Extended Abstract:

Not my job: automatic and interpersonal punishment in a public goods game in rural Uganda

Sehomi L. Fanou,* Ian Levely,* and Marrit van den Berg*

Well-functioning institutions that sanction free-riders can lead to the more efficient provision of public goods. However, evidence shows that when financial incentives are introduced, this can crowd out non-material incentives for cooperative behaviour (Bowles and Polania-Reyes, 2012). In extreme cases, this can make sanctions counter-productive. Since many individuals are willing to pay a cost to punish free riders, peer-to-peer punishment is a second way of maintaining cooperation in a public good environment. In this study, we use an lab-in-the-field experiment that consists of a standard public goods game with punishment, with an additional feature: an automatic punishment mechanism that imposes a fine on subjects whose contributions fall below a certain threshold. Our design models a scenario in which a government imposes a sanction on law-breakers, but does so imperfectly. Does the presence of the automatic sanction affect individuals' willingness to punish free riders? How does this affect cooperation? This dynamic is of particular importance for understanding how weak or sporadically enforced formal rules interact with informal institutions in developing countries. Our subjects live in rural villages in Uganda in which public goods and common-pool resources are often managed informally, with varying degrees of regulation from formal institutions. We find that certain individuals refrain from punishing group members in the experiment when the automatic sanction is present. We find a correlation between those whose punishing behaviour is affected by treatment and attitudes towards formal authorities outside of the lab. Interestingly, this effect persists even when the automatic sanction is not present. These findings have implications for understanding how incomplete formal institutions affect informal cooperation.

Our set-up resembles the public goods game with punishment after Fehr and Gächter (2000), who show that allowing subjects to punish one another after observing contributions can drastically increase cooperation (Fehr et al., 2002; Gächter and Herrmann, 2009; Chaudhuri, 2011). Subsequent studies have shown that when given the choice, subjects prefer having the option to punish peers, and are able to reach more efficient outcomes (Gürerk et al., 2006). There is also evidence on the benefits of automatic punishment in public goods provision, though interestingly, when given a choice, peer-to-peer punishment is more popular and effective than formal sanction schemes (Markussen et al., 2014; Kamei

*Development Economics Group, Wageningen University. Email: sehomi.fanou@wur.nl; ian.levely@wur.nl

et al., 2015). None of these studies have focused on the interaction between formal and informal sanctions.

On the other hand, a number of studies show that imposing financial sanctions to enforce cooperation can crowd out cooperation. Bowles and Polania-Reyes (2012) provide a theoretical framework for this phenomenon of "state-dependent preferences," and empirical studies demonstrate how sanctioning can lead to worse outcomes if it crowds out moral incentives for cooperative behavior (Titmuss, 1970; Gneezy and Rustichini, 2000; Fehr and Rockenbach, 2003; Fehr and List, 2004; Falk and Kosfeld, 2006). Furthermore, several studies have tested how punishment and rewards do not always complement cooperation in public good games (Irelenbusch and Ruchala, 2008; Nikiforakis, 2008a; Cardenas and Carpenter, 2008). For instance, allowing for counter-punishment deters interpersonal punishment and can lead to feuds (Nikiforakis, 2008b; Nikiforakis and Engelmann, 2011). Thus, peer-to-peer punishment outside of laboratory experiments might not always be so efficient.

A common feature of these studies is that they focus on a first-order public goods. However altruistic punishment, costly to the individual, but beneficial to the group can be considered a "2nd-order public good". The provision of this 2nd-order public good produces the same cooperation dilemma as a 1st-order public goods, and relies on prosocial preferences. Hence, if automatic sanctioning crowds out preferences for altruistic punishment, then formal rules could have a detrimental effect on societies. This has been seldom studied in the literature. Understanding this dynamic is interesting from a theoretical standpoint but relevant for institutional transitions especially when enforcement is imperfect or when former rules govern only limited aspects of social interaction.

Design

We use a public goods environment with peer-to-peer punishment and introduce an automatic punishment regime, where subjects are sanctioned by the experimenter if they contribute less than a specified amount. Subjects were matched in anonymous groups of four for the duration of the experiment and play three tasks within each treatment arm. In Task 1 subjects in all treatments played a standard public goods game without punishment. In each round, subjects were given an endowment $\omega_i = 10$ experimental units. Contributions to the public good, c_i, were doubled and evenly distributed among the four group members. Hence the marginal per capita ratio is equal to 0.5.

In Task 2, subjects in all treatments again played the public goods game, but were given the option of punishing group members in each round, after observing contributions. To do so, they were granted an extra endowment of three experimental units, which they could either keep or use to punish group members. For each experimental unit that i spent on punishment, i's payoff decreased by 1 unit, while a selected group member j's payoff decreased by 3.

Across treatments, we varied whether low contributers were automatically punished by the experimenter. In the *automatic punishment* (AI) treatment, if an individual contributed $c_i < 8$ experimental units, her payoff was automatically decreased by 3 units. This was common knowledge, and subjects could observe who received automatic punishment before making their interpersonal punishment decisions. Thus, the payoff for each subject in the automatic punishment treatment is:

$$\pi_i = \overbrace{\omega_i - c_i + 0.5\sum_{i=1}^{N} c_i}^{\text{Contribution}} \quad \overbrace{-3(\mathbf{1}\{c_i < 8\})}^{\text{Automatic punishment}} \quad \overbrace{+(3 - \sum_{j=1}^{N} p_{ij}) - \sum_{j=1}^{N} 3p_{ji}}^{\text{Inter-personal punishment}}$$

Where p_{ij} is the number of punishment points allocated by i to j, and p_{ji} the number of punishment points received by i from j. The automatic punishment, in effect, allocates an additional punishment point to any subject whose contribution falls below the threshold ($c_i = 8$).

In contrast, the *interpersonal punishment only* (IO) treatment serves as a control, and does not include that automatic punishment mechanism.[1] The treatments and timing are described in Table 1. Comparing the AI and IO in Task 2 allows us to observe whether the automatic punishment mechanism crowds out interpersonal punishment, and whether this has an effect on contributions.

In Task 3, subjects in the AI and IO treatments played another 4 rounds, this time with no automatic punishment mechanism in either treatment (i.e. as in the Interpersonal treatment in Task2). This allows us to study how experience with the automatic punishment institution affects subsequent punishment behaviour, in the absence of the institution.

We include two additional treatments which measure whether automatic punishment crowds out interpersonal punishment and how it affects behaviour in the absence of the institution. In Task 2, in the *High-* and *Low-detection* treatments, subjects who contributed less than 8 experimental units received an automatic deduction, but only 10% and 90% of the time, respectively. In these treatments, each subject first decided how much to contribute to the public good. Next, contributions from each group member were communicated. Then, using the strategy method, subjects decided how much to punish each group member in two cases: i) if shirkers would be detected and face an automatic deduction of three experimental units and ii) if shirkers would not be detected, and receive no automatic deduction. In case that round was chosen for payment, a drawing with the specified probability was conducted, and the corresponding decision was used to calculate payoffs.[2]

[1] Since the AI treatment strongly communicates in norm for expected contributions, we mitigate this potential confound in the IO treatment by calling subjects attention to contributions of less than eight experimental units with a neutral marker.

[2] Subjects in the *High* and *Low* detection treatments did not complete Task 3, though subjects were told that they would complete 3 tasks to keep expectations across treatments consistent, and played one round of the standard public goods game in order to avoid deception.

Procedures

The experiment was conducted in rural communities in Wakiso, a peri-urban and rural district in the central region of Uganda. The participants are a representative sample of individuals aged between 18 and 65 years. We ran 52 sessions with 16 subjects each which lasted around four hours. Subjects were matched in anonymous groups of four for the duration of the experiment. Each task consisted of four rounds, with one round from on task randomly chosen for payment at the end of the experiment. The protocol has been adapted to illiterate subjects, and the instructions used neutral language (i.e. " deduction" rather than "punishment"). At the end of each session, we randomly selected a round for payment and administered a short survey. Average earnings from the experiment was 5.18 Euro, compared to the median daily income of our subjects pool, 1.63 Euro.

Results

To begin, we compare the results between the AI and IO treatments. In Task 2, across the 4 rounds, the average contribution in the Automatic punishment treatment was 6.37 experimental units, compared to 5.95 units in the Interpersonal punishment only treatment (see figure 1). In contrast to standard assumptions about enforcement rules, we find no significant difference in contributions to the public good account (p=0.44).[3]

However, in Figure 2, we do see a difference in how this cooperation is maintained. When the automatic punishment mechanism is present, we see a drop in the amount of peer-to-peer punishment. In the IO treatment in Task 2, subjects allocated an average of 0.77 punishment points in each round, compared to only 0.48 punishment points on average in the AI treatment (p=0.00). This indicates that subjects treat the automatic punishment as a substitute for costly, peer-to-peer punishment. In Table 2, we confirm this result by regressing punishment in Task 2 on treatment, controlling for Task 1 results. We also test for whether automatic punishment*exactly* displaces interpersonal punishment, and find that it does not.[4] While automatic punishment lowers interpersonal punishment, it increases overall punishment (i.e. peer-to-peer plus automatic). This is consistent with no change in preferences, since in effect, there is a relaxed budget constraint in the AI treatment.[5]

Further analysis shows that this change is at the extensive margin: in Task 2 in the IO treatment, 1.12 subjects per group refrained from any punishment across the 4 rounds, compared to 1.98 in the AI treatment (p=0.00).

[3] All p-values are from Wilcoxon rank-sum tests, with one observation per group.

[4] If this is the case, then the sum of the parameter estimates of *Contribution*≤ 3 and *Aut. Pun*Contribu*≤ 3 in column 2 of table 2 should equal to one-third of a punishment point. We reject this hypothesis $p = 0.01$ and reach similar conclusion for contributions level that are between four and seven, $p = 0.00$.

[5] We confirm that the difference in punishment across treatments is not a result of differences in contributions by using a Oaxaca decomposition, in which we include the contribution level of each group member and treatment, as in column 2 of Table 2, we find that only the interaction between treatment and low contributions is statistically significant (p=0.02), while contribution on its own is not (p=0.610).

Next, we examine how experience with the automatic sanctioning regime affects behaviour in Task 3 (see table 2, column 3 & 4 and Figure 2), when all treatments played the same public goods game, with no automatic punishment. We find that while there is again no difference in contribution levels between the AI and IO treatments—5.54 and 6.71 units on average, respectively (p=0.72)—the gap in punishing behaviour that we observe in Task 2 remains. In the AI treatment, subjects allocated 0.38 punishment points on average, compared to 0.65 in the IO treatment (p=0.01). In other words, subjects who are accustomed to punishing less in Task 2, due to the presence of the automatic punishment mechanism, continue to punish less. Interestingly, they are able to maintain a similar level of cooperation despite this.

In Task 2 we do not find treatment difference in the rate of cooperation or punishment behaviour between the *low-* and *high-detection* treatments. However, when compared to the AI and IO treatments, we again find that the automatic sanction affects punishing behaviour. In the Low- and High- detection treatments, when shirkers escaped detection, this resembles the Task 2 decision faced by subjects in the *interpersonal* treatment. Comparing levels of punishment between these conditions allows us to assess how propensity to punish is affected by the fact that low contributors *could have* been punished automatically, but were not. We find that this matters, but only in the *high-detection* treatment, in which subjects assigned fewer punishment points than in the IO (p=0.02). Conversely, we examine the conditional decisions in which shirkers were detected in the Low- and High-detection treatments and compare this with the Automatic punishment treatment. This tells us how punishment behaviour differs if low contributors, *could have* escaped detection, but were punished. We find that the propensity to punish is higher than in the Automatic punishment treatment: we observe 0.21 (p=0.06) and 0.14 (p=0.16) additional punishment points allocated in the *low-* and *high-* detection treatments, respectively (see Figure 2).

We test for heterogeneous treatment effects to assess how subjects' personal characteristics affect their behaviour in the Automatic and Interpersonal punishment treatments. The results are presented in Table 3. We could observed that female subjects are less likely to engage in interpersonal punishment in the Automatic punishment treatment. Additionally subjects that report high trust are more not affected by treatment. Interestingly subjects that resort to formal institutions for the resolutions of problems related to the violation of wetlands use, or dispute with community member are less likely to punish when the formal sanction mechanism is in place.

Together, these findings suggest that institutions that impose automatic punishment on individuals for failing to contribute to a public good do indeed affect willingness to punish peers. However we do not find that automatic punishment lowers the overall amount of punishment that low contributors receive. Interestingly, there is a long-term effect of automatic punishment on interpersonal punishment, yet, the level of contributions to the public good remain equal in spite of the lower punishment levels. This effect is driven by certain individuals, who also prefer formal institutions for managing conflict and social dilemmas outside the experiment.

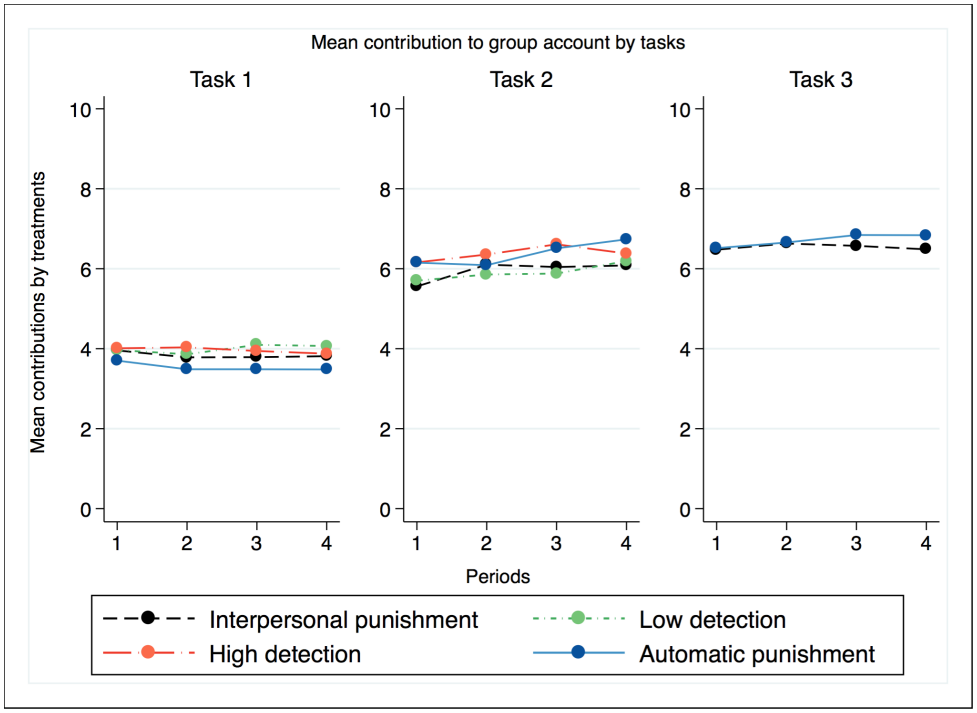

Figure 1: Time path of mean contribution to public good across treatment and task.

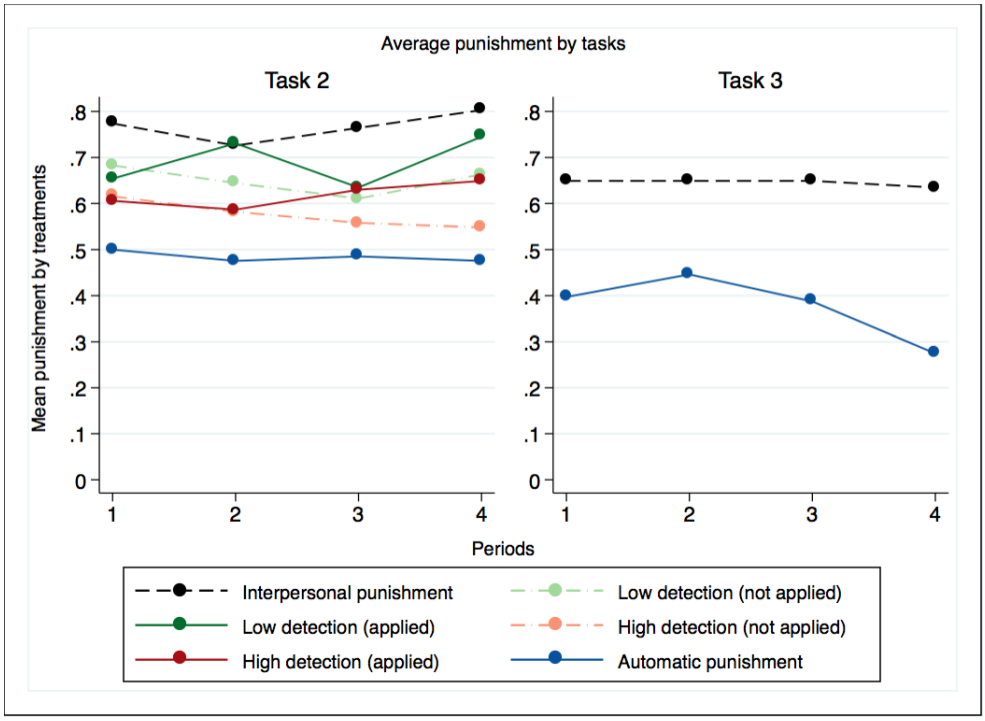

Figure 2: Time path of mean number of (peer-to-peer) punishment points assigned per subject across treatment and task.

Table 1: Experimental treatments

	Tasks		
Treatment	**1**	**2**	**3**
Interpersonal only (IO)	No pun.	Interpersonal only	Interpersonal only
Automatic + interpersonal (AI)	No pun.	Interpersonal & automatic	Interpersonal only
High detection (HD)	No pun.	Interpersonal & automatic (90%)	–
Low detection (LD)	No pun.	Interpersonal & automatic (10%)	–

Table 2: Treatment effects on interpersonal punishment: Task 2 & 3

	Task2		*Task 3*	
Sample	*Perfect detection + Interpersonal only*			
Dependent variable	**Punishment points allocated**			
	(1)	(2)	(3)	(4)
Automatic punishment treatment	-0.09***	0.03	-0.09***	-0.02
	(0.03)	(0.02)	(0.03)	(0.03)
Contribution<=3		0.46***		0.48***
		(0.04)		(0.08)
Contribution 4-7		0.27***		0.28***
		(0.03)		(0.05)
Aut. Pun*Contrib<=3		-0.20***		-0.17*
		(0.06)		(0.11)
Aut. Pun*Contrib 4-7		-0.20***		-0.14**
		(0.04)		(0.06)
Constant	0.25***	-0.03	0.24***	0.02
	(0.04)	(0.05)	(0.05)	(0.05)
Observations	4,944	4,944	4,944	4,944
Number of groups	103	103	103	103

Note: *** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. Nested random effects model at group level and individual levels. All specifications include controls for average group contribution in round 1; columns 1-4 include a control for round number; and a constant term.

Table 3: Heterogeneity analysis: Task 2

Sample	Automatic and interpersonal punishment treatments						
Dependent var.	Punishment points allocated						
		High score on trust index			Formal dispute resolution		
Personal control	Female	Outgroup	Ingroup	Courts + police	Wetland use violators	Problem w/ community member	High group index
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Task 2						
AI treat.	-0.04	-0.12***	-0.13***	-0.10***	-0.02	-0.05	-0.09***
	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)
Personal control	0.03	-0.05	-0.05	-0.04	0.09**	0.08**	-0.03
	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)
AI*Control	-0.09**	0.17***	0.09*	0.02	-0.13**	-0.10**	-0.05
	(0.04)	(0.06)	(0.05)	(0.04)	(0.05)	(0.05)	(0.05)
Observations	4,944	4,944	4,944	4,944	4,944	4,944	4,944
Number of groups	103	103	103	103	103	103	103

Note: *** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. Nested random effects model at group level and individual levels. All specifications include controls for average group contribution in round 1; control for round number;

References

Bowles, S., Polania-Reyes, S., 2012. Economic incentives and social preferences: substitutes or complements? Journal of Economic Literature 50(2), 368–425.

Cardenas, J. C., Carpenter, J., 2008. Behavioural Development Economics : Lessons from Field Labs in the Developing World. Journal of Development Studies 44(3), 37–41.

Chaudhuri, A., 2011. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. Experimental Economics 14(1), 47–83.

Falk, A., Kosfeld, M., 2006. The Hidden Costs of Control. American Economic Review 96(5), 1611–1630.

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

Fehr, E., Gächter, S., 2000. Cooperation and Punishment in Public Goods Experiments. American Economic Review 90(4), 980–994.

Fehr, E., List, J. A., 2004. The Hidden Costs and Returns of Incentives - Trust and Trustworthiness among CEOs. Journal of the European Economic Association 2(5), 743–771.

Fehr, E., Rockenbach, B., 2003. Detrimental Effects of Sanctions on Human Altruism. Nature 422(6928), 137–140.

Gächter, S., Herrmann, B., 2009. Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. Philosophical Transactions of the Royal Society of London B: Biological Sciences 364(1518), 791–806.

Gneezy, U., Rustichini, A., 2000. Pay enough or don't pay at all. The Quarterly Journal of Economics 115(3), 791–810.

Gürerk, Ö., Irlenbusch, B., Rockenbach, B., 2006. The competitive advantage of sanctioning institutions. Science 312(5770), 108–111.

Irelenbusch, B., Ruchala, G. K., 2008. Relative rewards within team-based compensation. Labour economics 15(2), 141–167.

Kamei, K., Putterman, L., Tyran, J.-R., 2015. State or nature? endogenous formal versus informal sanctions in the voluntary provision of public goods. Experimental Economics 18(1), 38–65.

Markussen, T., Putterman, L., Tyran, J.-R., 2014. Self-organization for collective action: An experimental study of voting on sanction regimes. Review of Economic Studies 81(1), 301–324.

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

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

Nikiforakis, N., Engelmann, D., 2011. Altruistic punishment and the threat of feuds. Journal of Economic Behavior & Organization 78(3), 319–332.

Titmuss, R., 1970. The gift relationship: from human blood to social policy. Allen & Unwin, London. .