971. The evolution of reciprocal altruism. *Q. Rev. Biol.* 46, 35–57.
- West, S.A., Gardner, A., 2010. Altruism, spite, and greenbeards. *Science* 327, 1341–1344.

Kuiying Deng

State Key Laboratory for Turbulence and Complex Systems,  
College of Engineering, Peking University, Beijing 100871, China  
E-mail address: RossDeng@pku.edu.cn

Herbert Gintis\*

Santa Fe Institute and Central European University,  
15 Forbes Avenue, Northampton, MA 01060, USA  
E-mail address: hgintis@comcast.net

Tianguang Chu

State Key Laboratory for Turbulence and Complex Systems,  
College of Engineering, Peking University, Beijing 100871, China  
E-mail address: chutg@pku.edu.cn

Received 26 July 2010

\* Corresponding author. Tel.: +1 413 586 7756; fax: +1 775 402 4921.