Distributed Punishment as a Norm-Signalling Tool

(Extended Abstract)

Daniel Villatoro
IIIA - CSIC
Bellatera, Barcelona, Spain
dvillatoro@iiia.csic.es

Giulia Andrighetto
ISTC - CNR
Rome, Italy
giulia.andrighetto@istc.cnr.it

Jordi Brandts
Business Econ. Dept. (UAB)
IAE - CSIC
jordi.brandts@iae.csic.es

Jordi Sabater-Mir
IIIA - CSIC
Bellatera, Barcelona, Spain
jsabater@iiia.csic.es

Rosaria Conte
ISTC - CNR
Rome, Italy
rosaria.conte@istc.cnr.it

ABSTRACT

Recent studies show that punishment plays a crucial role in favoring and maintaining social order. However, very little attention has been paid so far to the potential of distributed punishment. In this work we put forward the hypothesis that distributed punishment is more effective than individual punishment, because is more persuasive in conveying messages of peer condemnation and of shared norm defense. We test this hypothesis by presenting results from laboratory experiments with humans and virtual agents and from simulations solely with agents.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent Systems

General Terms

Experimentation

Keywords

Incentives for Cooperation, Normative systems, Social simulation, Modeling cognition and socio-cultural behavior

1. INTRODUCTION

Theoretical and laboratory studies indicate that cooperation and the maintenance of social order typically requires a punishment threat, as the temptation to cheat, free-ride and violate norms is always strong for autonomous agents [4, 5].

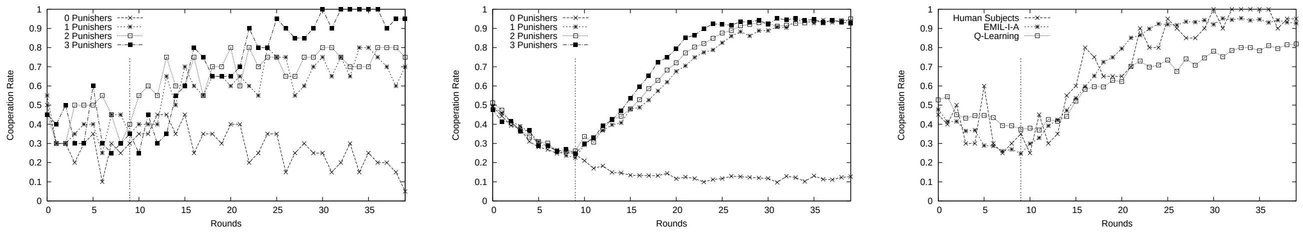
With few exceptions [3], punishment has been usually modeled as (a) a *material* damage, i.e. a cost inflicted to the target, by (b) a *single* agent, that (c) sustains *alone* all the costs of the punishing action (including those consequent to possible retaliations) [2]. On the contrary, ethnographic evidence shows that punishment is often *distributed*, i.e. performed by many, which share the costs of acting, and includes gossip and other forms of explicit or implicit communication. In this work, we focus on the potential of distributed punishment in promoting compliant conduct. With distributed punishment we refer to the practice that occurs when a number n of agents, where $n > 1$, inflicts the target a material damage, such that each punisher sustains a share of the punishment cost. In particular, we suggest that when distributed, punishment works as a *norm-signalling* tool and we put forward the hypothesis that distributed punishment may boost cooperation more than individual one because it is more effective in expressing cooperation norms, as it is more likely to be interpreted as a sanction (for an analysis of the differences between punishment and sanction [6]).

We present cross-methodological evidence supporting our hypothesis: a laboratory experiment with human subjects where we compare the respective effects of individual versus distributed punishment; and an agent-based simulation that allowed us to properly explore the power of "moral suasion" of distributed versus individual punishment

To test the viability of distributed punishment in achieving and maintaining cooperation, we conducted a laboratory experiment reproducing a social dilemma situation. In particular, participants (divided in groups of 4) played a *public goods game* in which they had to decide whether to invest or not their private endowment in a group fund. Payoffs are such that it is individually rational to abstain from investing in the group fund, yet the pro-social group best strategy would be investing in the group fund because this yields a bonus. After having decided whether to contribute or not to the group fund, participants have the possibility to punish. What is special to our set-up is that each group of 4 was composed of one human subject and three confederate virtual players. Human subjects were not informed of the fact that they were playing with confederate virtual players. The reason for putting each human subject in a group with three confederate virtual players is to be able to observe humans in a completely controlled situation.

The experiment consists of four treatments, which differ with respect to the number of the punishing subjects: (1) no punishment, (2) the subject is punished by *one* peer, (3) the subject is punished by *two* peers, (4) the subject is punished by *three* peers. The material damage imposed

Appears in: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (eds.), 4-8 June 2012, Valencia, Spain.



(a) Human Cooperation Rates (b) EMIL-I-A Cooperation Rates (c) Behavioural Comparison

Figure 1: Laboratory and Simulation Results.

on the punished agent in treatments 2, 3 and 4 is *identical* (i.e. it reduces the payoffs of the punished subject to zero) and the way the experiment has been implemented prevents the occurrence of reputational effects, as participants cannot identify one another. Thus, the material and symbolic incentives imposed in treatments 2, 3 and 4 are the same.

In Figure 1(a), the average cooperation rates obtained in the four treatments are shown. Only the behavior of human subjects is plotted. After run 10 the four treatments are actived and it is possible to observe their relative effects on the cooperation level. In the no punishment condition, the cooperation level rapidly collapses. On the contrary, in the three punishing treatments the cooperation level increases with respect to the first 10 rounds and is higher than the one obtained in the no punishment treatment. It is interesting to notice that being punished by three group members (i.e. 3 punishers treatment) leads to a higher cooperation level than when they are punished by two or just one subjects.

As the *same* material damage is imposed in all the treatments (except for the *0 Punishment Treatment* one), we hypothesize that the explanation for the difference on the cooperation rates has to be found in additional information that the punished players receive. We suggest that the higher the number of punishers, the less likely the observers will interpret their behaviors as dictated by the self-interest and, conversely, the more likely they will attribute the punishment to impersonal, possibly normative and legitimate reasons. In other words, our hypothesis is that distributed punishment is more likely to be interpreted as a norm-defending act than individual punishment, thus conveying a strong normative message of peer condemnation.

To test this hypothesis, we designed cognitively complex agents able to interpret as normative the social information they are exposed to and to include it into their decision-making. The agent architecture used for such task is EMIL-I-A [1, 6]. We then replicated the experiment conducted in the laboratory through agent-based simulation.

It is interesting to notice that the cooperation dynamics achieved in the simulation experiment with EMIL-I-As (see Figure 1(b)) are very similar to the ones obtained in the experiment with human subjects (see Figure 1(a)). However, the difference in the cooperation levels observed in the three punishment treatments in the laboratory experiment (with a higher level of cooperation when 3 punishers acted simultaneously) is stronger than the one achieved with EMIL-I-A agents). A possible explanation for this difference is that humans in addition to be sensitive to the fact that three punishers acted together, are also influenced by the fact that is the group as a whole that reacts against his conduct. This additional information is not taken into account by EMIL-I-As. Finally, we conducted a simulation experiment in which the game is played by *Reinforcement Learning* agents, not endowed with normative reasoning and driven only by utilitarian motivations. In Figure 1(c) the cooperation levels obtained in the 3 punishers treatment by human subjects, normative agents and Reinforcement Learning agents are confronted. Data show that Reinforcement Learning agents obtain cooperation levels similar to humans, confirming that the utilitarian motivation in humans is very strong, although the cooperation rates are not as high as the ones obtained by humans and EMIL-I-As.

In this study, we have provided some experimental evidence to show the viability of distributed punishment in promoting cooperation. Distributed punishment is shown to be a powerful tool through which messages of peer condemnation and of shared norm defense are conveyed. These data provide support for the hypothesis that punishment is effective in regulating people's behavior not only through the imposition of a material damage, but also thanks to the normative information it conveys and the normative requests it asks people.

Acknowledgments

This work was supported by the Spanish Education and Science Ministry [EVE project, TIN2009-14702-C02-01]; the Generalitat de Catalunya [2009-SGR-1434]; the Complexity-Net Project SEMIRA.

2. REFERENCES

[1] G. Andrighetto and D. Villatoro. Beyond the carrot and stick approach to enforcement: An agent-based model. In *European Perspectives on Cognitive Science*. New Bulgarian University Press, 2011.

[2] R. Boyd, H. Gintis, and S. Bowles. Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, 328(5978):617–620, 2010.

[3] R. Ellickson. *Order without Law : How Neighbors Settle Disputes*. Harvard University Press, June 2005.

[4] E. Fehr and S. Gachter. Altruistic punishment in humans. *Nature*, 415:137–140, 2002.

[5] E. Ostrom, J. Walker, and R. Gardner. Covenants with and without a sword: Self-governance is possible. *Am. Pol. Sci. Review*, 86(2):404–417, 1992.

[6] D. Villatoro, G. Andrighetto, R. Conte, and J. Sabater-Mir. Dynamic sanctioning for robust and cost-efficient norm compliance. In *Proceedings of the 22nd IJCAI*, 2011.