NBER WORKING PAPER SERIES

THE HIRED GUN MECHANISM

James Andreoni
Laura K. Gee

Working Paper 17032
http://www.nber.org/papers/w17032

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2011

We would like to thank our colleagues at UCSD, participants in the Social Dilemmas Conference, participants at the ESA 2010 Conference, participants at the SWET 2011, Kristy Buzzard, Simon Gachter, Nikos Nikiforakis and Anya Savikhina for helpful comments during the writing of this paper. Andreoni also gratefully acknowledges the National Science Foundation, grant 1024683, for financial support. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

The Hired Gun Mechanism
James Andreoni and Laura K. Gee
NBER Working Paper No. 17032
May 2011
JEL No. C72,C91,C92,D7,H41

ABSTRACT

We present and experimentally test a mechanism that provides a simple, natural, low cost, and realistic solution to the problem of compliance with socially determined efficient actions, such as contributing to a public good. We note that small self-governing organizations often place enforcement in the hands of an appointed leader–the department chair, the building superintendent, the team captain. This hired gun, we show, need only punish the least compliant group member, and then only punish this person enough so that the person would have rather been the second least compliant. We show experimentally this mechanism, despite having very small penalties out of equilibrium, reaches the full compliance equilibrium almost instantly.

James Andreoni
Department of Economics
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0508
and NBER
andreoni@ucsd.edu

Laura K. Gee
Department of Economics
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0508
l1gee@ucsd.edu

An online appendix is available at:
http://www.nber.org/data-appendix/w17032

1 Introduction

The problems of externalities and public goods are among the oldest and most widely studied market failures in economics. Numerous authors have proposed mechanisms to implement Pareto efficient allocations in these circumstances. These mechanisms, and their difficulties, are well known. Getting full revelation of preferences in a dominant strategy mechanism is difficult or impossible, and Nash equilibrium mechanisms are often not individually rational. Moreover, such mechanisms can be complex, abstract, and difficult to implement. (For a review, see Myerson (2008).)

Recently economists have been turning to a more practical formulation of this problem. Imagine a small group of individuals who have a clear notion of what an efficient allocation of costs or efforts would be. For instance, groups can easily agree to equal cost sharing for a public good, or cost sharing for maintaining a commons that is proportional to use. Communities adopt norms of proper communal behavior which limit or forbid externalities on neighbors or other group members. How might a group create a governance scheme that could be low cost to implement, easy to understand, result in sustained (second-best if it is costly) efficient allocations, and suffer only small costs of sanctions off the equilibrium path? See, also, Ostrom, Walker and Gardner (1992).

A special problem faced by small self-governing groups, such as neighborhoods, clubs, and academic departments is that punishments meted out by members, as suggested by Fehr and Gachter (2000), can often be quite deleterious. Peers often punish to the extent that they erase any gains brought on by the punishment resulting in net losses in the short run (see Egas and Riedl, 2008; Gachter, Renner and Sefton, 2008; Herrmann, Thoni and Gachter, 2008; Botelho, Harrison, Pinto and Rutstrom, 2007; Fehr and Gachter, 2002). And, since individuals apparently enjoy the act of punishing (Fudenberg and Pathak, 2010), this can result in counter-punishment and revenge cycles (Nikiforakis and Engelmann, 2011; Nikiforakis, 2008; Denant-Boemont, Masclet and Noussair, 2007) that further deplete the social surplus.

In contrast to this, we propose a mechanism that is based on a single simple rule of punishment of noncompliance. The rule is easy to understand, low cost to enforce, in equilibrium results in no punishments and full compliance, and when off the equilibrium path typically results in punishments that will be small. These are all advantages of the mechanism. However, to avoid revenge and retribution, the rule must be implemented in a way that isolates the enforcer. This means appointing or hiring someone from either within or outside the group to enforce the rule. For this reason we call our device the Hired Gun Mechanism.

A central feature of the Hired Gun is that the enforcer does not need to perfectly document all the noncompliance. The Hired Gun, only needs to know the exact actions of the two largest deviators from compliance. In many instances the biggest deviators (think of loudest neighbors, worst teachers, most truant volunteers) are easy to identify. Moreover, the enforcer does not need to punish all non-compliant people, just the single biggest cheater. Finally, the punishment need not be large. It only needs to be just big enough that the most non-compliant person would rather have been the second most non-compliant person. If the second most non-compliant person is best responding to his or her environment, the two most non-compliant choices should be nearly identical, meaning that in expectation this difference should be trivial. It follows that punishments off the equilibrium path will likely be small. Hence, even if our mechanism requires some experience to reach equilibrium, the costs along that road should be minimal.

Our intuition for the Hired Gun comes from two sources. First is simple observation of real life mechanisms. Speeding tickets are not generally issued to everyone on the freeway going over the speed limit, but rather tend to be assigned to the fastest car on the road. To avoid a speeding ticket, one only needs to be the second fastest car. Likewise, the noisiest neighbor will get the call from the superintendent, the neighborhood association will address the most egregious code violations first, and the department chair will deal with the worst teachers before going to others, the club will ostracize the member with the worst service record. That is, enforcement of compliance in the real world often focusses first, and often exclusively, on the most egregious violators.

The second source of intuition is from the Keynesian p-beauty contest games (Ho, Camerer and Weigelt, 1998; Nagel, 1995). Imagine a game in which the winner of a prize is the person who guesses a number between 0 and 100 that is closest to two-thirds of the average of the others' guesses. As long as there is common knowledge of rationality, people will realize that(through iterated deletion of dominated strategies) the only way for everyone to be two-thirds of the average is if they all guess 0, which is the Nash equilibrium. Our mechanism turns this intuition upside down. Here the loser will be the one who gained the most by deviating from full compliance, and the penalty will be enough to make them wish they had been the second biggest cheater rather than the biggest. The only equilibrium in which everyone can avoid being the biggest cheater (again with common knowledge of rationality) is full compliance.

Others have also discussed worlds in which punishment is focussed on the largest free rider. Boyd, Gintis and Bowles (2010) and Sigmund, Silva, Traulsen and Hauert (2010) present evolutionary models in which punishing free riders is "coordinated" and punishment has "increasing returns," which combine to show social forces will evolve to coordinate on the largest free riders. Steiner (2007) models a world similar to that of Fehr and Gachter (2002) in which he assumes punishment is motivated by (potentially small amounts of) anger, and that anger is directed at only the worst free rider. The study most similar to ours is by Yamagishi (1986). His mechanism differs from ours in that individuals first play a public goods game, then make voluntary contributions to a punishment mechanism. Here we assume individuals procure a hired gun before the contributions stage, hence institutionalizing the governance ex-ante. Although we do not discuss it in this paper, as we show in a companion paper (Andreoni and Gee, 2011), hiring a gun, for a small enough cost, can be welfare improving by allowing individuals to socially commit to a sub-game perfect enforcement regime. The companion paper shows that individuals in our experiment easily recognize this and endogenously acquire the delegated punishment scheme. The current paper provides the groundwork for this by explaining and testing the efficacy of the Gun for Hire mechanism.

In what follows we present a formal statement of the Hired Gun Mechanism and provide an experimental test. The experiment reveals both the power of the mechanism and, in one session, the reliance on (common knowledge of) rationality.

2 The Mechanism

We derive and test our game within the context of a linear public goods game (LPG). The LPG game is the workhorse of experiments on free riding and externalities. The linearity makes the ideas easier to present, however the model is easily demonstrated in a nonlinear environment.

In the LPG, each individual i is endowed with w_i. Each allocates g_i, $0 \leq g_i \leq w_i$, to the public good and consumes the rest, $w_i - g_i$. Each person receives one util for each unit of w_i kept, and α for each unit allocated to the public good by themselves and all others. So i's payoff is $\pi_i = w_i - g_i + \alpha \sum_{j=1}^{n} g_j$. Assuming, as is usual, $0 \leq \alpha < 1$ and $n\alpha > 1$ then the unique sub-game perfect Nash equilibrium of this game is total free riding, $g_i = 0$ by all players, while the socially agreed Pareto efficient allocation is complete contribution, $g_i = w_i$. Thus, full compliance with the socially proscribed optimum is for $g_i = w_i$ for all.[1]

We assume for our mechanism that members of the community commonly understand the socially desirable action of each individual, that is, they all agree that $w_i = g_i$ is socially desirable, and that deviation from this action is undesirable. The w_i could represent equal cost sharing for instance, or perhaps a proportional tax, or simply providing no negative externalities on others in the group. For these purposes, it becomes convenient to write our model in terms of the *deviations* from the socially desirable Pareto efficient allocation. Thus, define

$$d_i = w_i - g_i$$

as i's deviation from his assigned contribution. Define $d = (d_1, d_2, ..., d_n)$ as the vector of deviations from full compliance. Then payoffs to i can be written in terms of d_i rather than g_i :

$$\begin{aligned} \pi_i(d) &= d_i + \alpha \sum_{j=1}^{n} (w_j - d_j) \\ &= d_i - \alpha \sum_{j=1}^{n} d_j + \alpha W \end{aligned}$$

where $W = \sum_{i=1}^{n} w_i$. In this formulation, $d_i = 0$ for all i is clearly the socially desired outcome.

2.1 Size of Shot Fired by the Hired Gun

Define d_z as the *largest* element of the vector d, that is, $d_z \geq d_i$ for all i. Let S be the set of potential contributors and $L(d) \subseteq S$ be the set of contributors with the largest deviations. Finally, define d_y as the *second* largest deviation. In particular if $L(d) \subset S$, then d_y is the highest d_i for all i in $S \backslash L(d)$. However, if $L(d) = S$, then $d_i = d_j$ for all i and j and there is no second highest deviator.

In the Hired Gun Mechanism, the largest deviator(s) are the only one(s) punished and the punishment is just sufficient to make the biggest deviator strictly prefer to have been the *second* biggest deviator. Let $P(d_i, d)$ be the punishment required by the mechanism for an individual who chooses d_i when the vector of choices is d. Then the mechanism requires punishments such that $\pi_z(d_z, d) - P(d_z, d) < \pi_y(d_y, d)$, that is, the biggest deviator would rather have been the second biggest deviator. Assume for convenience, and realism, that d_i can only be chosen in discrete amounts, and for simplicity assume it takes integer values. Then, to assure punishments are as small

[1] More generally, let $g = g^*$ be any among the set of possible Pareto Effecient allocations, where g_i^* is perhaps chosen by some focal ideal, such as equal division ($g_i^* = g_j^*$, all i, j) or equal sacrifice ($g_i^*/w_i = g_j^*/w_j$, all i, j).

as possible, set the punishment on the biggest deviator to be $P(d_z, d) = \pi_z(d_z, d) - \pi_y(d_y, d) + 1 = d_z - d_y + 1$. Note, if there are ties for the biggest deviator, then all must be punished. In this case, punishment need only be high enough to break indifference, that is the smallest integer unit of the private good, thus $P = 1$. The exception is if all tie by being fully compliant. In this case, of course, there are no punishments.

In general, therefore, define punishment as follows:

$$P(d_i, d) = \begin{cases} 1 & \text{if} \quad L(d) = S, \quad \text{and} \quad d \neq 0 \\ 0 & \text{if} \quad L(d) = S, \quad \text{and} \quad d = 0 \\ d_z - d_y + 1 & \text{if} \quad L(d) \subset S, \quad \text{and} \quad i \in L(d) \\ 0 & \text{if} \quad L(d) \subset S, \quad \text{and} \quad i \notin L(d) \end{cases}$$

2.2 Equilibrium

It is easy to see that $d = 0$ is the unique equilibrium. Proof follows from elimination of dominated strategies. Let m' be the highest standard of compliance set for anyone (this allows for m' to be the same for all i). Then any $d_i = m'$ is sure to be the lowest compliance level and i is sure to receive a punishment of at least 1. As such d_i will never equal m'. In this case, anyone with compliance level m' will act as if their true compliance level is $m' - 1$. But if m' is never chosen, then $m' - 1$ is sure to be punished and anyone choosing it would be better off choosing $m' - 2$. Thus $d_i = m' - 1$ will never be chosen by any i. We can repeat this logic, eliminating dominated strategies until the only choice that is not eliminated is that $d_i = 0$ for all i. This demonstrates that $d = 0$ is the unique equilibrium.

3 Experimental Design and Results

Our experimental test of the Hired Gun Mechanism is a within subject comparison of behavior in the Linear Public Goods game (LPG) to the Hired Gun Mechanism. Each session of the experiment employs 12 subjects divided into three groups of four subjects. In each session, subjects first play 10 periods of a LPG with strangers matching (Andreoni, 1988). The same subjects then play another 10 periods, again with strangers matching, but now are subject to the Hired Gun Mechanism. The Hired Gun Mechanism was not explained to the subjects until period 10 was complete.

We chose this design for two reasons. First, we wanted to present our mechanism with a realistic and rugged terrain of a group that had become accustomed to significant degrees of free riding, hence creating the most challenging environment for the Hired Gun mechanism. Second, if the Hired Gun can turn free riders into cooperators, then this is the strongest kind of support one can get for the mechanism.

We conducted 4 sessions with 12 subjects in each session ($N = 48$). Subjects were paid for a single randomly selected period. The experiment was conducted in neutral language using z-tree software Fischbacher (2007). In each period subjects were endowed with 5 tokens, $w = 5$. Each token invested in the private good paid a return of \$3 to the individual who invested it, while each token invested in the public good paid a return of \$2 to all group members. In terms of the game above, $\alpha = 2/3$ and $n = 4$. If all subjects free ride, they earn \$15, while if they all contribute w they earn \$40 each.

In the Hired Gun game (periods 11-20) the player with the lowest contribution in the group is punished the difference between the first and second lowest contributors LPG payoff plus the value of one unit of the private good, which was worth \$3.[2] Full instructions and screen shots are available from the authors.[3]

3.1 Results

Figure 1, rounds 1 to 10, shows the results of the LPG game. The usual pattern emerges, with contributions at or near the Nash equilibrium of total free riding for many subjects by period 10. Figure 1, rounds 11 to 20, shows that employing the Hired Gun Mechanism immediately pushed the subjects to almost full contribution in periods 11-20. The LPG game reached only 27 percent of the efficient allocation, while with the Hired Gun efficiency averaged 91.2%. This difference is statistically significant (Kolmogorov-Smirnov Test, $p = 0.037$).[4]

Figure 1: Average Public Contribution Per Subject by Session

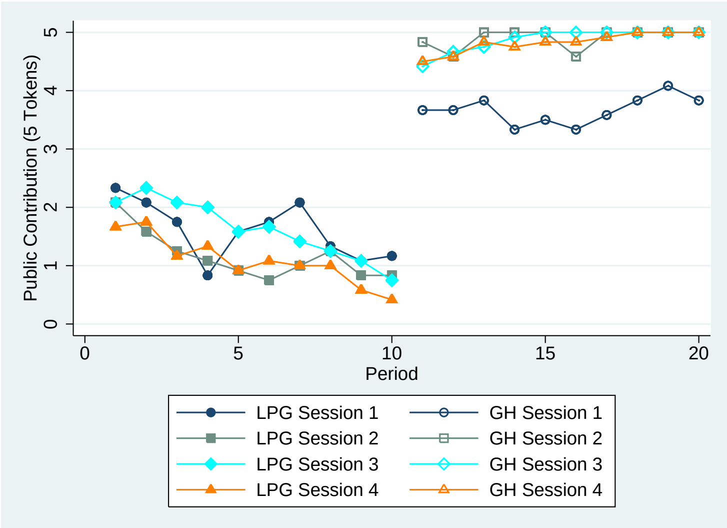


Figure 2 shows that net earnings were significantly improved with the Hired Gun. Average earnings were \$21.84 in the LPG periods, but jumped to \$36.77 (net of punishment costs) when there was a Hired Gun, again this difference is statistically significant (Kolmogorov-Smirnov Test $p = 0.037$ by session).

[2]In addition to the payoffs discussed here subjects in the experiment were also given a small lump sum payment each period. These were in place to make income comparable to other treatments discussed in our companion paper, Andreoni and Gee (2011). Since these do not alter the predictions, for expository convenience we do not discuss them here. In the LPG game subjects were given a lump-sum payment of \$1 per period, while in the GH game subjects were given a \$0.50 lump sum payment. These lump sum payments are not included in the discussion of earnings in this paper.

[3]`econ.ucsd.edu/~jandreon` or `econ.ucsd.edu/~l1gee`

[4]We treat each session average as an observation, so we have 4 observations for the LPG and 4 observations for the periods 11-20 of GH.

Figure 2: Average Earnings Per Subject by Session

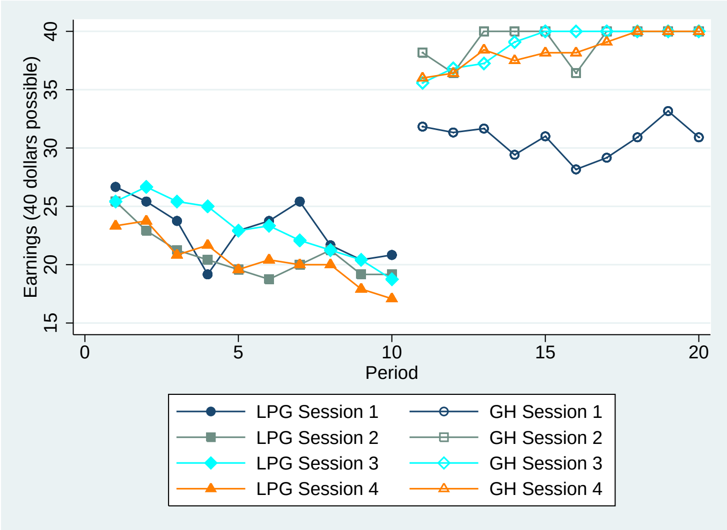
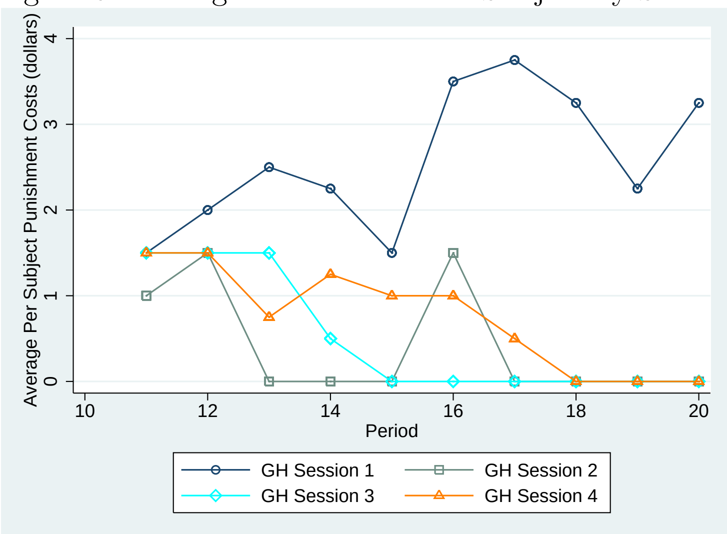


Figure 3 shows the average punishment costs per subject across rounds 11-20 were indeed very small. This is especially true for sessions 2, 3, and 4 where these costs drop to zero for periods 18 to 20.

Figure 3: Average Punishment Per Subject by Session



3.2 What was Different about Session 1?

One can see quite plainly in the figures above that Session 1 stands out for being less cooperative under the Hired Gun than the other three sessions. Why did this happen? One reason may simply be that it was the first session that we ran, and so we as experimenters had not honed our presentation of the stimulus. However, it also appears that the selection of subjects into Session 1 included an unusually high proportion of what we will call unimproving players.

Say that an action by a player is *unimproving* if the player was punished in the previous period, but she does not raise her next public contribution above her previous period's contribution. Then define an *unimproving player* as one who takes two or more unimproving actions during periods 12-20. In Session 1, one third of the subjects (4 of 12) were unimproving subjects, whereas in Sessions 2 and 3 there were no unimproving subjects and session 4 had only a single unimproving subject among the 12. Although the Hired Gun Mechanism still resulted in an improvement to public contributions and earnings in Session 1, it was not as pronounced as in sessions 2, 3 and 4. This result shows that the Hired Gun Mechanism is effective, but that its full potency is dependent upon the lack of unimproving actions of the players. Punishments are only effective if people want to avoid them, and realize how to change their actions to do so. Our Hired Gun Mechanism will not fare well if subjects either do not want to avoid punishment or do not understand how to avoid punishment. In other words, both the assumptions of common knowledge of rationality and rationality itself may have been violated in Session 1.

Although session 1 failed on one respect, it also showed an important advantage of the Hired Gun Mechanism. Even in the presence of subjects who refused to conform to the equilibrium, punishments stayed rather small—about 8% of gross earnings, and 25% of the increase in earnings relative to the LPG game. That is, equilibrium responses of the more rational players to the stubbornly self-destructive actions of the unimproving players assures that when things go wrong, they are not catastrophic.

4 Conclusion

We theoretically developed a simple mechanism for enforcing full compliance with a socially agreed standard of cooperation, which we call the Hired Gun Mechanism. The intuition for the Hired Gun is that, first, in the world around us enforcement focusses primarily if not exclusively on the worst violators; other violators go unpunished. Second, using the analogy of a p-beauty contest, if we punish the largest deviator so that the person would rather have been the second biggest deviator, the only point at which everyone can be the second biggest cheater is if everyone is fully compliant. Thus the mechanism is inexpensive to employ, since only the biggest deviants need to be monitored, and optimizing behavior implies that the path toward the equilibrium will have minimal punishments. Finally, since the punisher is a hired or appointed authority following a publicly accepted rule, we can potentially avoid the cycles of retribution and revenge that plague peer punishment.

We experimentally test our mechanism within subjects by first giving them 10 rounds of experience, with strangers matching, of a standard linear public goods game without any enforcement mechanism. As usual, by round 10 each session is rife with free riding. In rounds 11-20 we then introduce the Hired Gun Mechanism. The mechanism instantly raises contributions. In three of our four sessions, all 12 subjects converged to the equilibrium of full compliance, and in all sessions the payoffs under the Hired Gun were significantly above the earnings without the mechanism. In all sessions, punishments were small and did not have an appreciable impact on the efficiency gains of the mechanism.

The Hired Gun Mechanism has many ecologically attractive features. First, it is simple. As shown in the lab, most subjects catch on to the incentives in a few short rounds. Second, it is low cost to enforce. This is because the enforcer need only identify the two largest violators, and does not have to precisely monitor all players. In life, the most egregious violators are typically also the easiest to identify and with our mechanism there is no need to distinguish fine differences among any more than the top two deviators. Third, by delegating enforcement to the Hired Gun, the mechanism avoids problems of retribution and revenge. In this way it is a more realistic model of actual self-governing mechanisms where authority figures (building superintendents, department chairs, team captains, club presidents) enforce simple rules of compliance. Fourth, the mechanism results in all players being better off in equilibrium, thus is individually rational. Fifth, before reaching the equilibrium, punishments remain small. This is because the best reply of rational players to a player who stubbornly (and mistakenly) chooses to be noncompliant each period, is to be only slightly more compliant than this person, thus naturally putting limits on how costly punishments can be, but without erasing the information value of being punished which is necessary for people to learn to avoid it. That is, the "slap on the wrist" may be all that is needed to push society steadily and sustainably to Pareto efficient compliance. And sixth, the mechanism works quickly to reach efficiency and, once reached, efficiency is sustained. It is worth noting as well, that all of these effects were accomplished in an experiment that provided a hostile environment for the mechanism. Despite starting from a situation with rampant free riding, and being burdened with random rematching that made reputations nearly impossible to build, three of our four sessions exceeded 95% efficiency after fewer than three trials with the mechanism, and reached complete efficiency within several trials.

In sum, the Gun for Hire Mechanism is simple, individually rational, ecologically valid, and works nearly instantly in an experimental test to improve efficiency.

References

Andreoni, James, "Why free ride? : Strategies and learning in public goods experiments," *Journal of Public Economics*, 1988, *37* (3), 291 – 304.

_ **and Laura Gee**, "Governing Ourselves: Does Third Party Enforcement Crowd out Peer Punishment?," 2011.

Botelho, A., G.W. Harrison, L.M.C. Pinto, and E.E. Rutstrom, "Social Norms and Social Choice," *Working Paper*, 2007.

Boyd, Robert, Herbet Gintis, and Samuel Bowles, "Coordinated Punishment of Defectors Sustains Cooperation and Can Proliferate When Rare," *Science*, 2010.

Denant-Boemont, Laurent, David Masclet, and Charles N. Noussair, "Punishment, Counterpunishment, and Sanction Enforcement in a Social Dilemma Experiment," *Economic Theory*, 2007, *33* (1).

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

Fehr, Ernst and Simon Gachter, "Cooperation and Punishment in Public Goods Experiments," *The American Economic Review*, 2000, *90* (4), 980–994.

_ **and** _ , "Altruistic punishment in humans," *Nature*, 2002.

Fischbacher, Urs, "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 2007, *10* (2).

Fudenberg, Drew and Parag A. Pathak, "Unobserved punishment supports cooperation," *Journal of Public Economics*, 2010, *94* (1-2), 78 – 86.

Gachter, Simon, Elke Renner, and Martin Sefton, "The Long-Run Benefits of Punishment," *Science*, 2008, *322* (5907), 1510–.

Herrmann, Benedikt, Christian Thoni, and Simon Gachter, "Antisocial Punishment Across Societies," *Science*, 2008, *319* (5868), 1362–1367.

Ho, Teck-Hua, Colin Camerer, and Keith Weigelt, "Iterated Dominance and Iterated Best Response in Experimental "p-Beauty Contests"," *The American Economic Review*, 1998, *88* (4), pp. 947–969.

Myerson, Roger B., "Mechanism design," *The New Palgrave Dictionary of Economics*, 2008.

Nagel, Rosemarie, "Unraveling in Guessing Games: An Experimental Study," *The American Economic Review*, 1995, *85* (5), 1313–1326.

Nikiforakis, Nikos, "Punishment and counter-punishment in public good games: Can we really govern ourselves?," *Journal of Public Economics*, 2008, *92* (1-2), 91 – 112.

_ **and Dirk Engelmann**, "Altruistic punishment and the threat of feuds," *Journal of Economic Behavior and Organization*, 2011, *In Press, Corrected Proof*, –.

Ostrom, Elinor, James Walker, and Roy Gardner, "Covenants With and Without a Sword: Self-Governance is Possible," *The American Political Science Review*, 1992, *86* (2), 404–417.

Sigmund, Karl, Hannelore De Silva, Arne Traulsen, and Christoph Hauert, "Social learning promotes insistitutions for governing commons," *Nature*, 2010.

Steiner, Jakub, "A Trace of Anger is Enough: on the Enforcement of Social Norms," *Economics Bulletin*, 2007, *8*, 1–4.

Yamagishi, Toshio, "The Provision of a Sanctioning System as a Public Good," *Journal of Personality and Social Psychology*, 1986, *51* (1), 110 – 116.