The evolution of anti-social rewarding and its countermeasures in public goods games

Miguel dos Santos

Department of Ecology and Evolution, Biophore, University of Lausanne, Lausanne 1015, Switzerland

Cooperation in joint enterprises can easily break down when self-interests are in conflict with collective benefits, causing a tragedy of the commons. In such social dilemmas, the possibility for contributors to invest in a common pool-rewards fund, which will be shared exclusively among contributors, can be powerful for averting the tragedy, as long as the second-order dilemma (i.e. withdrawing contribution to reward funds) can be overcome (e.g. with second-order sanctions). However, the present paper reveals the vulnerability of such pool-rewarding mechanisms to the presence of reward funds raised by defectors and shared among them (i.e. anti-social rewarding), as it causes a cooperation breakdown, even when second-order sanctions are possible. I demonstrate that escaping this social trap requires the additional condition that coalitions of defectors fare poorly compared with pro-socials, with either (i) better rewarding abilities for the latter or (ii) reward funds that are contingent upon the public good produced beforehand, allowing groups of contributors to invest more in reward funds than groups of defectors. These results suggest that the establishment of cooperation through a collective positive incentive mechanism is highly vulnerable to anti-social rewarding and requires additional countermeasures to act in combination with second-order sanctions.

1. Introduction

Understanding the emergence and maintenance of cooperation when individual self-interests are in conflict with collective benefits has been a major challenge for social scientists, economists and biologists. Because contributing to a public good that will benefit others is usually costly, there is an incentive to 'free-ride' and profit from the cooperators' contributions. Cooperation cannot evolve in such social dilemmas, and cooperative societies are easily invaded by defectors, causing a tragedy of the commons [1].

Solutions to this tragedy rely on the presence of additional incentives [2–5]. While negative incentives such as punishment are used to reduce the defectors' payoff, positive ones such as rewards can be used to increase the contributors' welfare. Incentives are used at all levels of human organization in order to enhance cooperation, from gift giving and teamwork rewarding to community-based resource management and international organizations such as the UN [5–8]. Facing the threat of punishment or the promise of rewards can effectively induce individuals to contribute in public goods games (PGGs) [9–17]. Investing in incentives is individually costly and hence poses a second-order dilemma, as those who do not contribute to the incentive system fare better. Thus, individuals providing these incentives must benefit from them in the long run, for example, with increased cooperation [18] or reputational benefits [19–23]. Alternatively, the possibility to abstain from participating in the collective action can pave the way for cooperative behaviours [24].

While most theoretical and empirical studies focused on incentive systems where sanctions or rewards are administered by peers [11,13–24], it is only recently that scholars have started to investigate mechanisms that can be viewed as primitive institutions, namely 'pool' mechanisms [25–31]. The idea is that
individuals can decide to pool their effort and provision an incentive fund before contributing to the PGG. The fund is increased by a productivity factor and then used to administer equally the corresponding incentives, punishing defectors or exclusively rewarding cooperators (note that one of the first experiments on punishment in PGG actually considered such pool punishment [32]). For punishment, such a primitive institutional incentive mechanism has been shown to evolve and prevail over a peer-based one, provided participation to the public good is optional and non-punishers can be sanctioned [33].

There is growing evidence that pool-reward funds can also lead to fully cooperative societies, even when participation to the collective action is compulsory [26,27]. Rewards have several advantages over punishment, such as not suffering from the threat of retaliation as well as increasing total group earnings [16,17]. In addition, in a society of defectors, rare punishers will suffer enormous costs from punishing group members, whereas investing in pool rewards will yield net benefits when defectors are common. It will, however, become increasingly expensive as cooperators take over, resulting in a second-order dilemma, as non-rewarders would fare better. Therefore, the establishment of cooperation still requires solving second-order free-riding with, for example, second-order sanctions [26]. Furthermore, pool-rewarding mechanisms are based on a particular structure that broadens their general relevance for human social behaviours. While they are usually framed as a prearranged ‘agreed upon’ institution, where decisions to contribute to reward funds are made before the PGG and where the fund is considered to be an excludable good, they can just as well be interpreted from another mathematically similar point of view: two consecutive collective actions with partner choice or coalition formation in-between. Contributors to the PGG choose their cooperative partners to take part in a subsequent joint enterprise. Arguably, such partner choice based on past decisions is an important feature of human social interactions (e.g. friendship and alliances) [34–36], and might also have been an important driving force for cooperation during human evolutionary history [37]. The question that arises now is: why would defectors not invest in a profitable reward fund of their own? This is absent from pool-rewarding models, where rewards can only be given by cooperators for cooperators (i.e. pro-social reward fund) [26,27], or in other words, where only cooperators can group together to participate in another collective action. Yet joint enterprises can also be made by individuals who decided to violate social norms in previous collective action problems (e.g. gangs, mafia or cartels) [38–40].

Here, I show that letting defectors contribute to a reward fund of their own, namely ‘anti-social rewarding’, has dramatic effects for cooperation. It destabilizes cooperation in conditions that previously led populations to cooperative states and leads the dynamics towards either full defection or anti-social rewarding. I then expand the basic model and investigate potential solutions to escape from this social trap with new variants of the pool-rewards PGG. Specifically, I first allow different rewarding efficacies for cooperators and defectors. Second, I investigate a variant where rewarding players (either cooperators or defectors) can exclude to a certain extent their respective non-rewarding type from the reward fund, a countermeasure that has already proved effective for reaching cooperative societies [26]. Finally, I look at a variant where investment into the rewarding fund can only be made with a proportion of the share received from the PGG.

2. Material and methods

Consider an infinitely large, well-mixed population. In each generation, groups of $N \geq 2$ are randomly formed and play a one-shot interaction, consisting of a PGG preceded (or followed, depending on the interpretation; see above) by a pool-rewards mechanism. In the PGG, each member can decide to contribute or not an amount $c_1 > 0$ to a common pot. The total of the pot is then multiplied by a productivity factor $r_1$ (with $1 < r_1 < N$) and equally shared among all group members. By definition, public goods are non-excludable. Hence, even non-contributors will benefit from them. If all contribute, the group’s total payoff is maximized. However, as cooperators receive a share of their own contribution and $r_1 < N$, their payoff is reduced by $c_1 (1 - r_1/N) = \sigma$ from that of an individual who did not contribute. Therefore, individuals are better off not contributing, hence the dilemma.

In the pool-rewarding stage, a focal individual can decide whether or not to contribute an amount $c_2 > 0$ to a reward fund which will be shared exclusively among individuals who made the same decision as the focal in the PGG (i.e. non-contributing players to the PGG can reward other non-contributors, and contributing players can reward other contributing players; the alternative interpretation, as stated above, is that individuals displaying a similar behaviour in the PGG group together to participate in a consecutive social dilemma). In this way, I allow defectors to reward each other, which was lacking in previous models. I refer to the reward fund dedicated to defectors as the anti-social reward fund, and the one dedicated to cooperators as the pro-social reward fund. Each reward fund is multiplied by a productivity factor $r_2$ (with $1 < r_2 < N$) and then distributed equally among the $N_C$ cooperators for the pro-social reward fund, or among the $N_D$ defectors for the anti-social reward fund. Note that even if the productivity factor of the reward funds is smaller than $N$, contributing to rewards will not necessarily be a social dilemma as the number of contributors $N_C$ (or non-contributors $N_D$) for the anti-social reward fund can be smaller than $N$. Here, I will focus on this kind of weakly altruistic rewarding, and not consider ‘strongly altruistic rewarding’. In the latter, rewarding players can only reward their co-players and cannot profit from their own contribution to the reward fund. Sasaki & Unemi [26] have shown that defectors dominate under this condition. In the presence of anti-social rewarding, it is likely that cooperation would evolve under more constrained conditions than those determined in the present paper. In particular, larger exclusion rates $\alpha$ would probably be required for rewarding individuals to suppress second-order free-riding.

I consider the following four strategies: defectors (D) who contribute to neither the PGG nor the anti-social rewarding fund; rewarding defectors (RD) who do not contribute to the PGG but invest in the anti-social reward fund; cooperators (C) who contribute to the PGG but not to the pro-social reward fund; and rewarding cooperators (RC) who contribute to both the PGG and the pro-social reward fund. These strategies are denoted with frequencies $x, y, z, w$ respectively. Thus, $x + y + z + w = 1$. The replicator dynamics is used to model the evolutionary fate of the population [41]. Unless stated otherwise, results are proved analytically in the electronic supplementary material, §5. To test the robustness of the results, I also investigate numerically a large range of parameter values for each
scenario (electronic supplementary material, §5). Graphics are based on the DYNAMO software [42].

3. Results

The state space is represented by the simplex $S_4 = \{(x, y, z, w) \mid x, y, z, w \geq 0, x + y + z + w = 1\}$, where $x, y, z$ and $w$ are the frequencies of defectors, rewarding defectors, cooperators and rewarding cooperators, respectively. The four homogeneous states $D \ (x = 1)$, $R \ (y = 1)$, $C \ (z = 1)$ and $RC \ (w = 1)$ are trivial equilibria. The presence of anti-social rewarding is dramatic for cooperation in the PGG as it will lead the population to uncooperative states such as D or RD under most conditions. Specifically, whenever the optimal group reward $c_3(r_2 - 1)$ cannot outweigh the net cost for a contributor $\sigma$, pro-social rewarders are not able to outcompete defectors (electronic supplementary material, §1). Hence, the system will end up with a population of defectors (electronic supplementary material, figures S1 and S2). In the boundary case that net rewards exactly compensate $\sigma$, there can be a stable mixture of D and RC players (electronic supplementary material, §1), but most of the interior states will end up in the vertex D over the long term (electronic supplementary material, figures S2 and S3). If, however, cooperators can outweigh the net cost of the PGG with their rewards (i.e. $c_3(r_2 - 1) > \sigma$), numerical simulations show that interior states will end up in a rock–scissors–paper cycle around an unstable fixed point, either on the face $D–C–RC$ interior states will end up in a rock–scissors–paper cycle (figure 1; electronic supplementary material, figure S7). Consequently, exclusion no longer leads to fully cooperative societies in the presence of anti-social rewarding.

A second potential solution would be to allow second-order sanctions, as proposed by Sasaki & Unemi [26]. Indeed, it is reasonable to assume that non-rewarding players run the risk of sanctions by their corresponding rewarding peers. Hence, I assume here that non-rewarding players have their share of the reward reduced by a certain proportion $\alpha$ (with $0 \leq \alpha \leq 1$). This reflects situations where contributors to a public good would group their effort in a subsequent joint enterprise that yields an excludable good (e.g. hunting small rodents for their bones). This solution has proved effective in the absence of anti-social rewarding. Provided $\alpha$ reaches a certain threshold, in which case the population will attain a fully cooperative state [26]. Here, I will focus on the case where $\alpha$ is similar for non-contributors and contributors, although it might be plausible that contributors to the PGG would defend their reward more fiercely than others. Interestingly, this solution does not work for counter-anti-social rewarding. In fact, an unstable equilibrium can exist between rewarding individuals and their corresponding non-rewarding type whenever $\alpha > (N - r_2)/(c_3(N - 1))$ (electronic supplementary material, §3). Under this condition, rewarding types are immune to invasion from second-order free-riders. However, non-contributing players to the PGG are still better off. Therefore, the dynamics in the interior of $S_4$ will go towards a population entirely composed of anti-social rewarders, as shown by numerical simulations (figure 1; electronic supplementary material, figure S7). Note that for low values of $\alpha$ and $r_2$, contributors and non-contributors can coexist (electronic supplementary material, figure S7). Consequently, exclusion no longer leads to fully cooperative societies in the presence of anti-social rewarding.

A third potential variant is to render the investment in the reward fund conditional to the public good produced (or in other words to the presence of cooperators). In fact, it is reasonable to assume that a group constituted exclusively of defectors, which will not create any good to consume, will not be able to then participate in subsequent social dilemmas. Consequently, I introduce a new parameter $\beta$ (with $0 \leq \beta \leq 1$), which specifies the proportion of the public good per capita benefit $B$ (with $B = Nc_1/c_2(N - 1)$) that rewarding players are willing to invest in their respective reward fund. As a result, rewarding players will only keep $(1 - \beta)B$ from the public good. I also only focus on the case where $\beta$ is similar for non-contributors and contributors. In the absence of contributors (i.e. C and RC), rewarding defectors RD will not be able to invest in rewards and will get a similar zero payoff to defectors. Consequently, the edge D–RD will be a line of fixed points (electronic supplementary material, §4). For low values of $\beta$ and $r_2$, all interior states will end up on the edge D–RD, which is stable (electronic supplementary material, figures S8 and S9). Pro-social rewarders will increase in frequency with increasing $r_2$. However, I show in the electronic supplementary material, §4 that as $\beta$ crosses a threshold $\beta_1 = (N - r_1)/(c_2(r_2 - 1))$, which is less than or equal to 1 if $r_1r_2 \geq N$ (electronic supplementary material, figure S10) and $r_2$ is above a certain threshold $r_{2T}$, the system will end up in stable polymorphic equilibrium where contributors C and RC coexist with anti-social rewarders RD (electronic supplementary material, figures S9 and S11). Therefore, this variant provides stable, albeit not full, cooperation levels.
In this latter variant, I assumed that the fraction of the PGG share individuals are able to invest in rewards is fixed. However, it is more realistic to assume that individuals can decide themselves how much of the PGG share they are willing to invest. Hence, in the electronic supplementary material, I investigate this with individual-based simulations where \( g \) is modelled as a continuous and evolvable cultural trait (electronic supplementary material, §6). I therefore look at its coevolution with the discrete contribution trait \( \gamma = \{0,1\} \), which determines an individual’s decision to contribute or not to the PGG (note that lower cooperation rates were found if \( \gamma \) is modelled as a continuous trait). In this scenario, although a fully cooperative society could not be reached either, the coevolution of \( \gamma \) and \( \beta \) allows relatively high cooperation levels when both the PGG and reward fund are social dilemmas (i.e. \( r_1 \) and \( r_2 < N \)). Starting from populations of pure defectors, random drift will drive the average \( \beta \) above \( \beta_T \), in which case a single mutant who contributes to the PGG will profit from investing in rewards. This will allow her to invade the population and pave the way for cooperation (electronic supplementary material, figure S12). Once the population is in a fully cooperative state, there is selection for low values of \( \beta \) (i.e. second-order free-riding), but as contributors become less frequent, there will be selection for higher values of \( \beta \) (as less frequent cooperators would profit more from investing their PGG share in rewards). Consequently, the population average \( \beta \) value shows only little variation (electronic supplementary material, figure S13), allowing contributors to resist full invasion from defectors, provided \( r_2 \) is above \( r_{2T} \).

Finally, completely resolving the dilemma relies on the combination of second-order sanctions with either (i) lower rewarding abilities for defectors (figure 2; electronic supplementary material, figures S14–S16) or (ii) reward funds dependent on the PGG’s production (figure 3; electronic supplementary material, figure S17–S19). As before, for second-order sanctions to be favourable for rewarding individuals, the latter must exclude their corresponding non-rewarding type from the reward fund above a critical threshold \( \alpha_r = (N - r_2)/(r_2(N - 1)) \). Then, the condition for non-contributors to fare poorly compared with contributors is either (i) that the anti-social reward fund is less productive (i.e. \( c_2(r_2 - 1) < c_2(r_2 - 1) - \sigma \)) or (ii) if reward funds depend on the PGG’s production, rewarding players must be able to invest more than a certain proportion \( \beta_T = (N - r_1)/(r_1(r_2 - 1)) \) of their PGG share.

### 4. Discussion

The present research demonstrates that positive incentive systems face an important drawback: defectors can take advantage of them to reward each other. The presence of such anti-social rewarding has dramatic effects on cooperation, as it often leads societies to be dominated by free-riders or anti-social rewarders. I predict that completely escaping this social trap relies not only on (i) elimination of second-order free-riding (i.e. refraining from contributing to incentives), but also on (ii) pro-social rewarders outcompeting anti-social rewarders. Exclusion of non-rewarding individuals over a certain threshold helps fulfill the first condition [26], whereas alternatives such as lower rewarding abilities for defectors (e.g. less productive anti-social reward funds) and reward investments dependent on the production of the public good would support the second
condition (although sanctions might have similar effects). Those two conditions are not always found together in our society. Hence, coalitions of norm violators (e.g. gangs, mafia or cartels) still occur as their productivity can often outcompete that of pro-social individuals and sanctions are not always effective. Yet there can be cases where contributors have greater productivities in subsequent joint enterprises because, for example, they gained some experience by either producing the good and/or coordinating their actions [43]. However, in the present model I assumed costless exclusion. While opportunities for costless punishment or exclusion often occur in real life [44], inflicting sanctions might also entail spending resources or energy [8]. Therefore, future work should investigate, in the presence of anti-social rewarding, the effects on cooperation of including not only exclusion costs but also other sorts of punishment inflicted by contributors on wrong-doers.

These results also have implications for the study of partner choice or coalition formation in collective action dilemmas. Indeed, as reward funds are similar to PGGs (i.e. investing is costly, returns are shared among participants), pool-rewards

---

**Figure 2.** The interplay of exclusion and high rewarding efficacies for contributors in the PGG with reward funds. Full cooperation can be attained from any initial condition as rewarding players are able to sufficiently exclude second-order free-riders (i.e. non-rewarding individuals) from their corresponding reward fund (i.e. \( \alpha > \alpha_0 \)), and contributors are able to produce sufficient rewards to outcompete non-contributors (i.e. \( c_3(t_3 - 1) < c_2(t_2 - 1) - c_1 \)). (a) Unstable equilibria are possible between rewarding individuals and their corresponding non-rewarding type (i.e. \( Q_2 \) between D and RD and \( Q_1 \) between C and RC), but random perturbations will eventually lead the system to stable full cooperation. (b) Interior states of S4 also end up in the vertex RC. Parameters: \( N = 5, c_1 = 1, r_1 = 3, c_2 = 1, r_2 = 3, c_3 = 1, r_3 = 2 \) and \( \alpha = 0.5 \). (Online version in colour.)

**Figure 3.** The interplay of exclusion and reward funds dependent on the public good produced. Full cooperation can be attained from any initial condition as rewarding players are able to sufficiently exclude second-order free-riders (i.e. non-rewarding individuals) from their corresponding reward fund (i.e. \( \alpha > \alpha_0 \)), and reward funds are raised with a fraction \( \beta \) of the public good produced. (a) An unstable equilibrium \( Q_2 \) is possible between pro-social rewarders and their corresponding non-rewarding type. The whole edge D–RD is a line of fixed points because, in the absence of contributors, anti-social rewarders cannot invest in their reward fund. Random perturbations such as mutations will eventually lead the system to stable full cooperation. (b) Interior states of S4 also end up in the vertex RC. Parameters: \( N = 5, c_1 = 1, r_1 = 3, r_2 = 3, \beta = 0.5 \) and \( \alpha = 0.2 \). (Online version in colour.)
models reflect situations where individuals play different consecutive social dilemmas and could form coalitions in-between. This study suggests that if cooperators can group together, so can the remaining defectors, which in turn often undermines cooperation or leads to coexistence between cooperators and defectors. Group or alliance formation based on decisions in prior social situations occurs in all kinds of real-life interactions, from within the family and tribes to between nations [45,46], and also happens between norm violators (e.g. criminal enterprises, mafia or cartels) [38–40]. This interpretation is different from current models of pool rewards where they are usually framed as a prearranged institution where decisions to invest into a reward fund are made before and not after the PGG. Both alternatives, however, are mathematically similar. One could argue that an important distinction between those two points of view is that decisions after the PGG could be made conditional on the number of contributors (or defectors for anti-social rewarders) in the first PGG. However, the same argument holds if investing in reward funds was made before the PGG, because decisions to contribute to the PGG could be conditional on the reward fund investment. It would be interesting to investigate the effect of such strategies on the dynamics of cooperation, as individuals investing in rewards only when rare would initially be better off by getting a larger share of the reward fund, but would then quit investing as more and more group members would benefit from their contribution, which would probably lead to a cooperation breakdown.

The present findings demonstrate that a positive incentive system that depends on the collective action’s success, or when a consecutive joint action depends on the public good being created in the first place, could enhance cooperation, at least partially. In fact, it allows stable coexistence between anti-social rewarding and pro-social individuals. Such situations were probably frequent during human evolutionary history. For example, bone tools such as knives, fishhooks or boomerangs served in many collective actions (e.g. hunting, meat sharing, fishing [47,48]). However, a prior dilemma was to hunt animals (e.g. antelopes or rodents) in order to obtain their bones. Only then could individuals go for a larger, more difficult prey, or engage in war with other groups. The present model shows that in such situations, there are two crucial conditions for pro-social rewarders to outcompete defectors. First, every unit produced in the PGG must yield sufficient benefits if invested in rewards (in other words, the product of both the PGG and reward fund multiplication factors $r_1$ and $r_2$, respectively, must be greater than the group size, i.e. $r_1 r_2 > N$). Second, individuals must be able to invest a sufficient proportion of their PGG share in rewards (i.e. $\beta > \beta_2$). I have initially investigated this proportion as a fixed parameter. However, it is more likely that individuals could choose themselves how much they are willing to invest. In fact, the model of coevolution between contribution to the PGG and $\beta$ showed that letting $\beta$ evolve allows individuals to reach the critical threshold $\beta_2$, resulting in considerably more stable, albeit not full, cooperation than in other variants without second-order sanctions.

The present results also highlight the importance of choosing with care the assumptions as well as the types of strategies to include in theoretical models. An important assumption made here is that the public good was a linear function of the individual contributions. A recent study investigated reward funds with nonlinear PGGs where the public good cannot be produced unless a sufficient number of contributors are present in the group [27]. In such ‘threshold public goods games’, reward funds can be effective in escaping the tragedy of the commons, without the help of second-order sanctions. In fact, fully cooperative, yet non-rewarding societies can be achieved provided the public good can only be produced if all group members contribute. However, anti-social rewarding was not allowed in this study. It is likely that its presence would result in these cooperative societies being invaded by anti-social rewarders whenever net anti-social rewards outweigh the loss incurred from not producing the good. However, potential solutions to prevent a cooperation breakdown would be to incorporate second-order sanctions, or even nonlinear returns on investments into reward funds.

A type of individual that could potentially have even more of a dramatic effect on cooperation is one who contributes to the PGG but rewards defectors, namely a ‘cooperating anti-social rewarder’. Such individuals would profit from the pro-social reward fund and reward non-contributors, which could alter the emergence of cooperation. However, these individuals would face an additional second-order dilemma, as those who do not reward defectors would fare better. Thus, it is not clear whether this strategy would evolve. The presence of second-order sanctions would even hinder its evolution. Hence, introducing this strategy (as well as that of ‘defecting pro-social rewarders’) in future work should provide more insight into the evolution of positive incentive systems.

Another example of an initially unconsidered strategy that led to different dynamics is anti-social punishment (i.e. punishment aimed at contributors) [49]. Interestingly, anti-social punishment is still something missing from pool-punishment models so far [29,33]. Allowing this rather intriguing strategy in models of peer punishment drastically alters the dynamics, and often prevents cooperators from dominating [50,51]. However, this deleterious effect depends on the model’s settings. For example, in the presence of loners (i.e. individuals who do not take part in the PGG and prefer a fixed income), cooperation is not hampered if punishers can only punish those who participate in the PGG [52]. Also, if punitive behaviours are linked to reputation, responsible punishment (i.e. directed at defectors) prevails, and cooperation can emerge [19,21]. It is still unknown what the effects on the dynamics of pool punishment would be if anti-social punishment was allowed. However, its evolution would probably have milder effects than anti-social rewarding because, as shown by the present findings, the latter often provide immediate monetary benefits, which is not the case for anti-social punishment.

Institutional punishment and reward funds are powerful means to enhance cooperation in collective action dilemmas. I have shown, however, that the latter is highly vulnerable to anti-social rewarding, as it undermines cooperation under many circumstances. Nonetheless, countermeasures such as lower rewarding abilities for wrong-doers, or rewards that depend on the joint enterprise’s success, can lead to coexistence between defection and cooperation, or even to fully cooperative societies if used in combination with second-order sanctions.

Acknowledgements. I thank S. Dridi, L. Marques da Cunha, S. Powers, N. Raihani, A. Uppal, C. Wedekind and two anonymous reviewers for comments.