



Contents lists available at ScienceDirect

## Journal of Theoretical Biology

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

# Evolution of cooperation and altruistic punishment when retaliation is possible

Marco A. Janssen\*, Clint Bushman

School of Human Evolution and Social Change, Center for the Study of Institutional Diversity, Arizona State University, PO Box 872402, Tempe, AZ 85287-2402, USA

## ARTICLE INFO

### Article history:

Received 31 October 2007

Received in revised form

19 June 2008

Accepted 19 June 2008

### Keywords:

Public good games

Group selection

## ABSTRACT

Altruistic punishment is suggested to explain observed high levels of cooperation among non-kin related humans. However, laboratory experiments as well as ethnographic evidence suggest that people might retaliate if being punished, and that this reduces the level of cooperation. Building on existing models on the evolution of cooperation and altruistic punishment, we explore the consequences of the option of retaliation. We find that cooperation and altruistic punishment does not evolve with larger population levels if the option of retaliation is included.

© 2008 Elsevier Ltd. All rights reserved.

## 1. Introduction

The evolution of cooperation between non-kin in large human groups is puzzling. Several theories have been proposed to explain the evolution of cooperation among humans and other animals. The theory of kin selection explanation on cooperation relate to the theory of kin selection focuses on cooperation among organisms that are genetically closely related (Hamilton, 1964), whereas theories of direct reciprocity focus on the selfish incentives for cooperation in repeated interactions (Trivers, 1971; Axelrod, 1984). The theories of indirect reciprocity and costly signalling show cooperation can emerge in larger groups when the cooperators can build reputation (Alexander, 1987; Nowak and Sigmund, 1998; Lotem et al., 1999; Wedekind and Milinski, 2000; Zahavi, 1977).

Altruistic punishment is suggested to be an important factor in the evolution of cooperation in large human societies. Laboratory experiments on public goods and common pool resources have shown that participants are willing to give up monetary returns to punish noncooperators (Yamagishi, 1986; Ostrom et al., 1992; Fehr and Gächter, 2002). When the option to sanction others was imposed by the experimenters, overuse of the sanctioning option had been identified by Ostrom et al. (1992), but when the subjects could communicate and decide on the level of appropriation and sanctioning they would or would not use, the level of cooperation approached optimality and sanctioning was used primarily on those who broke cooperative agreements.

From an evolutionary perspective, it is puzzling why individuals would accept a reduction of payoff to decrease the payoff of other individuals. The possibility on the evolution of altruistic punishment in human societies has been demonstrated by various studies including Boyd et al. (2003) and Hauert et al. (2007).

Commonly absent amongst these models is reaction to punishment. This contrasts with the ethnographic record where vengeance is not uncommon. The response to punishment is not always compliance. In the US police officers wear bullet proof vests in fear of criminal reprisal. Among the Yanomamo the common response to a punishing raid is to attack the punishers with a return raid (Chagnon, 1988). A more mundane illustration is that workers misuse of company property is connected to injustice both perceived and real on the part of employers (Skarlicki and Folger, 1997) the persistence of retaliation across scales and complexities of societies argues its inclusion in models that include punishment.

Recent laboratory experiments which allow counter punishment similarly do not support the cooperation enhancing effect of punishment alone (Cinyabuguma et al., 2006; Denant-Boemont et al., 2007; Nikiforakis, 2008). In those experiments participants could make an additional punishment decision after they had a first decision round on punishment. They could punish back, which we refer to as retaliation. Note that this is different to spite, which explains unconditional punishers (Nakamaru and Iwasa, 2006). Those individuals may focus on relative earnings within the group, and can lower the earnings of others by punishment (Saijo and Nakamura, 1995).

Since the models on the evolution of altruistic punishment do not include the option of retaliation (Sigmund, 2007), we will explore the consequences of including the option to retaliate.

\* Corresponding author. Tel.: +1480 965 1369.  
E-mail address: [Marco.Janssen@asu.edu](mailto:Marco.Janssen@asu.edu) (M.A. Janssen).

We tested the consequences of retaliation on one of those models: Boyd et al. (2003). This model uses multi-level selection as a mechanism to mimic relevant human evolutionary processes, for which there is recently sufficient evidence as a plausible mechanism in human evolution (Bowles, 2006; Wilson and Wilson, 2007). Other models are based on local interaction (Brandt et al., 2003), or the possibility not to participate in social dilemma (Fowler, 2005; Hauert et al., 2007). Future analysis will address the robustness of our findings to other models on the evolution of altruistic punishment.

## 2. Model description

Our model is an extension of the model described in Boyd et al. (2003).<sup>1</sup> Consider a population which is divided into groups of size  $n$ . An agent can contribute or not to a public good. The probability that an agent contribute is  $p_c$  and is agent-specific. If an agent contributes it will incur a cost  $c$  to produce a total benefit  $b$  that is shared equally among group members (Fig. 1). If an agent does not contribute, but defect, it will incur no costs and produce no benefits. If the fraction of contributors in the group is  $x$ , the expected payoff for contributors is  $bx-c$  and the expected payoff for defectors is  $bx$ , so the payoff disadvantage of the contributors is a constant  $c$  independent of the distribution of types in the population.

After a decision is made to contribute or not, an agent makes a decision to punish each defector in their group or not with probability  $p_p$ . This will reduce each defector's payoff by  $p_1/n$  at a cost  $k_1/n$  to the punisher. If the frequency of punishers is  $y$ , the expected payoffs become  $bx-c$  to contributors,  $bx-p_1y$  to defectors, and  $bx-c-k_1(1-x)$  to punishers.

If the frequency of punishers  $y$  is sufficiently high, the cost of being punished exceeds the cost of cooperating ( $p_1y > c$ ). Hence contribution is more beneficial, especially when a contributor does not invest in punishment. In fact, punishers suffer a fitness disadvantage of  $k_1(1-x)$  compared with nonpunishing contributors. This is the reason punishment is considered to be altruistic and mere contributors are 'second-order free riders.' When there is almost no defection, the payoff disadvantage of punishers relative to contributors approaches zero.

The third decision agents can make is whether they will punish those who have punished them as a way of retaliation. We assume agents know who have punished them, and with probability  $p_r$  a defector who is punished will retaliate. This will reduce the payoff of the original punisher by  $p_2/n$  at a cost of  $k_2/n$  to the retaliator. The fraction of retaliators is denoted as  $z$ . The resulting payoffs will be  $bx-c-k_1(1-x)$  for agents who contributed and punished,  $bx-c$  for agents who only contributed,  $bx-(p_1+k_2)y$  for agents who do not contribute and who retaliated, and finally  $bx-p_1y$  for agents who did not contribute but not retaliated.

Like Boyd et al. (2003) we assume that there are  $N$  groups of population size  $n$ . Each individual decide first whether it will contribute or not using its individual probability value  $p_c$ . Secondly, whether it will penalize defectors, and third whether it will retaliate.

In the next step, individuals encounter other group members with probability  $1-m$  and an individual from another group with probability  $m$ . An individual  $i$  who encounters an individual  $j$  imitates  $j$  with probability  $w_j/(w_j+w_i)$ , where  $w_i$  is the payoff

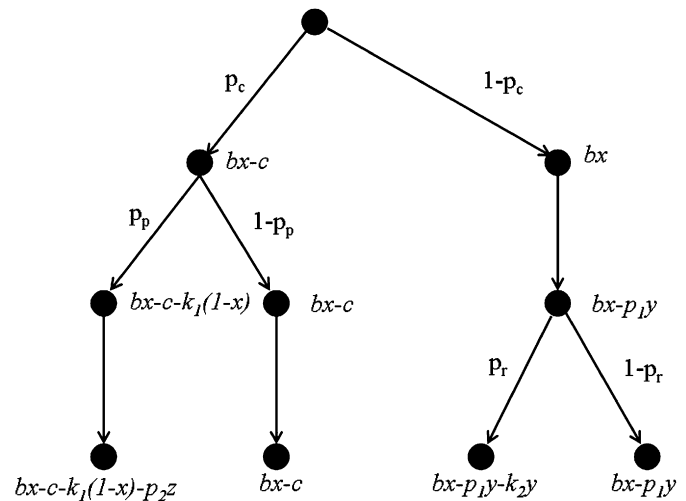


Fig. 1. Decision tree of the agents.

of individual  $l$  in the game, including the costs of any punishment received or delivered. This leads to selection of higher payoff deriving strategies and migration of strategies between groups.

Like Boyd et al. (2003) we assume that group selection occurs through intergroup conflict. In each time period, groups are paired at random, and with probability  $\varepsilon$  intergroup conflict results in one group defeating and replacing the other group. The probability that group  $i$  defeats group  $j$  is  $1/2(1+(W_j-W_i))$ , where  $W_i$  is the average payoff in group  $i$ . In Boyd et al. (2003) the level of cooperation was used to determine the likeliness of winning a conflict, but here we take into account the reduction of payoffs in a group with larger amount of retaliation. This means that the group with agents, who are more likely to defect and/or retaliate, is more likely to lose a conflict. As a consequence, cooperation is the sole target of the resulting group selection process. This is somewhat similar as the recent findings of West et al. (2006) who show that unrelated individuals cooperate when compete for resources at a larger scale they tend to cooperate more compared to interactions on a local scale.

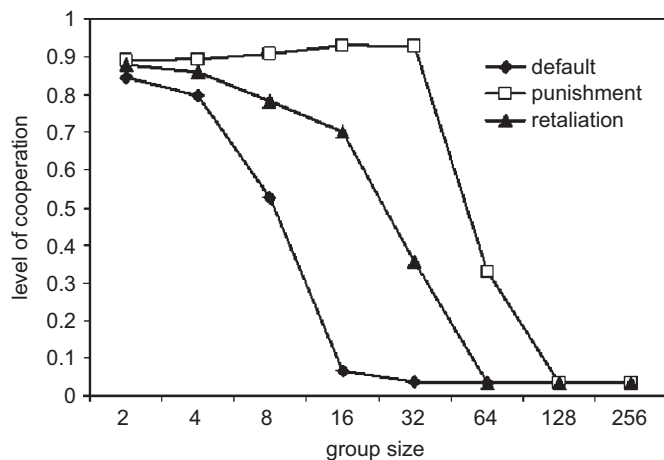
Finally, each generation a white noise  $n(0,\sigma)$  will affect the probabilities that define the strategies of the individuals.

## 3. Analysis

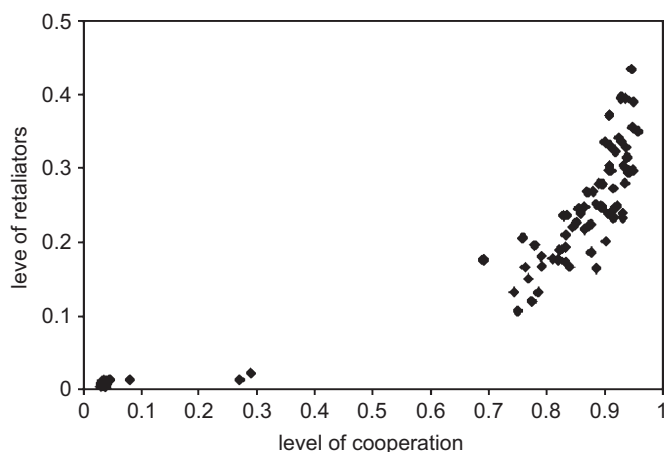
Our main interest is to explore the consequences of the inclusion of retaliation. Like Boyd et al. we run the model for 2000 timesteps and calculate the averages of cooperation, punishment and retaliation for the last 1000 timesteps. Each model run consists of 128 groups. We run the model hundred times for group sizes 2, 4, 8, 16, 32, 64, 128 and 256 agents. We use the following default values  $c = k_1 = k_2 = 0.2$ ;  $p_1 = p_2 = 0.8$ ;  $b = 0.5$ ;  $m = 0.01$ ;  $\varepsilon = 0.015$ ; and  $\sigma = 0.01$  in line with the values used by Boyd et al. (2003).

Fig. 2 shows the results of including the option of retaliation or not. Inclusion of this option has significant effects on the level of cooperation. Including retaliation makes the option of punishment less effective, and as a result the retaliation treatment lies between the default case and the case of punishment only. There is a persistent punishment of retaliation up to group size 32. Beyond group size 32 no cooperation evolves with the default parameter settings, and also the number of cooperators who punish declines. In the case of retaliation, the evolved level of punishment declines.

<sup>1</sup> Boyd et al. (2003) use three pure strategies: cooperators, cooperators who punish and defectors. We use a more general model, where decisions are made probabilistically. Each agent is assumed to have a probability to cooperate and a probability to punish. This formulation reproduces the basic results of the original Boyd et al. (2003) model.



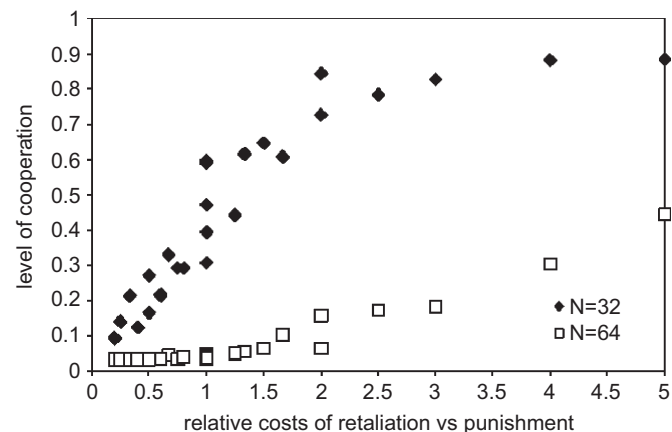
**Fig. 2.** The average level of cooperation for 100 runs for different group sizes  $n$ . Three treatments are distinguished: default (no punishment and no retaliation), punishment (which does not include retaliation) and retaliation (which includes punishment and retaliation).



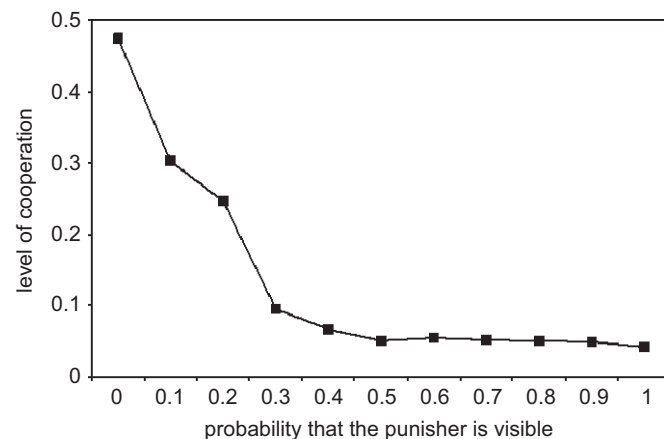
**Fig. 3.** For the case of group size 32 with retaliation we depict the levels of cooperation versus the level of retaliation at the end of the 100 simulation runs.

The average values in Fig. 2 do not provide a complete picture. Fig. 3 shows the results of 100 runs for a group size of 32. In about 60% of the runs the level of cooperation is minimal, and in the other 40% it is relatively high. In fact there are two types of outcomes, and the results in Fig. 2 show the relative amount of high levels of cooperation versus minimal cooperation. In simulations where high levels of cooperation evolve, the level of retaliation is low, since there are not many defectors. The level of retaliation is higher when the evolved level of cooperation is somewhat lower. When no cooperation evolves, the level of retaliation is low again, since there are not many punishers. The diversity of outcomes for group size 32 in Fig. 3 shows that there is a bifurcation of ending up with high or low levels of cooperation.

We also investigated the effect of different levels of costs of punishment,  $k_1$ , and retaliation,  $k_2$ . Using group sizes  $n = 32$  and 64, we systematically varied  $k_1$  and  $k_2$  from values 0.2 to 1.0 with steps of 0.2. We then represented the information as the relative cost of retaliation versus cost of punishment ( $k_2/k_1$ ) (Fig. 4). As would be expected, the level of cooperation increases when we increase the relative cost to retaliate (Fig. 2). Fig. 4 shows that the larger group is less sensitive to the option of retaliation, as the level of cooperation drops steeply with relative low costs of retaliation for group size 32.



**Fig. 4.** Level of cooperation for different values of  $k_1$  and  $k_2$  (average of 100 runs for each parameter combination), represented as  $k_2/k_1$ .



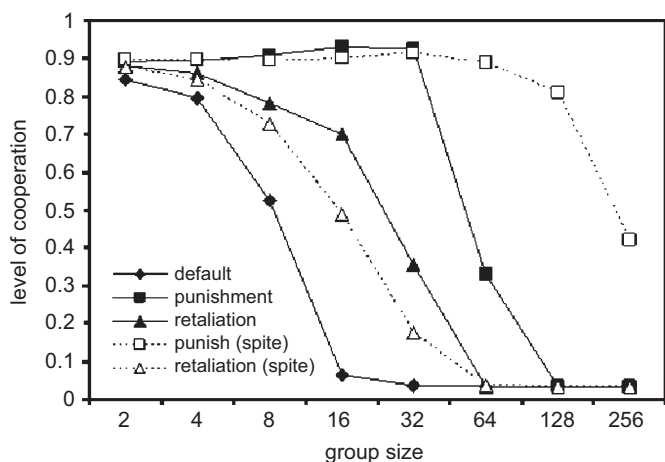
**Fig. 5.** Level of cooperation, punishment and retaliation for different probabilities of visibility of punishers in the situation of group size 32 and retaliation.

### 3.1. Sensitivity to visibility

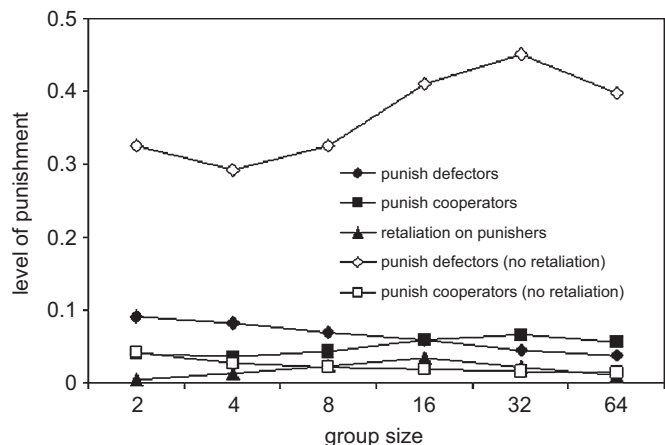
Visibility of the punisher might affect the level of cooperation. If punishers can be identified, and therefore agents know whom to retaliate, the level of punishment may be low. In empirical cases punishers are not always known. Rule breakers among Maine lobster fishermen may experience that their lines are cut, but they do not know who has done this (Acheson, 1988). Therefore we may expect an effect of the visibility on the level of cooperation. We included a probability  $p_v$  to determine whether the agent who is punished know the identity of the punisher or not. We show a typical result in Fig. 5 for the case of group size 32. As shown in Fig. 2, inclusion of retaliation leads to a very low level of cooperation. Fig. 5 shows that the visibility or traceability of the punisher needs to be below 30% before we experience a significant increase of cooperation.

### 3.2. Sensitivity to spite

Low contributors have observed to be punished in various public good experiments (Cinyabuguma et al., 2006). Falk et al. (2005) analysed possible causes of this phenomenon and found that when the cost of punishment for the punisher is equal to the penalty the punished player receives, no perverse punishment is observed again. In most experiments the cost of punishment to the punisher is smaller than the cost to the punished one, which may trigger spiteful



**Fig. 6.** The average level of cooperation for 100 runs for different group sizes  $n$ . Same as Fig. 2 but including two additional treatments: punishment with defectors who unconditionally punish, and retaliation with defectors who unconditionally punish.



**Fig. 7.** The level of punishment cooperators, defectors, and punishers experience. Average level of 100 runs for different group sizes.

behaviour. This spiteful behaviour is considered to be a defector who is punishing. Both defectors and cooperators are punished by spiteful punishers. Compared to retaliation, spiteful agents are defectors who unconditionally punish. We analysed the sensitivity of our results by including the option of defectors to punish a defector with probability  $p_{pd}$  and a cooperator with probability  $p_{pc}$ . Fig. 6 shows that our main conclusion remains the same: retaliation leads to a lower level of cooperation. Including spiteful behaviour without retaliation leads to high levels of cooperation since in the beginning of the simulation more defectors are punished (spiteful punishers also punish defectors) and with an increase of cooperation spiteful behaviour almost disappears. In Fig. 7 we see that for the condition where we have cooperators who punish, punished agents who retaliate, and defectors who punish, we have high levels of punishment of defectors in small groups. With larger groups, about 16 agents, the level of punishment of cooperators and defectors become equal and the level of retaliation is significant. With increasing size of the group we see a decline of cooperation as well as punishment and retaliation. This shows that retaliation of agents of being punished makes that the evolved level of agents who punish defectors decrease. Fig. 7 also shows the level of punishment of defectors when retaliation is not possible, and this level is much higher and explains the higher level of cooperation in Fig. 6.

#### 4. Discussion

Our analysis shows that inclusion the option of agents to retaliate punishers lead to lower levels of altruistic punishment and cooperation, especially when the cost of retaliation is at a similar or higher level as the cost of punishment. For the problem of human cooperation altruistic punishment is not a panacea. In order to explain observed high levels of cooperation in human society, we need to explore alternative mechanisms to explain how retaliation and punishment can be used effectively. The relative cost of retaliation versus punishment affect the use of retaliation and therefore the level of cooperation. Especially large groups are relative sensitive to a relative drop in the cost of retaliation. If an agent start punishing in larger group defectors, it might be hit hard by those defectors who retaliated.

Our results are robust to the possibility of spiteful behaviour which we define as defectors who unconditionally punish other agents. This observed phenomenon in experiments does not preclude the evolution of conditional punishment of defectors (retaliation) and the reduction of cooperation for larger groups if retaliation is possible.

Another possible mechanism we explored is the visibility of the punisher. There are situations where the strategy of retaliation is limited since it is not evident who punished. For example, in certain lobster fisheries in Maine, the cutting of the lines of non-compliant fisherman is used to enforce norms. This method is anonymous and cheap making retaliation comparatively more expensive (Acheson, 1988). Local conditions may affect the traceability of the punisher as well as the relative costs of retaliation versus punishment. Therefore, altruistic punishment as an explanation to self-organized institutions that overcome is dependent on a specific number of local conditions of the group.

#### Acknowledgement

We appreciate the support of the National Science Foundation for the Grant BCS-0432894.

#### References

- Acheson, J.M., 1988. The Lobster Gangs of Maine. New England Universities Press, Hanover, NH.
- Alexander, R.D., 1987. The Biology of Moral Systems. Aldine de Gruyter, New York.
- Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, New York.
- Bowles, S., 2006. Group competition, reproductive leveling, and the evolution of human altruism. *Science* 314, 1569–1572.
- Boyd, R., Gintis, H., Bowles, S., Richerson, P.J., 2003. The evolution of altruistic punishment. *Proc. Natl. Acad. Sci. USA* 100 (6), 3531–3535.
- Brandt, H., Hauert, C., Sigmund, K., 2003. Punishment and reputation in spatial public goods games. *Proc. R. Soc. London Ser. B–Biol. Sci.* 270, 1099–1104.
- Chagnon, N., 1988. Life histories, blood revenge, and warfare in a tribal population. *Science* 239, 985–992.
- Cinyabuguma, M., Page, T., Putterman, L., 2006. Can second-order punishment deter perverse punishment? *Exp. Econ.* 9 (3), 265–279.
- Denant-Boemont, L., Masclet, D., Noussair, C., 2007. Anonymity in punishment, revenge and cooperation: a public good experiment. *Econ. Theory* 33 (1), 145–167.
- Falk, A., Fehr, E., Fischbacher, U., 2005. Driving forces behind informal sanctions. *Econometrica* 73 (6), 2017–2030.
- Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. *Nature* 415, 137–140.
- Fowler, J.H., 2005. Altruistic punishment and the origin of cooperation. *Proc. Natl. Acad. Sci. USA* 102, 7047–7049.
- Hamilton, W.D., 1964. Genetical evolution of social behavior I and II. *J. Theor. Biol.* 7, 1–52.
- Hauert, C., Traulsen, A., Brandt, H., Nowak, M.A., Sigmund, K., 2007. Via freedom to coercion: the emergence of costly punishment. *Science* 316, 1905–1907.
- Lotem, A., Fishman, M.A., Stone, L., 1999. Evolution of cooperation between individuals. *Nature* 400, 226–227.
- Nakamaru, M., Iwasa, Y., 2006. The coevolution of altruism and punishment: role of the selfish punisher. *J. Theor. Biol.* 240 (3), 475–488.
- Nikiforakis, N., 2008. Punishment and counter-punishment in public good games: can we really govern ourselves? *J. Public Econ.* 92 (1–2), 91–112.

- Nowak, M.A., Sigmund, K., 1998. Evolution of indirect reciprocity by image scoring. *Nature* 393, 573–577.
- Ostrom, E., Walker, J., Gardner, R., 1992. Covenants with and without a sword: self-governance is possible. *Am. Pol. Sci. Rev.* 86 (2), 404–417.
- Saijo, T., Nakamura, H., 1995. The spite dilemma in voluntary contribution mechanism experiments. *J. Conflict Resol.* 39, 535–560.
- Sigmund, K., 2007. Punish or perish? Retaliation and collaboration and humans. *Trend Ecol. Evol.* 22 (11), 593–600.
- Skarlicki, D.P., Folger, R., 1997. Retaliation in the workplace: the roles of distributive, procedural and interactional justice. *J. Appl. Psychol.* 82, 434–443.
- Trivers, R., 1971. The evolution of reciprocal altruism. *Q. Rev. Biol.* 46, 35–57.
- Wedekind, C., Milinski, M., 2000. Cooperation through image scoring in humans. *Science* 288, 850–852.
- West, S.A., Gardner, A., Shuker, D.M., Reynolds, T., Burton-Chellow, M., Sykes, E.M., Guinnee, M.A., Griffin, A.S., 2006. Cooperation and the scale of competition in humans. *Curr. Biol.* 16, 1103–1106.
- Wilson, D.S., Wilson, E.O., 2007. Rethinking the theoretical foundation of socio-biology. *Q. Rev. Biol.* 82 (4), 327–348.
- Yamagishi, T., 1986. The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.* 51 (1), 110–116.
- Zahavi, A., 1977. The cost of honesty. *J. Theor. Biol.* 67, 603–605.