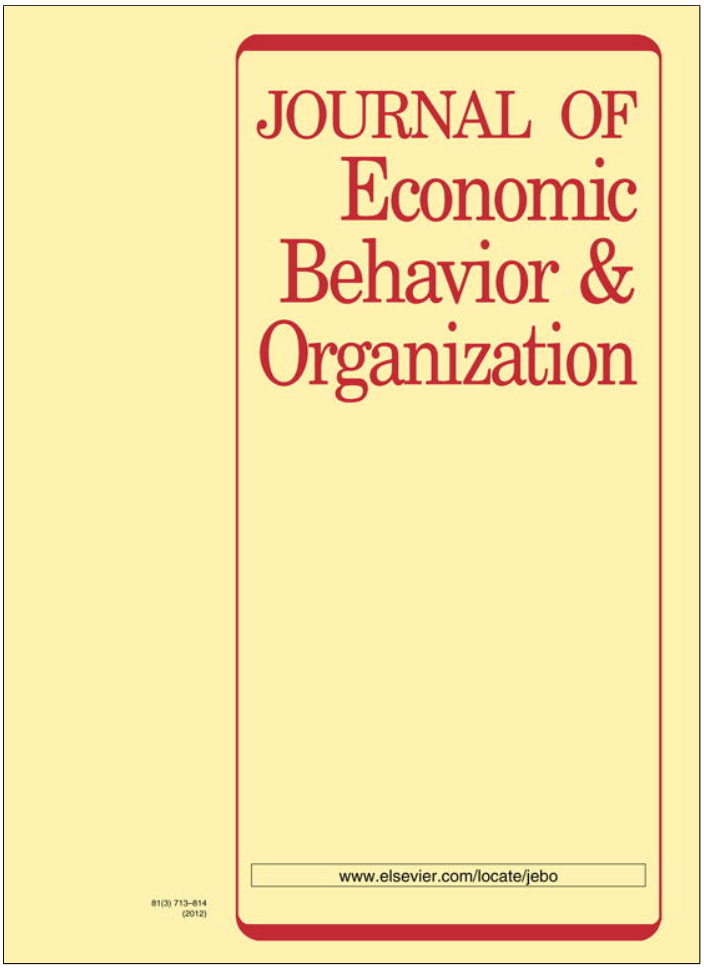


(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

Journal of Economic Behavior & Organization 83 (2012) 330–341

Contents lists available at SciVerse ScienceDirect

Journal of Economic Behavior & Organization

journal homepage: www.elsevier.com/locate/jebo

# Is altruism bad for cooperation?☆

Sung-Ha Hwang[a], Samuel Bowles[b,c,*]

[a] School of Economics, Sogang University, Republic of Korea
[b] Santa Fe Institute, USA
[c] University of Siena, Italy

ARTICLE INFO

*Article history:*
Received 11 July 2011
Received in revised form 29 May 2012
Accepted 20 June 2012
Available online xxx

*JEL classification:*
D64 (altruism)
H41 (public goods)
D03 (behavioral economics)

*Keywords:*
Public goods
Altruism
Spite
Reciprocity
Punishment
Cooperation

ABSTRACT

Some philosophers and social scientists have stressed the importance for good government of an altruistic citizenry that values the well being of fellow citizens. Economists, however, have emphasized the need for incentives that induce even the self-interested to contribute to the public good. Implicitly most have assumed that these two approaches are complementary or at worst additive. But this need not be the case. Behavioral experiments find that if reciprocity-minded subjects feel hostility towards free riders and enjoy inflicting harm on them, the incentives provided by the anticipated punishment support near efficient levels of contributions to a public good. Cooperation may also be supported if altruistic individuals internalize the group benefits that their contributions produce. But the effects of these two supports for high levels of cooperation may be less than additive. Using a utility function embodying both reciprocity and altruism we show that unconditional altruism attenuates the punishment motive and thus may reduce the level of punishment inflicted on defectors, resulting in lower levels of contribution. Increases in altruism may also reduce the level of benefits from the public project net of contribution costs and punishment costs. The range over which altruism inhibits cooperation and reduces material payoffs is greater, the stronger is the reciprocity motive among group members.



## 1. Introduction

When Adam Smith famously proposed that the actions of the self interested economic man might implement a socially desirable allocation of resources he added "Nor is it always worse for the society that. . .he intends only his own gain" (Smith, 1776). Smith was aware, of course, that in public goods settings and other social dilemmas, a concern for the well being of others may improve allocational efficiency. But economists have also elaborated the dark side of other-regarding preferences, including the way that inequity averse preferences may result in a smaller joint surplus, and hostility towards outsiders may restrict opportunities for mutually beneficial exchange. Simple, unconditional altruism, however, seems an unlikely candidate as the culprit in such deviations from efficient allocation. But we will show that altruism may reduce contributions to a public good, resulting in a smaller joint surplus than otherwise would be available.

Both altruism and reciprocity may motivate individuals to contribute to the provision of a public good. Altruism induces the individual to unconditionally value the payoff of other individuals, while reciprocity implies a valuation of the others' payoffs that is conditional on their contributions (or other indications of their type). Reciprocators may value the payoffs of

☆ We thank the Behavioral Science Program of the Santa Fe Institute, the U.S. National Science Foundation, the University of Siena and the European Science Foundation for support of this project and Roland Benabou, Theodore Bergstrom, Drew Fudenberg, Simon Gaechter, David Levine, Louis Putterman, Rajiv Sethi, Joel Sobel, Elisabeth Wood, and two anonymous referees for helpful comments.

* Corresponding author at: Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA.
*E-mail addresses:* sunghah@sogang.ac.kr (S.-H. Hwang), bowles@santafe.edu (S. Bowles).

0167-2681/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jebo.2012.06.001

low contributors negatively and be motivated to reduce the payoffs of defectors at a cost to themselves, when this option is available. The prospect of punishment of low contributions may induce individuals to contribute more than they otherwise would, thereby sustaining cooperation in groups where formal constraints and incentives are insufficient (Fehr and Gaechter, 2000; Anderson and Putterman, 2006; Boyd et al., 2010).

We explore the possibility that the two motives for contribution – a positive valuation of the payoffs of others and a desire to avoid the punishment induced by a negative valuation of one's payoffs by others – may work at cross purposes. Specifically we show that by attenuating the punishment motive, a general increase in the level of unconditional altruism may reduce rather than increase contributions.

Thus, while one often refers to individuals as being 'cooperative' or 'uncooperative', the motives supporting high levels of cooperation are heterogeneous, and they need not work synergistically. For example, experimental evidence indicates that unconditional altruists among American student subjects are significantly less likely to punish low contributors in a public goods game (Carpenter et al., 2009): a standard deviation increase in an individual's level of altruism reduced the amount he spent on punishment by 0.13 standard deviations. Also consistent with a possible conflict between altruistic motives for contributing to a public good and a willingness to discipline free riders is the fact that, among Russian urban and rural adults, those who contributed more to the public good punished low contributors significantly less, conditional on (a) the contribution level of others, (b) the amount by which the target of the punishment contributed less than the punisher, and (c) a large number of demographic and occupational controls (Gaechter and Hermann, 2011).

We model a public goods problem in which voluntary contributions are sustained both because altruistic citizens value the benefits conferred on others and reciprocal citizens punish free riders. We show that an increase in the level of altruism in a population may reduce the level of contributions and the benefits of the public project. This occurs because more altruistic individuals, while predisposed to contribute, are less willing to punish free riders. The idea behind the result – that seemingly good motives need not be synergistic – is as ancient as the contrast between the retribution-based morality typically attributed to the Old Testament and the unconditional generosity advocated in parts of the New Testament.

A key assumption in our model is that people have an intrinsic motivation to punish shirkers, not simply an instrumental desire to alter their behavior or to affect the distribution of payoffs so as to either reduce unfairness or to enhance the punisher's own relative payoffs. This is similar to what Boyd and Richerson (1992) call retribution punishment and the analogue of Andreoni's (1990) warm glow altruism. That subjects view punishment of shirkers also as retribution rather than simply as instrumental towards affecting behavior is consistent with the recent public goods with punishment experiment of Falk et al. (2005). The game was one shot, ruling out behavior modification as a motive for punishing low contributors, and the punishment technology was such that punishment could not alter the difference in payoff between the punisher and the target (the cost to the punisher was the same as that inflicted on the target). Nonetheless, sixty per cent of cooperators punished defectors.

Further evidence for our assumption that punishment is non-strategic comes from the public goods experiment of Fudenberg and Pathak (2010). As in the standard game, following each round of contributions subjects were given information on the contributions of fellow group members and had the opportunity to deduct some of their own payoffs in order to lower the payoffs of another in the group. But unlike the usual treatment in which the targets of punishment are informed of the level of punishment received after each round, in the Fudenberg and Pathak experiment the levels of punishment were not to be revealed until the experiment was over, and those who punished others knew this. Thus the experimental design ruled out modifying the behavior of shirkers as a motive for punishment. Consistent with what the authors term a "pure preference" motivation for punishment, subjects nonetheless punished shirkers, leading the authors to conclude that "agents enjoy punishment, where 'enjoyment' includes anger and a desire for retribution." There is considerable further evidence for our non-strategic modeling of punishment (de Quervain et al., 2004; Casari and Luini, 2012; Gaechter and Hermann, 2011; Anderson and Putterman, 2006).

In the next section we use the ideas of Levine (1998), Rabin (1993), and Falk and Fischbacher (2006) to explore the joint effects of altruism toward fellow group members and reciprocity-based hostility towards low contributors in a public goods game. In Section 3 we study the Nash equilibrium levels of punishment and contribution under varying levels of unconditional altruism of the members of a group. We show first that the relationship between the level of altruism and contributions is non-monotonic, and that there exists a range of levels of altruism over which increases in altruism reduce both equilibrium levels of contribution and the sum of benefits from the public project, net of the costs of contributing and the costs of punishing. Second, we show that the range for which altruism is bad for both cooperation and net benefits is larger the more reciprocal are the group members. For simplicity of exposition and clarity of the underlying causal mechanisms we initially assume a homogeneous population. In Section 4 we extend this model to a heterogeneous population and show that our main results and key insights still hold. In heterogeneous populations we can also show that the greater the frequency of altruistic reciprocators in the population the wider is the range for which increased levels of altruism in their functions will decrease average contributions. In the penultimate section we consider a number of caveats and possible extensions. In the conclusion we suggest some implications for how social preferences may support cooperation despite the sometimes counterproductive effects of increased altruism and the costly nature of punishment. In Appendix A we present the proofs of Propositions 1 and 2.

Similar in spirit to our first result is the finding of Bernheim and Stark (1988) that increased altruism among two family members in a repeated game setting may be welfare-reducing (see also Nakao, 2008; Alger and Weibull, 2010). However, our setting is a non-repeated public goods game rather than a repeated dyadic interaction; and rather than the simple

withdrawal of cooperation from a shirking partner, punishment in our model is explicitly modeled as costly to the punisher and motivated by reciprocal preferences. We are thus able to show that altruism and reciprocity – two social preferences thought to contribute to public goods provision – may interact in counter-productive ways. Other than showing that altruism may sometimes have unintended effects, our model is unrelated to the "Samaritan's dilemma" arising because generous acts may undermine the beneficiaries' incentives for self-improvement (Buchanan, 1975; Lindbeck and Weibull, 1988; Bruce and Waldman, 1990).

## 2. Altruism, reciprocity and cooperation

Consider a community of individuals indexed by $i = 1, \ldots, n$ who may contribute to a public project by supplying an amount of effort $e_i \in [0, 1]$. The total contributions, $\sum_k e_k$, result in a benefit of $q\sum_k e_k$ which is shared equally among individuals in the community, while each individual experiences the quadratic cost of contribution; i.e., $(1/2)e_i^2$. Letting $\phi := q/n$ be $i$'s material payoff from the project net of the cost of contribution is

$$\pi_i = \phi \sum_k e_k - \frac{1}{2} e_i^2.$$

(The results that follow do not depend on this particular functional form or on the quadratic cost function to follow; but use of general functions considerably complicates the presentation and obscures the underlying mechanisms at work.) We note that the marginal private benefit of contribution is $\phi$ and we suppose that $1/n < \phi < 1$; $1/n < \phi$ ensures that full contribution, $e_i = 1$, is socially optimal whereas $\phi < 1$ means that in the absence of punishment selfish individuals, namely those who vary $e$ to maximize $\pi$, under-contribute to the public project ($e_i = \phi < 1$).

After contributions have been observed, each individual $i$ can impose a cost on $j \neq i$ with monetary equivalent $s_{ij}$ at cost $c_{ij}(s_{ij})$ to himself. The cost $s_{ij}$ results from public criticism, shunning, ostracism, physical violence, exclusion from desirable side-deals, or another form of harm. We assume $c_{ij}$ is increasing and strictly convex, and $c_{ij}(0) = c'_{ij}(0) = 0$. Hence $s_i = \sum_{j \neq i} s_{ji}$ is the punishment inflicted upon $i$ by other community members and $c_i = \sum_{j \neq i} c_{ij}(s_{ij})$ is $i$'s cost of punishing others.

Individual $j$'s standing as a cooperative member of community, $b_j$, depends on $j$'s level of effort and the contribution that $j$ makes to the group, which we assume is public knowledge. Specifically, we assume

$$b_j = 2e_j - 1$$

so $b_j = -1$ if $j$ contributes nothing, and $b_j = 1$ if $j$ contributes fully. This means that $e_j = 1/2$ is the point at which $i$ evaluates $j$'s cooperative behavior as neither good nor bad (the 1/2 reflects what we suppose to be an exogenous social norm; the results are unaffected by the any norm specifying a positive level of effort.).

To model cooperative behavior with social preferences, we say that individual $i$'s utility depends on his own material payoff $\pi_i$, the punishment inflicted on $i$ by others, the cost of punishing others, the payoff $\pi_j$ to other individuals $j \neq i$ net of the punishment they receive weighted by $i$'s valuation of the payoffs received by others, which depends both on $i$'s altruism and if $i$ is reciprocal) the others' level of contribution to the public good. Thus

$$u_i = \pi_i - s_i - c_i + \frac{1}{n-1} \sum_{j \neq i} (a_i + \lambda_i b_j)(\pi_j - s_{ij}), \tag{1}$$

where the parameter $a_i$, $-1 < a_i < 1$, is $i$'s level of unconditional altruism if $a_i > 0$ and unconditional spite if $a_i < 0$ and $\lambda_i$, $0 \leq \lambda_i \leq 1$, is the strength of $i$'s reciprocity motive, valuing $j$'s payoffs more highly if $j$ conforms to $i$'s concept of good behavior, and conversely. (The function is similar to Levine (1998), but $i$'s evaluation of $j$'s type is here based on $j$'s actions in a particular game, rather than on $j$'s level of altruism.) The valuation of others' payoffs is weighted by the inverse of the number of other members so that changes in group size do not alter the importance of an individual's own payoffs relative to the payoffs of others.

The final term on the right hand side of (1) captures $i$'s motivation to reduce $j$'s payoffs by inflicting punishment on $j$, in case that $j$ has contributed so little that $i$'s (negative) reciprocity towards $j$ outweighs $i$'s unconditional altruism. We abstract from a second way that $i$'s actions may reduce $j$'s payoffs: if $j$'s reciprocity motive is strong enough, then by contributing less, $i$ can induce $j$ to punish $i$ and hence to incur a cost and to reduce $j$'s payoffs. We regard this motivation as cognitively implausible for it requires that to reduce your payoffs I induce you to punish me when I could have reduced your payoffs directly by punishing you. Moreover, our results hold (Proposition 1) when we reformulate (1) to take it into account (by subtracting $c_j$ from $j$'s payoffs in the final term in (1)). In any case, the size of the associated effect on effort levels is proportional to $1/n^2$ and hence is not substantial except for a very small group.

The cost to $i$ of punishing $j$, $c_{ij}$ is increasing in the level of punishment inflicted. We abstract from the fact that the cost of punishing may also increase with $i$'s level of altruism due to the discomfort that altruists may experience in punishing fellow group members. Taking account of this effect of altruism would contribute a further reason why altruism might be bad for cooperation. For simplicity, we adopt the following specific functional form:

$$c_{ij}\left(s_{ij}\right) = \frac{1}{2}\left(s_{ij}\right)^2. \tag{2}$$

Observe from (1) that consistent with our assumption of "retribution punishment", an individual punishing low contributors values the punishment *per se* rather than the benefits likely to accrue to the punisher or to others if the shirker responds positively to the punishment. Note that when punishment of defectors sustains high enough levels of cooperation to offset the costs of punishment and hence to increase group average payoffs, the reciprocity-based punishment of low contributors is a form of altruism as defined by biologists: altruistic actions increase average payoffs in the group but actors would enhance their individual payoffs were they to forgo punishing low contributors(Kerr et al., 2004). To avoid semantic confusion, we restrict the term altruism to its unconditional variant.

## 3. Altruism versus cooperation?

We model a two-stage optimization process in which individual $i$ selects an effort level taking account of the effect of this choice on the punishment inflicted on $i$ by other team members. To illustrate the effect of a general increase in the altruism of all group members clearly, in this section we suppose that individuals in the community are homogenous; i.e., $(a_i, \lambda_i) = (a, \lambda)$ for all $i$. To find the punishment inflicted on $i$, we first determine $j$'s decision concerning the punishment of $i$ depending on $i$'s contribution level:

$$s_{ji}^*(e_i) = \underset{s_{ji}}{\arg\max}\, u_j(e_j, e_{-j}, s_{ji}, s_{j,-i}) \text{ for all } j \neq i, \tag{3}$$

where $e_{-j} = (e_1, \ldots, e_{j-1}, e_{j+1}, \ldots, e_n)$ and $s_{j,-i} = (s_{j1}, \ldots, s_{j,i-1}, s_{j,i+1}, \ldots, s_{jn})$. Member $j$'s choice of $s_{ji}^*$ in (3) gives the first order condition for an interior solution as follows:

$$c'(s_{ji}^*) = s_{ji}^* = \frac{1}{n-1}(\lambda(1-2e_i) - a).$$

This means the marginal cost of punishing is equal to the marginal benefit of reducing $i$'s payoffs given $j$'s assessment of $i$'s type, net of the subjective costs of inflicting this punishment on $i$ given $j$'s level of unconditional altruism. When $\lambda > 0$ and

$$e_i \geq e_0 := \frac{1}{2\lambda}(\lambda - a), \tag{4}$$

member $j$ does not punish. However, if $\lambda > 0$ and $e_i < e_0$,

$$s_{ji}^* > 0 \text{ and } \frac{\partial s_{ji}^*}{\partial e_i} = -\frac{2\lambda}{n-1}. \tag{5}$$

Note from (4) that the level of contribution that $i$ must make to avoid punishment – which we term the no punishment threshold – is declining in the level of altruism, and that there exists a level of altruism ($a = \lambda$) such that even those who contribute nothing will not be punished. Correspondingly, a sufficient level of spite (or negative altruism, namely, $a = -\lambda$) will require full contribution ($e = 1$) to avoid punishment.

Next individual $i$ decides the level of effort by taking account of the effect of his effort choice on the level of punishment he will receive. Thus member $i$ will choose

$$e_i(e_{-i}, a) = \underset{e_i}{\arg\max}\, u_i(e_i, e_{-i}, s_{ij}^*, s_{i,-j}^*). \tag{6}$$

Eq. (6) defines member $i$'s best effort response to other's effort levels, $e_i = e_i(e_{-i}, a)$. To find $i$'s best response explicitly we proceed as follows. When there is no punishment of $i$, an interior solution $e_{N,i} = e_{N,i}(e_{-i}, a)$ for (6) satisfies the following first order condition (recall $b_j = 2e_j - 1$):

$$e_{N,i} = \phi + \frac{1}{n-1}\sum_{j \neq i}(a + \lambda(2e_j - 1))\phi. \tag{7}$$

Thus when no punishment is inflicted, $i$'s optimal choice of $e_i$ equates the marginal cost of contribution (the left hand side of (7)) to the direct benefits to $i$ of contributing to the project, $\phi$, plus $i$'s valuation on others' material payoffs. Since individuals are identical, from (7) we can find the Nash equilibrium level of contribution $e_N^*$ in the absence of punishment (see Appendix A for the expression) and the effect of increases in altruism on $e_N^*$ :

$$\frac{de_N^*}{da} = \frac{\phi}{1 - 2\lambda\phi}. \tag{8}$$

Similarly when $i$ is subject to punishment (hence $e_i < e_0$), the first order condition for an interior solution $e_{P,i} = e_{P,i}(e_{-i}, a)$ becomes:

$$e_{P,i} = \phi + \frac{1}{n-1}\sum_{j \neq i}(a + \lambda(2e_j - 1))\phi + 2\lambda. \tag{9}$$

Eq. (9) is identical to the no-punishment first order condition (7) except that in addition to the marginal costs and benefits of the project, $i$ must now take account of the effect of increased contribution in reducing punishment ($2\lambda$). Similarly, we

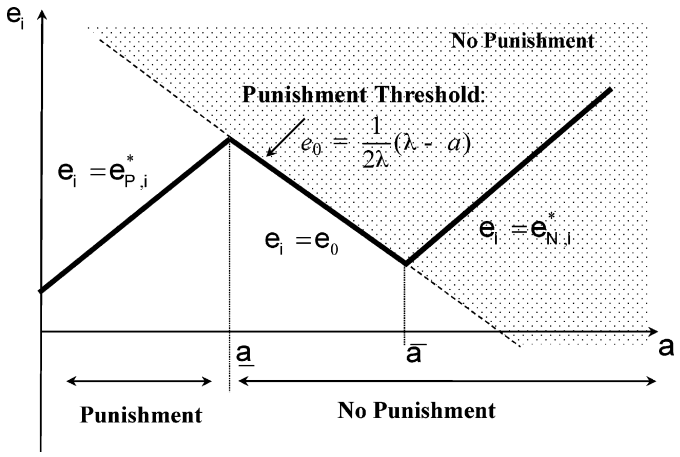


**Fig. 1.** Equilibrium contributions as a function of group members' altruism. For levels of altruism less than $\underline{a}$ contribution levels are low enough to provoke punishment; between $\underline{a}$ and $\overline{a}$ equilibrium contributions are just sufficient to deter punishment, and this critical level of contribution falls as $a$ increases; when altruism exceeds $\overline{a}$ individual's positive valuation of the payoffs of other members induces contribution levels in excess of the punishment threshold. In the shaded region, no punishment arises.

can find the Nash equilibrium level of contribution $e_P^*$ when punishment occurs. Since $\lambda > 0$, we see that $e_P^*(a) > e_N^*(a)$; punishment supports a higher contribution level. The amount contributed by $i$ will depend on whether punishment is present or not, and this will depend on the level of unconditional altruism of the members of the group. We show in Proposition 1 that there exist critical values, $\overline{a}$ and $\underline{a}$, such that for levels of altruism in the interval between them the best response for member $i$ is to contribute just enough to avoid punishment, and this punishment-avoiding level of contribution is declining in the level of altruism. That is:

$$e_i = \begin{cases} e_P^*(a) & \text{if } a < \underline{a} \\ \dfrac{1}{2\lambda}(\lambda - a) & \text{if } \underline{a} < a < \overline{a} \\ e_N^*(a) & \text{if } \overline{a} < a \end{cases} . \tag{10}$$

Fig. 1 illustrates Eq. (10).

If altruism is greater than $\overline{a}$, the expected positive effect of altruism occurs because altruism enhances the members' valuation of the external benefits that their contribution allow, while the offsetting effect (the reduced punishment avoidance motive) does not exist because contribution levels are high enough so that punishment does not occur. In the intermediate range of altruism, Eq. (4) is binding – members contribute just enough to avoid being punished – so an increase in altruism *decreases* the equilibrium effort level since altruism lowers the threshold level of effort required to escape punishment.

Does the 'altruism bad for cooperation' range occur for plausible parameter values? Recall when members have reciprocal motives ($\lambda > 0$), members may punish low contributors and punishment will induce higher effort levels. So we infer that $\lambda > 0$ is necessary for the existence of an interior equilibrium with positive punishment. And if the reciprocity motive $\lambda$ is sufficiently strong among community members and the marginal benefit of the public project $\phi$ is sufficiently great, specifically $2\lambda\phi > 1$, then at a symmetric equilibrium with no punishment, the marginal benefit of contribution is always greater than the marginal cost of contribution and thus every member fully contributes to the project. This is because for $e_j = e$ for all $j$ and $2\lambda\phi > 1$ from Eq. (9) we have:

$$\phi + \frac{1}{n-1}\sum_{j \neq i}(a + \lambda(2e_j - 1))\phi > \phi(1 - \lambda + a) + 2\lambda\phi e > e,$$

so all members contribute fully. When we exclude these uninteresting cases – in which punishment never occurs or in which full contribution is always the result – we obtain the following proposition.

**Proposition 1.** *We suppose that* $0 < \lambda$, $2\lambda\phi < 1$. *There exists* $\underline{a}$ *and* $\overline{a}$ *such that*

$$\frac{de^*}{da} < 0 \text{ for } \underline{a} < a < \overline{a}$$

*where* $e^*$ *is a Nash equilibrium. Furthermore we have*

$$\frac{d}{d\lambda}(\overline{a} - \underline{a}) > 0$$

**Proof.** See Appendix A. □

The intuition behind the first part is that because altruism diminishes the incentive to punish free riders it also reduces the level of contribution necessary to avoid punishment. The second part of the proposition – that the range over which altruism has a negative effect is increasing in the degree of reciprocity – occurs because over the critical range, an increase in reciprocity reduces the negative effect of altruism on contributions (the "no punishment threshold" in Fig. 1 is flatter) while also increasing the difference in contributions between the best responses with and without punishment (the vertical distance between the $e_p^*$ and $e_N^*$). (When $\underline{a} < 0$, contributions are declining not only over the range of altruism but also over some range of reductions in negative altruism, i.e. spite.) Note that while increases in altruism for values of $a$ below $\underline{a}$ and above $\overline{a}$ increase the benefits of the public project net of contribution costs and punishment costs, the reverse is true in the 'altruism unambiguously bad for cooperation' range. Here punishment costs are zero, but increases in altruism reduce contributions to the public good, thus lowering the net benefits.

## 4. A heterogeneous population

We again consider a population consisting of $n$ individuals, but each can be either an altruistic reciprocator or a selfish individual. So parameters describing an individual type $(a_i, \lambda_i)$ can be either $(a, \lambda)$ (an altruistic reciprocator) or $(0, 0)$ (a selfish individual). We denote the fraction of altruistic reciprocators in the population by $\alpha$, where $\alpha \in \{(k/n) : k = 0, \ldots, n\}$, so there are $\alpha n$ altruistic-reciprocators and $(1-\alpha)n$ selfish individuals. We order the indexing of individuals so that

$$(a_i, \lambda_i) = \begin{cases} (a, \lambda) & \text{if } \ i \le \alpha n \\ (0, 0) & \text{if } \ i > \alpha n \end{cases}.$$

The choice of punishment level at the second stage is given by

$$\text{for} j \le \alpha n \ s_{ji}^*(e_i) = \begin{cases} \dfrac{2\lambda}{n-1}\left(\dfrac{1}{2} - e_i - \dfrac{a}{2\lambda}\right) & \text{if} e_i \le \dfrac{1}{2} - \dfrac{a}{2\lambda} \\ 0 & \text{otherwise} \end{cases}, \ \text{for} j > \alpha n \ s_{ji}^*(e_i) = 0,$$

so an altruistic-reciprocators ($j \le \alpha n$) punishes an individual $i$ when the effort of $i$ is low, while a selfish individual never engages in punishment. We retain the quadratic function for the cost of punishment to simplify the exposition; i.e., $c(e_i) = (1/2)(e_i)^2$.

When an altruistic reciprocator $i$ is subject to punishment, the first order condition for an interior solution $e_i^R$ is given by

$$e_i^R = \phi + \frac{1}{n-1}\sum_{j \ne i}^{\alpha n}(a + \lambda(2e_j^R - 1))\phi + \frac{1}{n-1}\sum_{j=\alpha n+1}^{n}(a + \lambda(2e_j^S - 1))\phi + \frac{\alpha n - 1}{n-1}2\lambda. \tag{11}$$

Thus when an altruistic reciprocator $i$ considers a marginal increase in contribution, $i$ will equate the marginal costs (the left hand side of (11)) to (respectively) the marginal private returns from the project ($\phi$), the marginal social returns enjoyed by the fellow reciprocal altruists $(1/(n-1)\sum_{j \ne i}^{\alpha n}(a + \lambda(2e_j - 1))\phi)$, the marginal social returns enjoyed by selfish members of the group $(1/(n-1)\sum_{j=\alpha n+1}^{n}(a + \lambda(2e_j - 1))\phi)$, which may be valued negatively if the reciprocator is sufficiently reciprocal and the selfish individuals contribute sufficiently little effort, and the marginal reduction in punishment $((\alpha n - 1)/(n-1)2\lambda)$. We note that when $\alpha = 1$ (a homogeneous population of altruistic reciprocators) (11) reproduces (9). The expression for the case of no punishment can be found by simply dropping the last term in the right hand side of (11). By contrast, for the case of a selfish individual $i$, the first order condition for an interior solution $e_i^s$ under punishment simply equates the marginal cost of contributing to the marginal private benefits from the project and from the lesser level of punishment associated with a marginal increase in contribution or

$$e_i^S = \phi + \frac{\alpha n}{n-1}2\lambda.$$

Note again that when $\alpha = 0$, the contribution by a selfish individual $i$ is solely determined by the material benefit of the project, $\phi$. In this way, the current heterogeneous population model generalizes the one in Section 3 as well as incorporates as a special case the classical public goods game where no social preferences and punishment are considered.

To find an equilibrium contribution by each type, we use the same method as in Section 3. The results are illustrated in Fig. 2 and stated in Proposition 2.

**Proposition 2.** *Suppose that* $1/4 < \alpha$, $\phi < 1/2$, *and* $1 - 2\phi < \lambda < 3/4$. *Then for the sufficiently large n, there exists* $\underline{a}_R$, $\underline{a}_S$, *and* $\overline{a}$ *such that equilibrium effort is given by Fig. 2. Furthermore, when* $e_R(\underline{a}_R)$ *lies in the interior of the unit interval, the interval* $[\underline{a}_R, \overline{a}]$ *becomes larger as* $\alpha$ *increases.*

**Proof.** See Appendix A. □

Proposition 2 confirms that our intuition obtained from the analysis of Section 3 remains valid for the heterogeneous population. Note that when $\alpha = 1$ Proposition 1 in Section 3 holds, and so Proposition 2 asserts that as long as there are some fraction of altruistic reciprocators in the population ($1/4 < \alpha$), an increase in altruism among altruistic reciprocators

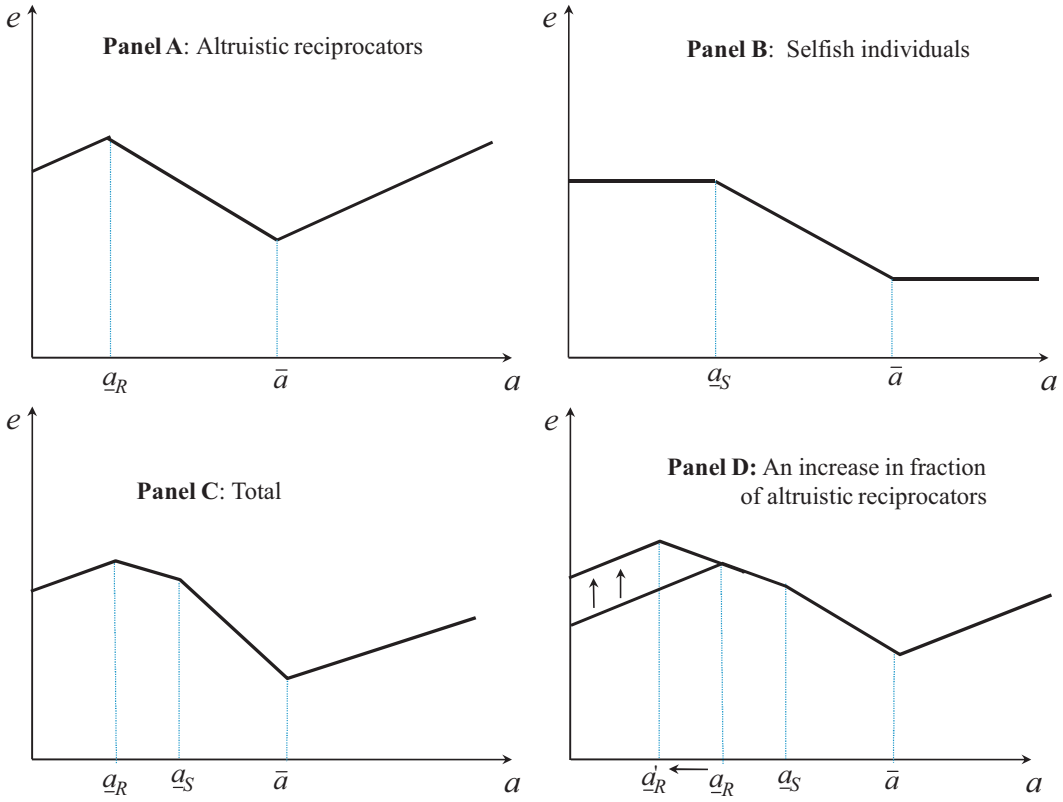


**Fig. 2.** Equilibrium contributions in a heterogeneous population and an increase in the fraction of altruistic reciprocators. Panels A, B show the levels of equilibrium contribution of two types and Panels C and D show the total level of contribution. When the level of altruism $a$ lies between $\underline{a}_R$ and $\bar{a}$, altruistic reciprocators' weights on others' payoffs, $a_i + \lambda_i b_k$, become zero since both types contribute the punishment threshold amount $e_0(a)$ as a result. Therefore when the level of altruism is such that the threshold amount is the same as the selfish individuals' contribution level in the absence of punishment, namely $\phi$ ($\bar{a}$ in Panel B), altruistic reciprocators contribute the same amount as selfish individuals do. For this reason, the critical value of $a$ from which altruistic reciprocators' contribution increases ($\bar{a}$ in Panel A) coincides with the point $\bar{a}$ in Panel B. In Panel D an increase in the fraction of altruistic reciprocators widens the range over which more altruistic preferences imply lower contributions.

will be counter-productive at some range of $a$. The "altruism bad for cooperation" range of values of $a$ increases when the population share of reciprocal altruists increase for the following reason. An increase in the fraction of the population who are reciprocal altruists has no effect on the upper limit of the range for which increased altruism reduces effort, $\bar{a}_S$, namely, the level of altruism above which both selfish and reciprocal agents provide more than enough effort to avoid punishment. But it reduces $\underline{a}_R$, the least level of altruism for which this altruism-bad effect holds. The reason why this is so is illustrated in Panel D of Fig. 2.

For any given level of altruism an increase in the fraction of altruistic reciprocators raises the level of effort that reciprocal agents provide when they are subject to punishment (the upward shift in the $e^*_{P,i}$ function in Panel D of Fig. 2). This reduces the level of altruism for which the no-punishment threshold is a binding and therefore expands the range of levels of altruism over which the level of contribution required to avoid punishment is declining in the level of altruism.

## 5. Caveats and extensions

We do not explore the conceptually challenging effect of an increase in altruism on subjective welfare, given that the change in altruism is itself a change in preferences (Bergstrom, 2006) analogous to a free resource allowing costless increases in subjective well being. Nor do we address the hypothesis that if incentive mechanisms other than peer punishment were allowed, a general increase in altruism might not be bad for cooperation. If the set of alternative mechanisms is unrestricted, the hypothesis is trivially the case: an appropriate subsidy for contributions in a complete information setting would implement the social optimum, making punishment redundant. We can think of no non-arbitrary way to expand the set of alternative mechanisms while retaining the underlying problem of public goods provision.

Our representation of the motive for punishment – hostility toward those who violate cooperative norms – could be expanded so that the extent of hostility is enhanced by feelings of altruism towards those that the defector has harmed. In this case a general increase in altruism would (as in the current model) make individuals more reluctant to harm defectors, but it would also increase hostility toward defectors, thereby possibly offsetting the first effect. Finally, we could have assumed a more sophisticated motive, one in which punishment was instrumental with behavior modification of the free riders as the objective. In this case, increased altruism might (but need not) enhance punishment and contributions. The

reason is that in this (we think empirically implausible) 'strategic punishing' model, the prospective punisher takes account of the other members' prospective gains resulting from the reduced costs of punishment that they will bear given the target's expected positive contribution response to the punishment. For sufficient levels of altruism these gains might outweigh the negative effect of altruism on the non-strategic punishment motive.

An interesting extension of our treatment of heterogeneous populations would de-link altruism and reciprocity so that there could be two new behavioral types in addition to the altruistic reciprocators and the entirely self interested agents studied here, namely altruistic non reciprocators ($a_i > 0$, $\lambda_i = 0$ in Eq. (1)) and reciprocating non altruists ($a_i = 0$, $\lambda_i > 0$). Then, if the traits "altruistic" and "reciprocal" are limited in supply, voluntary public goods provision may be greater if individuals have one or the other trait but not both, and that under these conditions an increase in the degree of altruism has unambiguously positive effects. The moral of the story would be that to foster a cooperative culture, the young should be raised either on the Old Testament or the New Testament, but not on both!

The intuition is that if altruists are never also reciprocal, then altruism cannot attenuate the motive to punish free riders. This is just a polar case illustrating the second part of Proposition 1, namely that the range over which increased altruism reduces contributions is greater, the more reciprocal are the citizens (and disappears when $\lambda = 0$). Analysis of the many possible equilibria for this problem is quite complex and depends critically on the extent of public and private information and the availability of a common culture or other coordinating mechanisms. But we have not studied this case in detail.

A further extension would be to consider the effects of individuals rewarding high contributing group-mates rather than punishing free riders. But the results of this exercise are less than startling: an increase in altruism could not reduce contributions, illustrating a common finding common in the behavioral economics literature, namely that rewards are not simply punishments with a sign reversal.

Finally, our result that altruism may be bad for cooperation because it deters individuals from the punishment of free riders has an evolutionary analogue, that can be expressed in the following way. A population has three types: unconditional altruists who contribute to a public good, free riders who do not, and reciprocators who contribute and punish free riders. In a standard replicator dynamic model of cultural evolution, the population may sustain high levels of cooperation when reciprocators are prevalent. But it can be invaded by unconditional altruists, who having replaced the reciprocators, are then replaced by the free riders. The reason is that the when free riders are rare, altruists benefit from the high levels of cooperation without paying the costs of punishing the free riders, and so they replace the reciprocators. The unconditional altruists are effectively parasites on the reciprocators, leading to their mutual elimination. The free riders then exploit the altruists and take over the population. Examples of this dynamic are Bowles (2004) and Bowles and Gintis (2004). In this population the expected level of cooperation in the very long run will be increased if a cost is imposed on unconditional altruists to deter their free riding on the punishment activities of the reciprocators.

Though the main result is the same – altruism is bad for cooperation – the evolutionary process differs from the one we have modeled here, in which individual behavior varies with the degree of altruism in the utility function. In the evolutionary model, individual behavior is fixed, and utility maximization is replaced by a payoff monotonic replicator dynamic based on differences in 'cultural fitness'. The casual mechanism is also different: in the evolutionary model, altruism has its deleterious effects because it reduces the number of reciprocators, leading to a collapse of cooperation, while in the model presented here, in the altruism-bad-for-cooperation range an increase in individual altruism reduces the level of effort required to avoid punishment, while an increase in the reciprocity parameter of the individual's utility function exacerbates the problem by expanding the range over which altruism is bad for cooperation.

## 6. Conclusion

Some philosophers and social scientists have stressed the importance for good government of an altruistic citizenry that values the well-being of fellow citizens (Mill, 1861; Rawls, 1971; Schumpeter, 1950; Almond and Verba, 1963). Economists, however, have emphasized the need for incentives that induce even the self-interested to contribute to the public good. Implicitly most have assumed that these two approaches are complementary or at worst additive. It is now recognized that this assumption may fail where the presence of monetary or other explicit incentives reduces the salience of altruistic or other public-spirited motives (Benabou and Tirole, 2003, 2006; Bowles, 2008; Falk and Kosfeld, 2006; Sliwka, 2007; Bowles and Hwang, 2008; Bowles and Polania Reyes, 2012). But as we have seen, the assumption that the effects of incentives and social preferences are at worst additive need not hold even in the absence of such motivational crowding out (namely, where preferences are exogenous, as in the model presented here).

Our results suggest that for a community wishing to sustain high levels of cooperation, efforts to enhance unconditional altruism may be counter-productive, and that enhancing the level of citizen reciprocity may exacerbate the negative effects of altruism. Other social preferences, however, may be synergistic with reciprocity. For example, inequality aversion as proposed by Fehr and Schmidt (1999) can enhance the motivation to punish those who make relatively high payoffs by free riding on the cooperation of others. This mechanism differs from ours in that the punisher is seeking to rectify an unfair outcome, while in our model the reciprocator wishes to punish the violation of a social norm.

But punishment is often (as in our model) resource-using; costs are imposed both on the target and the punisher. Unless or until levels of contribution sufficient to make punishment rare are achieved, the costs associated with punishment of low contributors may more than offset the gains to cooperation that the punishment allows (Herrmann et al., 2008; Gaechter

et al., 2008). This is particularly true in a case we have not considered, namely when vendetta-like cycles of punishment and counter punishment occur (Hopfensitz and Reuben, 2009; Nikiforakis and Engelmann, 2011).

Nonetheless, cooperation sustained by a combination of altruism and reciprocity-based punishment may be welfare enhancing. This is true in part because punishment is not only an incentive; it is also a signal. The incentive-based response to punishment may be enhanced by the feelings of shame that punishment by peers triggers (Bowles and Gintis, 2005). In part for this reason, disapproval by peers may induce members to contribute even when it is expressed in non-resource-using ways such as gossip, ridicule or the simple statement that the individual has violated a norm (Masclet et al., 2003; Barr, 2001; Wiessner, 2005).

## Appendix A. Proofs of Proposition 1 and 2

### A.1. Proof of Proposition 1

We first find the interior equilibrium candidates $e_P^*(a)$ and $e_N^*(a)$ :

$$e_N^*(a) = \frac{\phi}{1-2\lambda\phi}a + \frac{1-\lambda}{1-2\lambda\phi}\phi$$
$$e_P^*(a) = \frac{\phi}{1-2\lambda\phi}a + \frac{2\lambda+(1-\lambda)\phi}{1-2\lambda\phi}$$

and denote by $e_0(a)$ the threshold effort level for punishment:

$$e_0(a) := \frac{1}{2\lambda}(\lambda - a).$$

Note that we have $e_0(\lambda)=0$ and $e_0(-\lambda)=1$. Then from $2\lambda\phi<1$ we have

$$e_N^*(\lambda) = \frac{\phi}{1-2\lambda\phi} > 0 \text{ and } e_N^*(-\lambda) = \frac{1-2\lambda}{1-2\lambda\phi}\phi < 1.$$

Thus we choose $\overline{a}$ such that $e_N^*(\overline{a}) = e_0(\overline{a})$ and this gives $\overline{a} = \lambda(1-2\phi)$. Next if $e_P^*(-\lambda) < 1$, we can choose $\underline{a}$ such that $e_P^*(\underline{a}) = e_0(\underline{a})$ and this yields $\underline{a} = \lambda(1-2\phi-4\lambda)$. Otherwise we set $\underline{a} := -\lambda$. Note that when $2\lambda+\phi>1$, $e_P^*(-\lambda)>1$. Concerning the second part of the claim, using the expressions we have either

$$\overline{a} - \underline{a} = 4\lambda^2 \text{ or } = 2\lambda(1-\phi).$$

Hence the result follows.

### A.2. Heterogeneous population

#### A.2.1. Equilibrium contributions

Since only the altruistic reciprocator type punishes, depending on which type is punished we have four possibilities of punishment pattern at equilibrium: (1) no type is punished (**NP**), (2) only the selfish type is punished (**PS**), (3) only the altruistic reciprocator type is punished (**NSPR**), (4) both types are punished (**PA**). For each case, we find the interior equilibrium candidate of the effort level as follows:

**NP** :
$$e_i^S = \phi \quad \text{for } i > \alpha n$$
$$e_i^R = \frac{1}{1-(n\alpha-1)/(n-1)2\lambda\phi}\left(\phi + \phi(a-\lambda) + \frac{(1-\alpha)n}{n-1}2\lambda\phi^2\right) \quad \text{for } i \le \alpha n$$

**NSPR** :
$$e_i^S = \phi \quad \text{for } i > \alpha n$$
$$e_i^R = \frac{1}{1-(n\alpha-1)/(n-1)2\lambda\phi}\left(\phi + \phi(a-\lambda) + \frac{(1-\alpha)n}{n-1}2\lambda\phi^2 + \frac{\alpha n-1}{n-1}2\lambda\right) \quad \text{for } i \le \alpha n$$

**PS** :
$$e_i^S = \phi + \frac{\alpha n}{n-1}2\lambda \quad \text{for } i > \alpha n$$
$$e_i^R = \frac{1}{1-(n\alpha-1)/(n-1)2\lambda\phi}\left(\phi + \phi(a-\lambda) + \frac{(1-\alpha)n}{n-1}2\lambda\phi^2 + \frac{\alpha(1-\alpha)n^2}{(n-1)^2}(2\lambda)^2\phi\right) \quad \text{for } i \le \alpha n$$

**PA** :
$$e_i^S = \phi + \frac{\alpha n}{n-1}2\lambda \quad \text{for } i > \alpha n$$
$$e_i^R = \frac{1}{1-(n\alpha-1)/(n-1)2\lambda\phi}\left(\phi + \phi(a-\lambda) + \frac{(1-\alpha)n}{n-1}2\lambda\phi^2 + \frac{\alpha(1-\alpha)n^2}{(n-1)^2}(2\lambda)^2\phi + \frac{\alpha n-1}{n-1}2\lambda\right) \quad \text{for } i \le \alpha n.$$

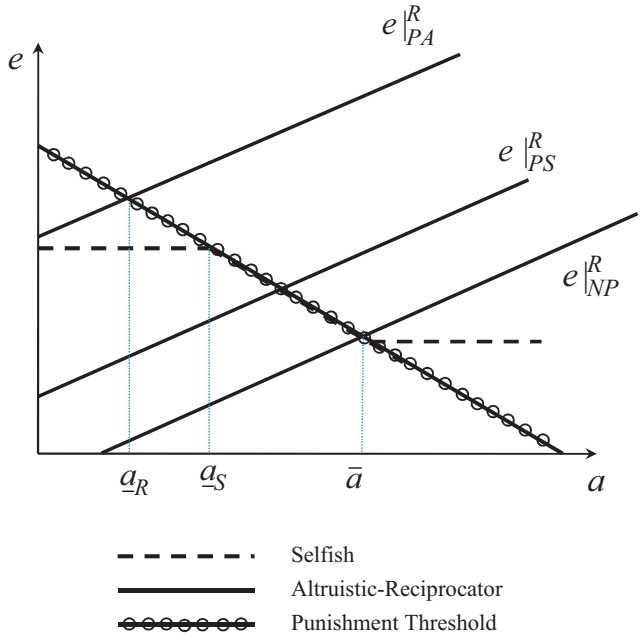


**Fig. A3.** Possible values of $\underline{a}_S, \bar{a}_S, \bar{a}_R, \underline{a}_R$.

Note that from the hypothesis the following stability condition is satisfied:

$$\frac{n\alpha-1}{n-1}2\lambda\phi < 1. \tag{A.1}$$

Also we have

$$\begin{aligned}(1-\frac{n\alpha-1}{n-1}2\lambda\phi)(e_i^R|_{\mathbf{NSPR}}-e_i^S|_{\mathbf{NSPR}}) &= \phi(a-\lambda)+\frac{n\alpha-1}{n-1}2\lambda\phi^2+\frac{(1-\alpha)n}{n-1}2\lambda\phi^2+\frac{n\alpha-1}{n-1}2\lambda\\ &\geq \phi a-\phi\lambda+2\lambda\phi^2+\frac{n\alpha-1}{n-1}2\lambda\\ &\geq \phi a+2\lambda\phi^2+(2\frac{n\alpha-1}{n-1}-\frac{1}{2})\lambda>0.\end{aligned}$$

where in the third line we choose $n$ large such that $(n\alpha-1)/(n-1)>(1/4)$ holds from $\alpha>(1/4)$. Thus we have $e_i^R|_{\mathbf{NSPR}} > e_i^S|_{\mathbf{NSPR}}$, so the **NSPR** case does not occur at equilibrium. We note that for altruistic reciprocators, $e_i^R|_{\mathbf{NP}} < e_i^R|_{\mathbf{PS}} < e_i^R|_{\mathbf{PA}}$. Next we find the condition under which the selfish type is not punished. This case occurs if a selfish individual's contribution is greater than the threshold level of effort:

$$e_i^S|_{\mathbf{NP}}(a) > \frac{1}{2\lambda}(\lambda-a) \Leftrightarrow a>\lambda(1-2\phi) := \bar{a}.$$

On the other hand when $a$ is sufficiently low, an increase in $a$ does not affect the effort level of a selfish individual. This happens if

$$e_i^S|_{\mathbf{PS}}(a)=e_i^S|_{\mathbf{PA}}(a)<\frac{1}{2\lambda}(\lambda-a)\Leftrightarrow a<\lambda\left(1-2\phi-\frac{\alpha n}{n-1}4\lambda\right) := \underline{a}_S.$$

So we have

$$e_*^S(a)=\begin{cases}\phi+\dfrac{\alpha n}{n-1}2\lambda & \text{if } a<\underline{a}_S\\ \dfrac{1}{2\lambda}(\lambda-a) & \text{if } \underline{a}_S<a<\bar{a}\\ \phi & \text{if } \bar{a}<a\end{cases}. \tag{A.2}$$

Notice from the discussion of the text we have

$$e_i^R|_{\mathbf{NP}}(\bar{a})=\frac{1}{2\lambda}(\lambda-\bar{a})=\phi$$

Fig. A3 shows possible values of $\underline{a}_S$, $\bar{a}$, and $\underline{a}_R$.

Recall that in the **PS** case only the selfish type is punished, so the line $e^R|_{\mathbf{PS}}$ is only meaningful when $a \in (\underline{a}_S, \bar{a})$. In Fig. A3 the intersection between $e^R|_{\mathbf{PS}}$ and the punishment threshold $((1/2\lambda)(\lambda - a))$ is located at the right side of $\underline{a}_S$, so $e^R|_{\mathbf{PS}} < e^S|_{\mathbf{PS}}$ for all $a < \underline{a}_S$ and this, in turn, implies that the altruistic reciprocator type, as well as the selfish type, is subject to punishment. However this is impossible by our definition of the **PS** case. Therefore the situation where the altruistic-reciprocator type chooses $e^R|_{\mathbf{PS}}(a)$ does not occur at equilibrium; i.e., the line $e^R|_{\mathbf{PS}}(a)$ is non-binding at equilibrium. Note that the equilibrium profile in Fig. A3 provides equilibrium contributions depicted by Fig. 2 in the text. To prove Proposition 2, we first define the effort levels of altruistic reciprocators at equilibrium, $e^R_*(a)$, when the equilibrium profile is given by Fig. A3:

$$e^R_*(a) = \begin{cases} \dfrac{1}{1-(n\alpha-1)/(n-1)2\lambda\phi}(\phi + \dfrac{n\alpha-1}{n-1}\phi(a-\lambda) + \dfrac{(1-\alpha)n}{n-1}2\lambda\phi^2 + \dfrac{\alpha(1-\alpha)n^2}{(n-1)^2}(2\lambda)^2\phi + \dfrac{\alpha n-1}{n-1}2\lambda) & \text{if } \ a < \underline{a}_R \\ \dfrac{1}{2\lambda}(\lambda - a) & \text{if } \underline{a}_R < a < \bar{a} \\ \dfrac{1}{1-(n\alpha-1)/(n-1)2\lambda\phi}(\phi + \dfrac{n\alpha-1}{n-1}\phi(a-\lambda) + \dfrac{(1-\alpha)n}{n-1}2\lambda\phi^2) & \text{if } \underline{a} < a \end{cases} \tag{A.3}$$

*A.2.2. Proof of Proposition 2*

Proof of Proposition 2 To have the equilibrium profile described in Figs. 2 and A3, we need

$$e^S_i|_{\mathbf{PA}}(0) < e^R_i|_{\mathbf{PA}}(0) \tag{A.4}$$

$$e^R_i|_{\mathbf{PS}}(\underline{a}_S) < e^S_i|_{\mathbf{PS}}(\underline{a}_S) \tag{A.5}$$

First, for (A.5) we have

$$e^S_i|_{\mathbf{PS}}(\underline{a}_S) - e^R_i|_{\mathbf{PS}}(\underline{a}_S) = \frac{2n\alpha\lambda}{n(1-2\alpha\lambda\phi) + 2\lambda\phi - 1}$$

From $\lambda\phi < 1/2$, $n(1-2\alpha\lambda\phi) + 2\lambda\phi - 1 > n(1-2\lambda\phi) + 2\lambda\phi - 1 = (n-1)(1-2\lambda\phi) > 0$, so $e^S_i|_{\mathbf{PS}}(\underline{a}_S) - e^R_i|_{\mathbf{PS}}(\underline{a}_S) > 0$. Concerning (A.4) we find

$$e^R_i|_{\mathbf{PA}}(0) - e^S_i|_{\mathbf{PA}}(0) = \frac{\lambda((2\phi^2 - \phi + 4\alpha\lambda\phi)n - 2\phi^2 + \phi - 2)}{n(1-2\alpha\lambda\phi) + 2\lambda\phi - 1}$$

Since $\alpha > 1/4$ and $2\phi - 1 + \lambda > 0$, we have $2\phi^2 - \phi + 4\alpha\lambda\phi > 2\phi^2 - \phi + \lambda\phi > 0$. We choose $\beta := 2\phi^2 - \phi + \phi\lambda > 0$. Also for $\phi < 1/2$, $2\phi^2 - \phi + 2 < 2$. So, we can choose $n$ such that $n\beta > 2$. Then

$$(2\phi^2 - \phi + 4\alpha\lambda\phi)n - 2\phi^2 + \phi - 2 > 0.$$

Note that when $e^R_i|_{\mathbf{PA}}(\underline{a}_R) > 1$ (when the equilibrium at $\underline{a}_R$ occurs at the corner of the unit interval 1), we can simply redefine $\underline{a}_R = -\lambda$ (from $e_0(-\lambda) = 1$) and the result similarly follows. Thus we obtain the first result. Next we show that the interval $[\underline{a}_R, \bar{a}]$ becomes larger as $\alpha$ increases when $e^R_i|_{\mathbf{PA}}(\underline{a}_R)$ lies in the interior of the unit interval. First we note that $\bar{a}$ and $\underline{a}_S$ does not depend on $\alpha$. Thus it is enough to show that $\underline{a}_R$ decreases as $\alpha$ increases; or equivalently, an increase in $\alpha$ shifts up $e|^R_{\mathbf{PA}}$ which is given by

$$e^R|_{\mathbf{PA}} = \underbrace{\frac{1}{1-(n\alpha-1)/(n-1)2\lambda\phi}}_{(i)}\underbrace{(\phi + \phi(a-\lambda) + \frac{(1-\alpha)n}{n-1}2\lambda\phi^2 + \frac{\alpha(1-\alpha)n^2}{(n-1)^2}(2\lambda)^2\phi + \frac{\alpha n-1}{n-1}2\lambda)}_{(ii)}. \tag{A.6}$$

By treating $\alpha$ in (A.6) as a continuous variable $\alpha \in [0, 1]$, first we have

$$\frac{\partial}{\partial\alpha}(i) > 0$$

and by differentiating the terms in the parenthesis in (A.6) with respect to $\alpha$, we find that

$$\begin{aligned} \frac{\partial}{\partial\alpha}(ii) &= -\frac{n}{n-1}2\lambda\phi^2 + \frac{n^2}{(n-1)^2}(1-2\alpha)(2\lambda)^2\phi + \frac{n}{n-1}2\lambda \\ &\geq \frac{2\lambda n}{n-1}(-\phi^2 + (1-2\alpha)(2\lambda)\phi + 1) \\ &\geq \frac{2\lambda n}{n-1}(\frac{3}{4} - 2\lambda\phi) \geq \frac{2\lambda n}{n-1}(\frac{3}{4} - \lambda) \\ &> 0. \end{aligned}$$

Since $(i)$ and $(ii)$ in (A.6) are positive, an increase in $\alpha$ shifts up $e|^R_{PA}$ and the second claim follows. □

## References


Alger, I., Weibull, J., 2010. Kinship, incentives, and evolution. American Economic Review 100 (4), 1725–1758.
Almond, G.A., Verba, S., 1963. The Civil Culture: Political Attitudes and Democracy in Five Nations. Princeton University Press.
Anderson, C., Putterman, L., 2006. Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. Games and Economic Behavior 54, 1–24.
Andreoni, J., 1990. Impure altruism and donations to public goods: a theory of warm glow giving. Economic Journal 100, 464–477.
Barr, A., 2001. Social dilemmas, shame-based sanctions, and shamelessness: experimental results from rural Zimbabwe. Center for the Study of African Economies Working Paper WPS/2001.11. Oxford University.
Benabou, R., Tirole, J., 2003. Intrinsic and extrinsic motivation. Review of Economic Studies 70 (33), 489–520.
Benabou, R., Tirole, J., 2006. Incentives and prosocial behavior. American Economic Review 96 (5), 1652–1678.
Bergstrom, T., 2006. Benefit-cost in a benevolent society. American Economic Review 96 (1), 339–351.
Bernheim, D., Stark, O., 1988. Altruism within the family reconsidered: do nice guys finish last? American Economic Review 78 (5), 1034–1045.
Bowles, S., 2004. Microeconomics: Behavior, Institutions, and Evolution. Princeton University Press.
Bowles, S., 2008. Policies designed for self-interested citizens may undermine "the moral sentiments:" evidence from experiments. Science 320 (5883), 1605–1609.
Bowles, S., Gintis, H., 2004. The evolution of strong reciprocity: cooperation in a heterogeneous population. Theoretical Population Biology 65, 17–28.
Bowles, S., Gintis, H., 2005. Prosocial emotions. In: Blume, L.E., Durlauf, S.N. (Eds.), The Economy as a Complex Evolving System III: Essays in Honor of Kenneth Arrow. Oxford University Press, Oxford, pp. 337–367.
Bowles, S., Hwang, S.-H., 2008. Social preference and public economics: mechanism design when preferences depend on incentives. Journal of Public Economics 92 (8–9), 1811–1820.
Bowles, S., Polania Reyes, S., 2012. Economic incentives and pro-social behavior. Journal of Economic Literature 50, 1–57.
Boyd, R., Gintis, H., Bowles, S., 2010. Coordinated punishment of defectors sustains cooperation and can proliferate when rare. Science 328, 617–620.
Boyd, R., Richerson, P.J., 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. Evolution and Human Behavior 13 (3), 171–195.
Bruce, N., Waldman, M., 1990. The rotten-kid theorem meets the Samaritan's dilemma. Quarterly Journal of Economics 105 (1), 155–165.
Buchanan, J.M., 1975. The Samaritan's dilemma. In: Pelphs, E.S. (Ed.), Altruism, Morality and Economic Theory. Russell Sage Foundation, New York.
Carpenter, J., Bowles, S., Gintis, H., Hwang, S.-H., 2009. Strong reciprocity and team production: theory and evidence. Journal of Economic Behavior and Organization 71 (2), 221–232.
Casari, M., Luini, L., 2012. Peer punishment in teams: expressive or instrumental choice. Experimental Economics 15, 241–259.
de Quervain, D., Fishbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., Fehr, E., 2004. The neural basis of altruistic punishment. Science 305, 1254–1258.
Falk, A., Fehr, E., Fischbacher, U., 2005. Driving forces behind informal sanctions. Econometrica 73, 2017–2030.
Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior 52 (2), 293–315.
Falk, A., Kosfeld, M., 2006. The hidden costs of control. American Economic Review 96 (5), 1611–1630.
Fehr, E., Gaechter, S., 2000. Cooperation and punishment in public goods experiments. American Economic Review 90 (4), 980–994.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Quarterly Journal of Economics 114, 817–868.
Fudenberg, D., Pathak, P., 2010. Unobserved punishment supports cooperation. Journal of Public Economics 94, 78–86.
Gaechter, S., Hermann, B., 2011. The limits of self-governance when cooperator get punished: experimental evidence from urban and rural Russia. European Economic Review 55, 193–210.
Gaechter, S., Renner, E., Sefton, M., 2008. The long-run benefits of punishment. Science 322 (5907), 1510.
Herrmann, B., Thoni, C., Gaechter, S., 2008. Antisocial punishment across societies. Science 319 (7), 1362–1367.
Hopfensitz, A., Reuben, E., 2009. The importance of emotions for the effectiveness of social punishment. Economic Journal 119 (540), 1534–1559.
Kerr, B., Godfrey-Smith, P., Feldman, M., 2004. What is altruism. Trends in Ecology and Evolution 19 (3), 135–140.
Levine, D.K., 1998. Modeling altruism and spitefulness in experiments. Review of Economic Dynamics 1 (3), 593–622.
Lindbeck, A., Weibull, J.W., 1988. Altruism and time consistency: the economics of fait accompli. Journal of Political Economy 96 (6), 1165–1182.
Masclet, D., Noussair, C., Tucker, S., Villeval, M.-C., 2003. Monetary and non-monetary punishment in the voluntary contributions mechanism. American Economic Review 93 (1), 366–380.
Mill, J.S., 1998/1861. Utilitarianism. Oxford University Press.
Nakao, K., 2008. Can altruism hinder cooperation? Economics Bulletin 4 (26), 1–6.
Nikiforakis, N., Engelmann, D., 2011. Altruistic punishment and the threat of feuds. Journal of Economic Behavior and Organization 78, 319–332.
Rabin, M., 1993. Incorporating fairness into game theory and economics. American Economic Review 83 (5), 1281–1302.
Rawls, J., 1971. A Theory of Justice. Harvard University Press.
Schumpeter, J., 1950. The march into socialism. American Economic Review 40, 446–456.
Sliwka, D., 2007. Trust as a signal of a social norm and the hidden costs of incentive schemes. American Economic Review 97 (3), 999–1012.
Smith, A., 1976/1776. An Inquiry into the Nature and Causes of the Wealth of Nations. Clarendon Press.
Wiessner, P., 2005. Norm enforcement among the ju/'hoansi bushmen: a case of strong reciprocity? Human Nature 16 (2), 115–145.