Noname manuscript No.
(will be inserted by the editor)

Punishment and the Potency of Group Selection

Richard Povey

Received: date / Accepted: date

Abstract It is known that altruism can be sustained in an evolving population by a process of group selection. There is also existing research on the role that punishment can play in inducing selfish agents to behave more co-operatively or in preventing selfish agents from evolving, and the limitations upon this mechanism. This paper embeds a simple model of a punishment system within an indirect cultural evolution framework. The use of punishment is shown to reduce the potency of the group selection mechanism, and thus the level of evolved altruism. This presents a novel reason why the use of punishment may have negative dynamic welfare implications.

Keywords altruism · punishment · social preferences · group selection · multilevel selection · indirect evolution · cultural evolution

1 Overview

The central problem of economic and social policy, indeed the essential prerequisite of the social order itself, is that of bestowing upon the individual agent the incentive to act in a manner which is beneficial for society as a whole. Such incentives can be *intrinsic* to the individual (altruistic preferences) or *extrinsic* (threats of punishment). This paper analyses the interaction between these two alternative "technologies". We explore the possibility that they cannot be freely mixed, since the use of extrinsic incentives can undermine the evolution of intrinsic incentives.

Group selection provides a framework for modelling this perverse dynamic effect. Societies evolve through imitation and propagation of the mores of their own most successful members (and those of other societies). This process of cultural evolution offers a functional explanation both for the individual "social preferences" which evolve (i.e. altruistic motivations) and for the institutional frameworks which are

The final publication is available at http://link.springer.com/article/10.1007/s00191-014-0375-3

Brasenose College and Hertford College, Oxford University
E-mail: richard.povey@hertford.ox.ac.uk

developed (i.e. the state and legal system). However, since group selection operates via the improved relative success of groups which adopt altruistic morals, and since the use of punishment systems primarily enhances the performance of groups where such intrinsic altruism is lacking, the use of such extrinsic punishment can weaken the group selection mechanism and hence make it more difficult for altruism to evolve.

Let us propose a few concrete examples of this phenomenon from the world of economics: Individuals arguably pay their taxes for a combination of intrinsic and extrinsic motives. They feel a moral obligation to contribute to the broader society. At the same time, they fear a fine or imprisonment for tax evasion. By making it easier to collect taxes in a society with low intrinsic motivation, use of extrinsic punishment may further weaken intrinsic moral norms regarding the paying of taxes. Other examples could include contract law, environmental protection and antitrust law. State-enforced environmental or competition law may weaken norms of corporate social responsibility which prevent firms from pursuing unrestrained self-interest. State enforcement of contracts might undermine individual honesty.

2 Relevant Literature

It is a well-established result in evolutionary theory that altruism can in principle be sustained by a process of group selection if a population is split into groups whose members interact disproportionately with one another, provided that there is migration between groups. The level of altruism which can be sustained depends upon the relative strength of the evolutionary forces benefiting the more selfish individuals at the expense of altruists within groups, and that favouring the more altruistic groups over the less altruistic ones. There is an evolutionary "tug-of-war" between individual-level and group-level selection [Sober and Wilson (1999)]. Direct inter-group competition (in the starkest case, lethal conflict and war) can further augment this process by providing a significant additional relative performance boost for groups of co-operative altruists in such multilevel selection models, further enhancing the evolution of human altruism [Bowles (2006)] [Bowles (2009)].

The idea of group selection originates with Darwin [Darwin (1872)], but the contemporary formulation was developed in the twentieth century literature on evolutionary biology, most famously in the work of W. D. Hamilton [Hamilton (1963)] [Hamilton (1972)]. The mathematical framework was originally devised by Price [Price (1970)]. Still controversial among some biologists (but more widely accepted as a useful practical theory in social science fields), the multilevel selection paradigm has recently been popularised within and beyond the biological field by Sober and Wilson. They have provided a survey article [Sober and Wilson (1994)] and a book-length treatment of the subject [Sober and Wilson (1999)].

Group selection has also become a popular framework in theoretical anthropology, from which the fruitful suggestion that we may see cultural as well as genetic characteristics as evolving through natural selection has been developed and employed [Boyd and Richerson (1982)] [Soltis et al (1995)] [Blackmore (1999)]. This idea also has a pedigree going back to Darwin [Darwin (1872)], but arguably he was heavily influenced [Hirshleifer (1977)] [Hayek (1988)] by the application

of the same principle to social institutions by the philosophers of the Scottish Enlightenment, most famously Adam Smith [Smith (1976)]. Economists have also made important contributions to the theory of group selection, particularly in clarifying the mathematical analysis of the different types of group structure that can enable this phenomenon to arise [Bergstrom (2002)] [Cooper and Wallace (2004)].

There is also an existing literature on the role that punishment, in the form of informal sanctions or a legal system, can play as an "altruism amplification device". It has been shown that, because punishing others is often less costly than benefiting them, the emergence of the ability to carry out altruistic punishment (secondary altruism) can explain how the evolution of primary altruism is made possible in a much wider variety of cases. This hypothesis fits the empirically-observed phenomenon that the ability to punish transgressors in simple experimental games such as the public goods game results in more co-operation being sustained [Fehr and Gächter (2000b)] [Fehr and Gächter (2000a)] [Fehr and Gächter (2002a)] [Fehr and Fischbacher (2003)] [Fehr and Gächter (2002b)]. There are two dimensions to this impact. Firstly, altruistic punishment improves "static" outcomes by making selfish individuals behave better, because they are afraid of being punished. Secondly, the evolution of altruistic punishment can also make it easier for altruism to evolve as a primary behaviour, by reducing the gain in fitness by selfish individuals relative to the altruists in the group [Sober and Wilson (1999)] [Boyd et al (2003)].

More recently, researchers have explored the limitations upon the potential for punishment to enhance social co-operation from both empirical and theoretical perspectives. It has been found empirically that altruistic punishment only enables co-operation to be sustained in public goods games if the cost to the punished agent is quite high relative to cost to the punisher, and that the deadweight loss from punishment can often outweigh the efficiency gains from greater co-operation [Egas and Riedl (2008)] [Nikiforakis and Normann (2008)]. It has also been found that the possibility of counter-punishment [Denant-Boemont et al (2007)] [Nikiforakis (2008)] or antisocial punishment[1] [Herrmann et al (2008)] [Rand et al (2010)] can play a significant role in undermining co-operation.

This paper aims to make a contribution to the theoretical understanding of the connection between group selection and punishment by applying a third conceptual strand; that of indirect evolution. Most models of the cultural evolution of altruism model cultural norms in a "mechanical" way in the sense that individuals blindly carry out their "programmed" behaviour, whereas economic theory seeks to explain phenomena from a wide variety of cultural scenarios as caused by the same underlying human rationality. The alternative is to assume that it is the weightings that individuals place on the welfare of others that form the evolving phenotype, rather than specific altruistic behaviours directly. In other words, *preferences* evolve but behaviour within the games being played is rational and forward looking, and therefore modelled using standard game theoretic concepts.

[1] Antisocial punishment involves sanctions such as altruistic punishment being directed against co-operators instead of non-co-operators. (On one explanatory hypothesis, it may be carried out by non-co-operators retaliating against observed previous punishment which they believe to have been carried out by co-operators.) The prevalence of this phenomenon is socially dependent and negatively correlated to the depth of the rule of law and civic society.

The model presented will embed a simple punishment system within an indirect cultural evolution framework. The indirect evolution approach was first proposed by Güth and Yaari [Güth and Yaari (1992)] following a suggestion originally made by Becker [Becker (1976)]. It has been profitably applied to explaining the evolution of preferences for fairness in the ultimatum game [Huck and Oechssler (1999)]. In this context, it has been shown that "vengeful" individuals, who gain utility from reducing the payoffs of others at cost to themselves, can propagate.

In the standard direct evolution group selection models, there is no distinction between altruistic preferences and altruistic behaviour. The presence of altruistic punishers in the population can only lead to more altruistic behaviour in the evolutionary equilibrium by causing the evolved proportion of selfish types to reduce. By contrast, an indirect evolution approach can recognise and model this distinction. Punishment is not carried out "blindly", but when the evolved preferences of the punisher make it rational to do so. This means that there is no longer a simple connection between altruistic preferences and altruistic behaviour. More individuals with altruistic preferences in a population will not necessarily lead to more altruistic behaviour, because altruistic individuals who care about others may not be willing to go through with punishment. On the other hand, more altruistic behaviour may occur without an increase in altruistic preferences, because selfish individuals may be incentivised to behave better by the credible threat of punishment.

The sequential punishment model which forms the workhorse model in this paper has the stark property that *only* the outcome for the selfish phenotypes is improved by the availability of punishment equilibria, because the altruists are totally unwilling to carry out punishment. The key result, proven analytically for any population structure, is thus unambiguous that fewer altruists are able to survive if punishment is used. Although this result is a strong one, and dependent on the specific model used, the phenomenon encapsulated by the model is arguably quite general. Even in other more complex models, this effect should therefore be one of the key factors in play.

3 The Sequential Punishment Model

In the sequential punishment model there are three players. Players 1 and 2 each receive an opportunity in sequence to punish (inflict harm upon) another individual. If they take the opportunity, they gain a felicity benefit $\hat{\pi}$ (where $0 < \hat{\pi} < 1$) and the individual they punish suffers a felicity loss of 1. (By "felicity", we mean the part of an individual's utility which is generated by the individual's *private* consumption of economic goods. Altruistic individuals' utility will thus depend on the felicity of others.) Player 1 first chooses whether or not to inflict harm upon player 2. Player 2, observing player 1's action, can either choose not to inflict harm, to harm player 3, or to harm player 1. Player 2 is assumed to be indifferent as to whom they harm. Player 2's ability to focus harm onto player 1 *if* player 1 inflicts harm creates the potential for a simple punishment scheme to be used to support a subgame-perfect Nash equilibrium where a selfish player 1 is deterred from inflicting harm. (If individuals are indifferent between inflicting harm and not doing so, we assume that they do not.)

Punishment as a useful "social technology" relies upon the sequential structure of interactions. If interactions were simultaneous, the ability and willingness of individuals to impose harm externalities upon others would create deadweight losses with no potential social gain. Once interactions are sequential, punishment can be socially useful and provide a potential substitute to primary altruism. Even when interactions are sequential however, the possibility remains open for agents to use their harm opportunities in the same inefficient way as in simultaneous interactions. Thus, for a fixed level of altruism, there generally exist multiple subgame-perfect Nash equilibria, with the "sequential" one involving use of punishment being Pareto-dominant over the "simultaneous" one.

The sequential punishment model differs substantially from the "public goods with punishment" model which has been used in much of the research on group selection and punishment in experimental economics and anthropology. A recent experimental study has found evidence that direct group competition and the ability to punish under-contributors have a mutually-reinforcing positive effect on the level of co-operation [Sääksvuori et al (2011)].[2] This evidence backs up the theoretical analyses of the evolutionary role of inter-group competition cited at the start of section 2. However, although such models certainly offer a plausible and useful representation of group selection in pre-industrial societies and the origins of the morality contemporary humans have inherited from them, the assumption that *all* members of the group potentially participate in the punishment of *every* undercontributor and that individual payoffs within a group depend upon relative "group success" modelled in a non-individualistic way, can be seen in a broader context as being quite specific and contrived. The sequential punishment model, although highly stylized, has the great advantage of being simple and tractable enough to allow the group selection dynamics to be analytically derived. The model abstracts away from the effects of direct inter-group conflict and also assumes that punishment is costless and takes the form of a single one-off cost to the punishee instead of other sanctions such as ostracism. (Section 6, in concluding, intuitively discusses some of the likely consequences of introducing these complicating factors.)

Imagine there is a large population of individuals, who differ in their level of altruism. This is designated by the coefficient of altruism, θ_i, which is the weighting placed on the felicity of other individuals in individual i's utility function. Since the benefit from punishing always takes the value $\hat{\pi}$, we only need two distinct phenotypes, H and L, which correspond to $\theta_H \geq \hat{\pi}$ and $\theta_L < \hat{\pi}$ respectively. Suppose that the proportion of individuals in the population with phenotype H is q, so that $(1-q)$ have phenotype L. Each period, individuals are randomly chosen to play the sequential punishment game. Individuals are formed into triplets, where two of the individuals are able to actually make a move whilst a third individual is randomly selected to be player 3. This third individual does not play any role except to act as a

[2] In this study, entire groups play a public goods game where each individual starts with an amount of money to split between an individual and a group account. The group account is then doubled and split between the group members. Each individual then observes the contributions to the group account made by others, and can impose a punishment which harms both the punisher and the punishee, at a cost ratio of 1:3. Finally, group competition is introduced via a mechanism whereby a bonus or penalty is applied to each group based upon the *relative* size of the group account after contributions have been made.

passive receptacle for person 2's punishment if person 1 co-operates by not punishing. Nature randomly determines, with equal probability, which individuals will receive their punishment opportunity first and second. All individuals are assumed to have full knowledge of the coefficient of altruism of the others with whom they interact.

4 Derivation of Payoff Matrices

We can think of the sequential punishment game as a subgame nested within a supergame in which the coefficients of altruism chosen by individuals A and B are simultaneously chosen before the sequential punishment game is played. (We refer to the two individuals who are chosen to be players 1 and 2 as A and B before they know who will go first. Player A and player B both have a probability of 0.5 of being in each position.) The choice of the coefficients of altruism by the players then determines the payoffs, and therefore the outcome, of the sequential punishment game nested within. It is, of course, not really appropriate to think of the coefficient of altruism as a strategy chosen, but rather as a phenotype which can be altered via mutations. Also, whereas it is the utility payoff that determines each player's behaviour in the nested game, it is the felicity payoff that determines the evolutionary dynamics.

Since there are two possible phenotypes for each individual, there are four possibilities when three individuals meet and interact. (The phenotype of the individual selected to be player 3 is unimportant because they do not have any opportunity to act.) Firstly, if both individuals have high altruism (i.e. they both "play" strategy H in the supergame) then they both behave efficiently by never punishing. Therefore whichever individual goes first, the felicity payoff to each individual is zero. The value of the social welfare function is also zero because no punishment occurs, and therefore all three individuals get a felicity payoff of 0. (The per-period social welfare function sums the felicity of the two individuals who get a punishment opportunity along with the felicity of the third individual.) This can be seen in the upper payoff matrix in figure 1 in which the top left square shows the zero payoffs of individuals 1 and 2, and the resulting zero social welfare in the middle of the square. Similarly, the corresponding square in the lower matrix (which shows the felicity payoffs for players A and B, once the chance of being player 1 or 2 has been randomized, and is therefore symmetric) is identical, because the payoffs are still zero whoever gets to move first.

Suppose instead that player 2 has phenotype L and player 1 has phenotype H. Since there is no future in which she can be punished, player 2 will inefficiently punish. Player 1 will still not punish because he is sufficiently altruistic not to do this in a single-move game anyway. Therefore player 2 will get a felicity payoff of $\hat{\pi}$ and player 1 will get a felicity payoff of 0, because he co-operates and so player 2 follows her default behaviour and punishes player 3. Total social welfare is therefore $\hat{\pi} - 1$. On the other hand, supposing that player 1 has phenotype L and player 2 has phenotype H, player 2 will not punish, and so there is then no credible threat to punish player 1 for inflicting harm, and so player 1 will do so. In this case, player 1 gets a felicity payoff of $\hat{\pi}$ and player 2 gets -1 because she is punished by player 1. Again, social welfare is $\hat{\pi} - 1$. In the lower matrix, the payoffs for individuals A

1's Phenotype		**H**	**L**
2's Phenotype		$\theta_1 \geq \hat{\pi}$	$\theta_1 < \hat{\pi}$
H	$\theta_2 \geq \hat{\pi}$	0 0 0	$\hat{\pi}$ $\hat{\pi}-1$ -1
L	$\theta_2 < \hat{\pi}$	0 $\hat{\pi}-1$ $\hat{\pi}$	0 $\hat{\pi}-1$ $\hat{\pi}$

A's Phenotype		**H**	**L**
B's Phenotype		$\theta_A \geq \hat{\pi}$	$\theta_A < \hat{\pi}$
H	$\theta_B \geq \hat{\pi}$	0 0 0	$\underline{\hat{\pi}}$ $\hat{\pi}-1$ $-\frac{1}{2}$
L	$\theta_B < \hat{\pi}$	$-\frac{1}{2}$ $\hat{\pi}-1$ $\underline{\hat{\pi}}$	$\underline{\frac{\hat{\pi}}{2}}$ $\hat{\pi}-1$ $\underline{\frac{\hat{\pi}}{2}}$

Fig. 1 Strategic form for sequential-move supergame. (Row player receives bottom left payoff in each cell, column player top right. Social welfare is shown in the centre of each cell. Players A and B in the lower matrix each have a 50% chance of being Player 1 or Player 2 in the sequential punishment game.)

and B in the bottom left and top right squares are found by averaging the payoffs in the corresponding squares from the first matrix to produce a new symmetric matrix, because players A and B have an equal chance of being player 1 or 2.

Finally, we have the case where both individuals have phenotype L. Here, player 2 will definitely punish because there is no future. However, this allows a credible threat to be made to player 1 that if he punishes socially inefficiently, player 2's punishment will be switched from player 3 onto him. If this occurs, player 1 loses social utility of $1-\theta_1$.[3] However, the gain in social utility he gets by punishing is only $\hat{\pi}-\theta_1$. Player 1 will therefore be effectively deterred from punishing. Player 1's felicity payoff is therefore 0 and player 2's is $\hat{\pi}$. Social welfare will be $\hat{\pi}-1$. The payoffs for the second matrix are again found by averaging, in order to take into account the equal chance of players A and B being player 1 or 2 in the first matrix.

The best response payoffs in the second matrix are underlined, and the unique (dominance solvable) pure strategy Nash equilibrium is for both individuals A and B to have phenotype L. Since each player is always better off in felicity terms by having low altruism, regardless of whether the other individual has high or low altruism, the individual level selection pressure in this model leads to a socially inefficient evolutionary equilibrium, in the same way as in the standard prisoners' dilemma.

Before further analysing the properties of this evolutionary equilibrium, it is instructive to compare it to that of an almost identical model, except that rather than

[3] This is assuming, for simplicity, no discounting. Permitting discounting would be problematic because we would then have to decide whether or not to discount felicity payoffs as well as social utility payoffs. It would also not really add anything insightful to the analysis of a finite-move sequential game.

A's Phenotype		**H**	**L**
B's Phenotype		$\theta_A \geq \hat{\pi}$	$\theta_A < \hat{\pi}$
H	$\theta_B \geq \hat{\pi}$	0 0 0	$\underline{\hat{\pi}}$ $\hat{\pi}-1$ $-\frac{1}{2}$
L	$\theta_B < \hat{\pi}$	$-\frac{1}{2}$ $\hat{\pi}-1$ $\underline{\hat{\pi}}$	$\underline{\hat{\pi}-\frac{1}{2}}$ $2\hat{\pi}-2$ $\underline{\hat{\pi}-\frac{1}{2}}$

Fig. 2 Strategic form for simultaneous-move supergame. (Row player receives bottom left payoff in each cell, column player top right. Social welfare is shown in the centre of each cell. Players A and B each have a 50% chance of being Player 1 or Player 2 in the sequential punishment game.)

having a two-move sequential punishment model nested within the supergame, there is instead a game where each individual chooses whether or not to punish in a single-move game simultaneously. So, person A punishes if and only if $\theta_A < \hat{\pi}$ and person B if and only if $\theta_B < \hat{\pi}$. (We continue to assume that person 1 will punish person 2 by default and that person 2 will punish person 3 by default. Thus individuals A and B only take the felicity loss of -1 if they turn out to be person 2, with probability $\frac{1}{2}$.) The payoff matrix for this model is shown in figure 2. Although the evolutionarily stable equilibrium is again for all individuals to have phenotype L, the important difference compared to the case where the nested game is sequential is that in the evolutionary equilibrium for this model, both individuals will punish, whereas in the case of the two-move sequential move game, although all individuals have low altruism in the evolutionary equilibrium, the individual who has a chance to punish first does not punish, due to the threat of having player 2's punishment focused on to him if he defects by punishing. This difference between the two models will turn out to be of crucial importance in determining the nature of their group selection dynamics.

The relevant difference between the sequential-move and simultaneous-move versions of the model can be brought out if we consider the effect on social welfare of a marginal increase in the proportion of the population with high altruism (phenotype H) from the evolutionarily stable equilibrium in a single homogeneous population. The expected value of the social welfare function, $E(W)$, depends upon the proportion of each phenotype in the population. In the case of the nested sequential-move punishment model, in a finite population of size n, this will be:

$$E\left(W_{seq}\right) = \frac{q\left(nq-1\right)0}{n-1} + 2\,\frac{q\left(1-q\right)n\left(\hat{\pi}-1\right)}{n-1} + \frac{\left(1-q\right)\left(n\left(1-q\right)-1\right)\left(\hat{\pi}-1\right)}{n-1} \quad (1)$$

In the case of the nested simultaneous-move punishment model, this will be:

$$E\left(W_{sim}\right) = \frac{q\left(nq-1\right)0}{n-1} + 2\,\frac{q\left(1-q\right)n\left(\hat{\pi}-1\right)}{n-1} + \frac{\left(1-q\right)\left(n\left(1-q\right)-1\right)\left(2\,\hat{\pi}-2\right)}{n-1} \quad (2)$$

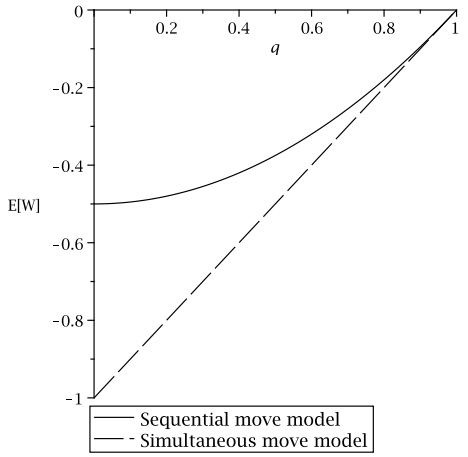

Fig. 3 Comparison of social welfare in simultaneous-move and sequential-move models

If we now differentiate these expressions with respect to q, we find an expression for the gains in social welfare from a marginal increase in the proportion of altruists:

$$\frac{d}{dq}E\left(W_{seq}\right) = \frac{(2nq-1)}{n-1}\left(1-\hat{\pi}\right) \qquad \frac{d}{dq}E\left(W_{sim}\right) = 2\left(1-\hat{\pi}\right) \tag{3}$$

As $n \longrightarrow \infty$ so that we get a continuous population, (3) goes to:

$$\frac{d}{dq}E\left(W_{seq}\right) = 2q\left(1-\hat{\pi}\right) \qquad \frac{d}{dq}E\left(W_{sim}\right) = 2\left(1-\hat{\pi}\right) \tag{4}$$

Figure 3 shows social welfare as a function of q for both types of nested model, letting $\hat{\pi} = \frac{1}{2}$ and taking the limit as $n \longrightarrow \infty$. We see that at the evolutionary equilibrium where $q = 0$, the marginal increase in social welfare when q increases is positive for the simultaneous-move model but 0 for the sequential-move model. This is because introducing a small number of high altruism individuals into a population of low altruism individuals means that they are almost certain to interact with low altruism individuals. In the sequential-move game, however, if the new high altruism individual moves first then they do not change their behaviour, whereas if they go second then although they do not punish this causes the low altruism individual to switch to punishing. So, altruism is only socially beneficial in the sequential punishment model when altruists encounter each other rather than low altruism individuals. In the simultaneous-move game, by contrast, the presence of even a small number of high altruism individuals is socially beneficial because even if they do interact with a low altruism individual, their behaviour is changed because they now do not punish, and this increases social efficiency even though the low altruism individual they interact with still punishes. Marginal injections of altruism are therefore not as socially beneficial in the sequential-move game. This gives us the intuition for why the group selection mechanism is weaker.

5 Price Equations

The conditions under which group-level selection pressures will dominate, and altruism will evolve, can be described using Price equations [Price (1970)]. It is clear that altruists will always be wiped out in the long run in a single isolated group because the selfish individuals always get better expected felicity payoffs. However, if altruists are sufficiently concentrated together in sub-groups within the population, whose members interact only (or mainly) with one another, then altruists can get better expected payoffs than selfish types, because the benefits of altruism will be focused mainly upon other altruists. For this to work, there must be a dispersal mechanism which has the property that it allows altruists to migrate between groups whilst maintaining sufficient inter-group variance of the altruism level relative to the intra-group variance. The Price Equation for a particular model establishes the minimum variance ratio required to enable altruism to survive.

We will show that the sequential-move game described above leads to a more stringent requirement on the variance ratio achieved by the dispersal mechanism than the simultaneous-move game. This means that altruism will evolve to a higher degree if the social control mechanism provided by person 2's threat to punish person 1 is removed. Although this threat prevents person 1 from punishing, and therefore causes a social welfare gain, *ceteris paribus*, the use of punishment also weakens the group selection mechanism, and thus the chance of achieving a high altruism equilibrium "anarchically". We will proceed by first showing that the simultaneous-move version of the sequential punishment model (which parallels *one* of the multiple equilibria in the sequential punishment model, where punishment is not used) has essentially the same Price equation as the standard prisoners' dilemma.[4] We will then derive the Price equation for the sequential-move version of the model.

The standard model of group selection involves a population split into m groups with average size n, so that $n = \frac{1}{m}\sum_{i=1}^{m} n_i$, where n_i is the number of individuals in group i. Individuals in each group only interact with members of their own group. If all individuals are in a single group then since the unique Nash equilibrium in the supergame is where both individuals play L, then (just like the prisoners' dilemma) the only evolutionarily stable state is for all individuals to have phenotype L. With multiple groups, however, we can show that conditions exist under which the altruistic phenotype can propagate. By deriving the change which will occur in the number of high altruism individuals and the total population, we are able to derive the condition for the proportion of high altruism individuals to increase. The payoffs for a member of a particular group depend upon the proportion in the group of each type, with q_i being the proportion of high altruism types in group i and $q = \frac{1}{mn}\sum_{i=1}^{m} n_i q_i$ being the proportion of altruists in the population. Let strategy notation $A \to i, B \to j$ mean "if player 1 punishes me (player 2) [history A] then punish player i, if player 1 does not punish me [history B] then punish player j".

[4] A high altruism individual *refraining* from punishing and imposing the cost of 1 on the other individual at benefit $\hat{\pi}$ to herself is equivalent to bestowing a benefit of 1 upon the other at a cost of $\hat{\pi}$ to herself. The simultaneous-move punishment game therefore has almost the same payoffs as the prisoners' dilemma, except that person 2, if she has phenotype L, punishes person 3 instead of person 1.

5.1 Selfish player 2 plays strategy $A \to 1, B \to 1$

Suppose first of all that a selfish player 2 (a player 2 with a low altruism phenotype, L) chooses to play the strategy $A \to 1, B \to 1$ so that she always punishes player 1 regardless of whether player 1 punishes her or not (as we have already seen, since there is no future an altruistic player 2 with phenotype H will never punish). A selfish player 1 will therefore definitely choose to punish player 2, because he will be punished anyway and so will optimally wish to take his opportunity to punish for a gain in his utility. This is a subgame-perfect Nash equilibrium because once player 1 has moved, player 2 will be indifferent about punishing player 1 or player 3.

We derive the Price equation condition by finding the expected felicity payoff of an altruistic individual and a selfish individual. Altruistic individuals have a $\frac{1}{3}$ chance of being player 1, 2 or 3. If they are player 1, then if player 2 is selfish (with probability $\frac{(1-q_i)n_i}{n_i-1}$), they will receive a felicity payoff of -1. Otherwise they will receive a felicity payoff of 0. If they are player 2, they also will receive a felicity payoff of -1 if player 1 is selfish (probability $\frac{(1-q_i)n_i}{n_i-1}$), and 0 otherwise. If they are player 3, they will always receive a payoff of 0, because they are never punished. So:[5]

$$U_i^H = f - \frac{2}{3}\frac{(1-q_i)n_i}{n_i-1} \tag{5}$$

If players are selfish, then if they turn out to be player 1, they will definitely choose to punish player 2, who will punish them in turn if they too are selfish. The expected payoff if they are player 1 would therefore be $\hat{\pi} - \frac{(1-q_i)n_i-1}{n_i-1}$. If they turn out to be player 2, they will again definitely punish, and player 1 will punish them if they are selfish. Again, the expected payoff would be $\hat{\pi} - \frac{(1-q_i)n_i-1}{n_i-1}$. As before, if they turn out to be player 3, their expected payoff is definitely 0. So:

$$U_i^L = f + \frac{2}{3}\hat{\pi} - \frac{2}{3}\frac{(1-q_i)n_i-1}{n_i-1} \tag{6}$$

The new proportion of altruists after one stage of interaction will be:

$$q' = \frac{\sum_{i=1}^{m}\left(f - \frac{2}{3}\frac{(1-q_i)n_i}{n_i-1}\right)q_i n_i}{\sum_{i=1}^{m}\left(\left(f - \frac{2}{3}\frac{(1-q_i)n_i}{n_i-1}\right)q_i n_i + \left(f + \frac{2}{3}\hat{\pi} - \frac{2}{3}\frac{(1-q_i)n_i-1}{n_i-1}\right)(1-q_i)n_i\right)}$$

Multiplying out, dividing numerator and denominator by n and collecting terms:

$$q' = \frac{\left(\sum_{i=1}^{m}\frac{3}{2}\frac{q_i n_i f}{n} - \sum_{i=1}^{m}\frac{q_i {n_i}^2}{n(n_i-1)} + \sum_{i=1}^{m}\frac{{q_i}^2 {n_i}^2}{n(n_i-1)}\right)}{\left(\sum_{i=1}^{m}\frac{3}{2}\frac{f n_i}{n} + \sum_{i=1}^{m}\frac{q_i {n_i}^2}{n(n_i-1)} + \sum_{i=1}^{m}\frac{\hat{\pi} n_i}{n} - \sum_{i=1}^{m}\frac{\hat{\pi} n_i q_i}{n} - \sum_{i=1}^{m}\frac{n_i}{n} - \sum_{i=1}^{m}\frac{q_i n_i}{n(n_i-1)}\right)}$$

[5] U_i^H and U_i^L are the expected felicity payoffs in group i. Note also the introduction of a fixed payoff f. This is the same for both phenotypes and thus has no effect on relative fitness, but is needed to ensure that both types always gain a strictly positive payoff.

The following expressions are now used to reformulate the above expression:

$$\begin{gathered}
\sum_{i=1}^{m} \frac{n_i}{n} = \sum_{i=1}^{m} \frac{n_i(n_i-1)}{n(n_i-1)} = \sum_{i=1}^{m} \frac{{n_i}^2}{n(n_i-1)} - \sum_{i=1}^{m} \frac{n_i}{n(n_i-1)} \\
\sum_{i=1}^{m} \frac{q_i {n_i}^2}{n_i-1} = mCov\left(\frac{q_i n_i}{n_i-1}, n_i\right) + \sum_{i=1}^{m} \frac{q_i n_i}{n_i-1} n \\
\sum_{i=1}^{m} \frac{{q_i}^2 {n_i}^2}{n_i-1} = mCov\left(\frac{q_i n_i}{n_i-1}, q_i n_i\right) + \sum_{i=1}^{m} \frac{q_i n_i}{n_i-1} nq \\
\sum_{i=1}^{m} \frac{{n_i}^2}{n_i-1} = mCov\left(\frac{n_i}{n_i-1}, n_i\right) + n\sum_{i=1}^{m} \frac{n_i}{n_i-1} \\
\sum_{i=1}^{m} \frac{q_i n_i}{n_i-1} = mE\left(\frac{q_i n_i}{n_i-1}\right) \qquad \sum_{i=1}^{m} q_i n_i = qnm \\
\sum_{i=1}^{m} n_i = mn \qquad \sum_{i=1}^{m} \frac{n_i}{n_i-1} = mE\left(\frac{n_i}{n_i-1}\right)
\end{gathered} \tag{7}$$

We can now apply (7) to derive the following expression for the change in the proportion of altruists in the overall population:

$$q' - q = \frac{\left(Cov\left(\frac{q_i n_i}{n_i-1}, q_i n_i\right) - (1+q)Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right) - (n-q)E\left(\frac{q_i n_i}{n_i-1}\right) + q(n-1)E\left(\frac{n_i}{n_i-1}\right) + q\left(Cov\left(\frac{n_i}{n_i-1}, n_i\right) - \hat{\pi} n(1-q)\right)\right)}{\left(\frac{3}{2} fn + Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right) + (n-1)\left(E\left(\frac{q_i n_i}{n_i-1}\right) - E\left(\frac{n_i}{n_i-1}\right)\right) - Cov\left(\frac{n_i}{n_i-1}, n_i\right) + \hat{\pi} n(1-q)\right)} \tag{8}$$

Provided f is high enough so that both types always get a positive payoff, the denominator of (8) will be positive, and $q' - q > 0$ if and only if the following holds:

$$\hat{\pi} < -\frac{(1+q)Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right)}{qn(1-q)} + \frac{Cov\left(\frac{q_i n_i}{n_i-1}, q_i n_i\right)}{qn(1-q)} - \frac{(n-q)E\left(\frac{q_i n_i}{n_i-1}\right)}{qn(1-q)} + \frac{(n-1)E\left(\frac{n_i}{n_i-1}\right)}{n(1-q)} + \frac{Cov\left(\frac{n_i}{n_i-1}, n_i\right)}{n(1-q)} \tag{9}$$

This result can be most easily interpreted in the situation where all groups are of equal size, so $E\left(\frac{q_i n_i}{n_i-1}\right) = \frac{qn}{n-1}$, $E\left(\frac{n_i}{n_i-1}\right) = \frac{n}{n-1}$, $Cov\left(\frac{q_i n_i}{n_i-1}, q_i n_i\right) = \frac{n^2}{n-1} Var(q_i)$, $Cov\left(\frac{n_i}{n_i-1}, n_i\right) = 0$ and $Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right) = 0$. Conditions (8) and (9) then become:

$$q' - q = \frac{2\left(nVar\left(q_i\right) - q\left(n-1\right)\left(1-q\right)\hat{\pi} - q(1-q)\right)}{\left(n-1\right)\left(3f - 2(1-q)(1-\hat{\pi})\right)} \tag{10}$$

$$\hat{\pi} < \frac{nVar\left(q_i\right)}{q\left(n-1\right)\left(1-q\right)} - \frac{1}{\left(n-1\right)} \tag{11}$$

The intuition for this result is that altruism is able to survive if altruists are sufficiently concentrated together that they have a higher average fitness level than the selfish types. Within a particular group, selfish individuals still do better than altruistic individuals, but across the population, altruists are able to do better than selfish individuals because the altruistic groups spread more rapidly. The $Var_i(q_i)$ part of the above condition is the inter-group variance of the level of altruism. The $q(1-q)$ part is the intra-group variance (the variance of the random variable formed by taking a single individual from the population and assigning a value of 1 if they have phenotype H and 0 if they have phenotype L). As $n \longrightarrow \infty$, (11) simplifies even further to give $\hat{\pi} < \frac{Var(q_i)}{q(1-q)}$; the variance ratio must be greater than the ratio of the cost of co-operating ($\hat{\pi}$) to the benefit bestowed upon the other individual (i.e. 1).

5.2 Selfish player 2 plays strategy $A \to 1, B \to 3$

We will now assume that player 2 always plays strategy $A \to 1, B \to 3$, so that player 1 is always induced not to punish if player 2 is of type L. Taking first the expected felicity payoff of an altruistic individual, they have a $\frac{1}{3}$ chance of being player 1 in their interaction. In this case, whether or not player 2 is altruistic, the individual will not punish, and so will receive a payoff of 0. If, on the other hand, they turn out to be player 2 (probability $\frac{1}{3}$), they will be punished by player 1 if player 1 is selfish (probability $\frac{(1-q_i)n_i}{n_i-1}$ and suffering a loss of 1), because they can make no credible threat to punish a selfish player 1 for doing this. The third possibility is that they will be player 3, in which case they will be punished by player 2 if player 2 turns out to be selfish (probability $\frac{(1-q_i)n_i}{n_i-1}$ and suffering a loss of 1). This is because even if player 1 turns out to be selfish, he will never choose to punish player 2 due to his fear of being punished by player 2. Hence, a selfish player 2 will always punish player 3. So, the expected felicity payoff of an altruistic individual in this model is:

$$U_i^H = f - \frac{2}{3}\frac{(1-q_i)n_i}{n_i-1} \tag{12}$$

Now we take the case of selfish individuals. If they turn out to be player 1, they will choose to punish player 2 if and only if player 2 is altruistic (the unconditional probability of this scenario is $\frac{1}{3}\frac{q_i n_i}{n_i-1}$ and the felicity payoff would be $\hat{\pi}$). If selfish individuals turn out to be player 2 (probability $\frac{1}{3}$), then they will definitely punish either player 1 or player 3, gaining a felicity payoff of $\hat{\pi}$. If they turn out to be player 3, they are in the same situation as they would be if they were altruistic, except that the probability that player 2 is selfish and punishes them is now $\frac{(1-q_i)n_i-1}{n_i-1}$. The expected utility payoff of an altruistic individual will therefore be:

$$U_i^L = f + \frac{1}{3}\frac{\hat{\pi} q_i n_i}{n_i-1} + \frac{1}{3}\hat{\pi} - \frac{1}{3}\frac{(1-q_i)n_i-1}{n_i-1} \tag{13}$$

From the above, we can see that, once interactions have occurred and reproduction has taken place, the new proportion of high altruism individuals will be:

$$q' = \frac{\sum_{i=1}^m \left(f - \frac{2}{3}\frac{(1-q_i)n_i}{n_i-1}\right) q_i n_i}{\sum_{i=1}^m \left(\left(f - \frac{2}{3}\frac{(1-q_i)n_i}{n_i-1}\right) q_i n_i + \left(f + \frac{1}{3}\frac{\hat{\pi} n_i q_i}{n_i-1} + \frac{1}{3}\hat{\pi} - \frac{1}{3}\frac{(1-q_i)n_i-1}{n_i-1}\right)(1-q_i)n_i\right)} \tag{14}$$

Multiplying out, dividing numerator and denominator by n and collecting terms:

$$q' = \frac{2\left(\sum_{i=1}^m \frac{3}{2}\frac{q_i n_i f}{n} - \sum_{i=1}^m \frac{q_i {n_i}^2}{n(n_i-1)} + \sum_{i=1}^m \frac{{q_i}^2 {n_i}^2}{n(n_i-1)}\right)}{\left(3\sum_{i=1}^m \frac{f n_i}{n} + \sum_{i=1}^m \frac{q_i {n_i}^2(1-\hat{\pi})}{n(n_i-1)} + \sum_{i=1}^m \frac{\hat{\pi} q_i {n_i}^2}{n(n_i-1)} + \sum_{i=1}^m \frac{\hat{\pi} n_i}{n} - \sum_{i=1}^m \frac{\hat{\pi} n_i q_i}{n} - \sum_{i=1}^m \frac{n_i}{n} - \sum_{i=1}^m \frac{q_i n_i}{n(n_i-1)}\right)}$$

The expressions from (7) can now be applied to derive the following:

$$\begin{aligned} q'-q = & \Bigg((2-q(1-\hat{\pi}))Cov\left(\tfrac{q_i n_i}{n_i-1}, q_i n_i\right) - (2+\hat{\pi}q)Cov\left(\tfrac{q_i n_i}{n_i-1}, n_i\right) \\ & -\left(\left((1-q)(q\hat{\pi}+2)+q^2\right)n-q\right)E\left(\tfrac{q_i n_i}{n_i-1}\right) + (n-1)qE\left(\tfrac{n_i}{n_i-1}\right) - q(1-q)\hat{\pi}n + qCov\left(\tfrac{n_i}{n_i-1}, n_i\right)\Bigg) \\ & \Bigg(3fn + \hat{\pi}\,Cov\left(\tfrac{q_i n_i}{n_i-1}, n_i\right) + (1-\hat{\pi})Cov\left(\tfrac{q_i n_i}{n_i-1}, q_i n_i\right) \\ & -(1-((1-q)\hat{\pi}+q)n)E\left(\tfrac{q_i n_i}{n_i-1}\right) - (n-1)E\left(\tfrac{n_i}{n_i-1}\right) + (1-q)\hat{\pi}n - Cov\left(\tfrac{n_i}{n_i-1}, n_i\right)\Bigg)^{-1} \end{aligned} \tag{15}$$

Assuming f is high enough to make the denominator of the RHS of (15) positive, $q'-q$ will be positive if and only if the following condition holds:

$$\hat{\pi} < \frac{(2-q)Cov\left(\tfrac{q_i n_i}{n_i-1}, q_i n_i\right) - 2\,Cov\left(\tfrac{q_i n_i}{n_i-1}, n_i\right) - \left(n(1-q)^2+n-q\right)E\left(\tfrac{q_i n_i}{n_i-1}\right) + q(n-1)E\left(\tfrac{n_i}{n_i-1}\right) + qCov\left(\tfrac{n_i}{n_i-1}, n_i\right)}{q\left(Cov\left(\tfrac{q_i n_i}{n_i-1}, n_i\right) - Cov\left(\tfrac{q_i n_i}{n_i-1}, q_i n_i\right) + n(1-q)\left(E\left(\tfrac{q_i n_i}{n_i-1}\right)+1\right)\right)} \tag{16}$$

Condition (16) will be shown in Theorem 1 to be more stringent than the equivalent condition (9) for the simultaneous-move game:

Theorem 1 *If $\hat{\pi}'_{seq} > 0$, then $\hat{\pi}'_{sim} > \hat{\pi}'_{seq}$.*

Proof Let $\hat{\pi}'_{sim}$ be the RHS of (9) and $\hat{\pi}'_{seq}$ be the RHS of (16). The following substitutions can be used to rewrite (9) and (16) in a more easily comparable form:

$$\alpha_i = Cov\left(\frac{n_i}{n_i-1}, n_i\right) + nE\left(\frac{n_i}{n_i-1}\right) - Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right) - nE\left(\frac{q_i n_i}{n_i-1}\right) \tag{17}$$

$$\beta_i = E\left(\frac{n_i}{n_i-1}\right) - E\left(\frac{q_i n_i}{n_i-1}\right) \tag{18}$$

$$\gamma_i = Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right) + nE\left(\frac{q_i n_i}{n_i-1}\right) - Cov\left(\frac{q_i n_i}{n_i-1}, q_i n_i\right) - qnE\left(\frac{q_i n_i}{n_i-1}\right) \tag{19}$$

Note that $\alpha_i > \beta_i > 0$ and that $\alpha_i > \gamma_i > 0$. Using these substitutions, (9) and (16) become:

$$\hat{\pi}'_{sim} = \frac{(\alpha_i-\beta_i)\,q-\gamma_i}{qn\,(1-q)} \qquad \hat{\pi}'_{seq} = \frac{(\alpha_i-\beta_i)\,q-(2-q)\,\gamma_i}{(\gamma_i+n\,(1-q))\,q} \tag{20}$$

It can now be seen clearly by observation that if $\hat{\pi}'_{seq} > 0$, then $\hat{\pi}'_{sim} > \hat{\pi}'_{seq}$. □

6 Conclusion

This paper has argued that the use of punishment in combination with altruism may be a "double-edged sword" for successful social co-operation. This is because selfish phenotypes are more willing to make use of opportunities to harm others, which in turn facilitates a more efficient static equilibrium, and thus improves the relative performance of selfish groups. Some concluding remarks should, however, be made about the robustness and empirical implications of this result.

The sequential punishment model abstracts away from the deadweight loss of punishment (which is zero because a selfish player 2 would harm player 3 even in a simultaneous equilibrium), the cost of altruistic punishment (which is zero because players are indifferent as to whom they punish) and evolution of altruism via mechanisms other than random dispersal (e.g. direct inter-group competition, group punishment in public goods games or assortative selection such as ostracism). Under these strong assumptions, the result is unambiguous that the evolution of altruism is undermined by the use of punishment.

Introducing costly punishment, direct inter-group competition and assortative selection is likely to make the result ambiguous. Greater benefits to altruism through aiding group victory in conflict, alongside the presence of costs to punishment, will intuitively increase the performance differential between more altruistic and less altruistic groups. Assortative selection increases the cost of being selfish in an altruistic group, thus likewise also improving the relative performance of altruistic groups. The dynamic effect identified in its purest form using the sequential punishment model presented in this paper should nonetheless be a key element at work in other models. There is clear potential for further research to explore the interaction of these various factors in detail.

In terms of empirical implications, it is important to understand that the sequential punishment model predicts that the use of punishment has a statically beneficial effect upon observed behaviour. The model is thus consistent with the empirical observation that co-operation is more easily achieved in experimental games with punishment. Where the model does have empirical implications and explanatory power is in its prediction that the use of different regimes of extrinsic incentives will result over time in different levels of altruistic preference evolving. The model can therefore potentially help to explain the wide variety of different preferences implied by experimental observations, and in particular the wide variability *between societies*.

The broader implications of the result are potentially very significant. It helps to shed light upon the historical transition away from pre-industrial societies based upon rigid codes of morality. Societies which develop more extensive systems of extrinsic social control have less need of, but simultaneously undermine, intrinsic morality. (Whether this results in greater social welfare is beyond the scope of this paper. There does appear, however, to be a trade-off, since punishment has beneficial static but detrimental dynamic effects.) This theory of the relationship between individual morals and society contradicts the traditional view that the enforcement of the law strengthens the inner morality of the citizen. Whilst we will not go so far as to conclude that this traditional view is totally mistaken, the results of this paper suggest that matters are at least more complicated than it recognises. Even advanced industrial societies still rely upon intrinsic morality as well as extrinsic incentive systems. The result thus also has important potential policy implications in that the introduction of new extrinsic control systems may have negative dynamic welfare effects by undermining intrinsic moral self-restraint.

Acknowledgements I would like to express my gratitude to the United Kingdom Economic and Social Research Council and to the Royal Economic Society for funding this research. I would also like to thank Kevin Roberts, Chris Wallace, Peter Hammond and the referee for their useful comments, suggestions and feedback on various drafts of this paper. The usual disclaimers apply.

References

Becker GS (1976) Altruism, egoism, and genetic fitness: Economics and sociobiology. Journal of Economic Literature 14(3):817–826

Bergstrom TC (2002) Evolution of social behavior: Individual and group selection. Journal of Economic Perspectives 16(2):67–88

Blackmore SJ (1999) The Meme Machine. Oxford University Press, Oxford.

Bowles S (2006) Group competition, Reproductive Leveling, and the Evolution of Human Altruism. Science 314(5805):1569–1572

Bowles S (2009) Did Warfare Among Ancestral Hunter-gatherers Affect the Evolution of Human Social Behaviors? Science 324(5932):1293–1298

Boyd R, Richerson PJ (1982) Cultural transmission and the evolution of cooperative behavior. Human Ecology 10(3):325–351

Boyd R, Gintis H, Bowles S, Richerson PJ (2003) The evolution of altruistic punishment. Proceedings of the National Academy of Sciences 100(6)

Cooper B, Wallace C (2004) Group selection and the evolution of altruism. Oxford Economic Papers 56:307–330

Darwin C (1872) On the origin of species by means of natural selection: Or, the preservation of favored races in the struggle for life, 5th edn. D. Appleton, New York

Denant-Boemont L, Masclet D, Noussair CN (2007) Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. Economic theory 33(1):145–167

Egas M, Riedl A (2008) The economics of altruistic punishment and the maintenance of cooperation. Proceedings of the Royal Society B: Biological Sciences 275(1637):871–878

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

Fehr E, Gächter S (2000a) Cooperation and punishment in public goods experiments. The American Economic Review 90(4):980–994

Fehr E, Gächter S (2000b) Fairness and retaliation: The economics of reciprocity. The Journal of Economic Perspectives 14(3):159–181

Fehr E, Gächter S (2002a) Altruistic punishment in humans. Nature 415:137–140

Fehr E, Gächter S (2002b) Strong reciprocity, human cooperation and the enforcement of social norms. Human Nature 13:1–25

Güth W, Yaari M (1992) An evolutionary approach to explain reciprocal behavior in a simple strategic game. In: Witt U (ed) Explaining Process and Change: Approaches to Evolutionary Economics, University of Michigan Press, pp 23–34

Hamilton WD (1963) The evolution of altruistic behavior. The American Naturalist 97(896):354–356

Hamilton WD (1972) Altruism and related phenomena, mainly in social insects. Annual Review of Ecology and Systematics 3(1):193–232

Hayek FA (1988) The Fatal Conceit: The Errors of Socialism. Routledge

Herrmann B, Thöni C, Gächter S (2008) Antisocial punishment across societies. Science 319(5868):1362–1367

Hirshleifer J (1977) Economics from a biological viewpoint. The Journal of Law and Economics 20(1):1

Huck S, Oechssler J (1999) The indirect evolutionary approach to explaining fair allocations. Games and Economic Behavior 28(1):13–24

Nikiforakis N (2008) Punishment and counter-punishment in public good games: Can we really govern ourselves? Journal of Public Economics 92(1):91–112

Nikiforakis N, Normann HT (2008) A comparative statics analysis of punishment in public-good experiments. Experimental Economics 11(4):358–369

Price GR (1970) Selection and covariance. Nature 227:520–521

Rand DG, Armao IV JJ, Nakamaru M, Ohtsuki H (2010) Anti-social punishment can prevent the co-evolution of punishment and cooperation. Journal of theoretical biology 265(4):624–632

Sääksvuori L, Mappes T, Puurtinen M (2011) Costly punishment prevails in intergroup conflict. Proceedings of the Royal Society B: Biological Sciences 278(1723):3428–3436

Smith A (1976) An Inquiry into the Nature and Causes of the Wealth of Nations. Oxford University Press, Oxford.

Sober E, Wilson DS (1994) Reintroducing group selection to the human behavioral sciences. Behavioral and Brain Sciences 17(4):585–654+

Sober E, Wilson DS (1999) Unto Others. Harvard University Press, Cambridge, Massachussets.

Soltis J, Boyd R, Richerson P (1995) Can group-functional behaviors evolve by cultural group selection?: An empirical test. Current Anthropology 36(3):473–494