Games and Economic Behavior 77 (2013) 271–283

Contents lists available at SciVerse ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Legitimate punishment, feedback, and the enforcement of cooperation

Marco Faillo [a], Daniela Grieco [b], Luca Zarri [b,*]

[a] *Department of Economics, University of Trento, Via Inama 5, 38122 Trento, Italy*
[b] *Department of Economics, University of Verona, Via dell' Artigliere 19, 37129 Verona, Italy*

ARTICLE INFO

Article history:
Received 6 December 2011
Available online 1 November 2012

JEL classification:
C73
C91
D02
D63

Keywords:
Public goods games
Cooperation
Legitimate punishment
Feedback
Behavioral mechanism design

ABSTRACT

In dealing with peer punishment as a cooperation enforcement device, laboratory studies have typically concentrated on discretionary sanctioning, allowing players to castigate each other arbitrarily. By contrast, in real life punishments are often meted out only insofar as punishers are entitled to punish and punishees deserve to be punished. We provide an experimental test for this 'legitimate punishment' institution and show that it yields substantial benefits to cooperation and efficiency gains, compared to a classic, 'vigilante justice' institution. We also focus on the role of feedback and we interestingly find that removing the information over high contributors' choices is sufficient to generate a dramatic decline in cooperation rates and earnings. This interaction result implies that providing feedback over virtuous behavior in the group is necessary to make a legitimate punishment scheme effective.

1. Introduction

In naturally occurring environments, punishment is a widespread phenomenon. The ubiquity of sanctioning is due to a significant degree to its importance for the proper functioning of society at large as well as the efficiency of small-scale organizations and communities. Since Fehr and Gächter's (2000) seminal paper, experimental economists have typically concentrated on a cooperation enforcement device such as *discretionary* peer-to-peer sanctioning, where players are free to castigate each other arbitrarily. By contrast, since in many environments meting out punishments implies inflicting (sometimes extremely severe) costs on the punishees, a defining feature of real-life sanctioning mechanisms is that their usage is far from being unrestricted. In modern societies, punishment is usually viewed as socially and ethically acceptable only insofar as it is 'legitimate'. *Legitimate punishment* means that specific requirements have to be met for a person or an institution to be viewed as a potential punisher as well as a potential punishee.[1]

Everyday life abounds in situations where sanctioning systems have to be legitimate in order to be implemented. In the case of centralized punishment, only some institutions (for instance, police and courts) are entitled to impose sanctions on wrongdoers. Trade unions and employers' associations usually set up arbitration boards entitled to monitor and enforce the compliance of their members. However, it is important to note that legitimate punishment often takes the form of

* Corresponding author.
E-mail addresses: marco.faillo@unitn.it (M. Faillo), daniela.grieco@univr.it (D. Grieco), luca.zarri@univr.it (L. Zarri).

[1] In the Western world, centuries of normative argument in applied ethics, philosophy of law and political philosophy (with classical contributions from prominent thinkers such as John Stuart Mill and, more recently, John Rawls, Jurgen Habermas and Ronald Dworkin, among many others) have convincingly made clear that in a liberal democracy punishment needs to be legitimate, in order to be theoretically justified. In his influential classical paper on crime and punishment, also Becker (1968) takes for granted that punishment must be legitimate in order to be permitted.

0899-8256/$ – see front matter
http://dx.doi.org/10.1016/j.geb.2012.10.011

decentralized, peer-to-peer sanctioning. In some situations, legitimate peer punishment characterizes the functioning of formal institutions, as it is the case with the well-known peer review system prevailing in the academic community at large. In other cases, legitimate peer punishment operates *informally*, as there are domains where this principle holds though it lacks formalization. In many parts of the world it is often the case that serious difficulties in establishing a central authority arise due to lack or weakness of the rule of law. Under these circumstances, social norms typically act as substitutes of formal institutions and decentralized legitimate punishment seems to be a better norm enforcement device than arbitrary sanctioning.[2] In the case of Turkana, a large-scale, politically uncentralized, egalitarian, nomadic pastoral community engaged in warfare in East Africa (Mathew and Boyd, 2011), antisocial behaviors like cowardice and desertions are severely punished by community-imposed sanctions.[3] At a lower scale, in small groups where problems due to principal–agent relationships frequently arise, self-monitoring via a fair rule appears to be a promising route to informally solve the free rider problem in the absence of monitoring opportunities on the part of the principal (Mas and Moretti, 2009).

What these otherwise distant situations where peer punishment is at work have in common is an underlying principle of *legitimacy*: only some people or institutions, thanks to their virtuous behavior within a given context, are entitled to sanction and only those who misbehave deserve to be sanctioned. We investigate this legitimacy–punishment nexus experimentally within a finitely repeated public goods game framework. We aim at comparing the performance, in terms of both cooperation and efficiency, of two decentralized sanctioning mechanisms such as a classic arbitrary punishment institution and a legitimate punishment one. While under arbitrary sanctioning everyone can punish everyone else in the group, so that cooperators and free riders alike could be sanctioned, the institution studied in this paper prescribes that people can punish only those who contribute to the public good *less than they do*. Our data provide clear evidence that legitimacy yields substantial benefits to both cooperation and welfare. We also focus on the role of information and find that restrictions on the punishment activity are effective only when feedback over how the most virtuous individuals do actually behave (in terms of contribution choices) is provided. On the whole, then, our results interestingly suggest that it is the *interaction* between the legitimate nature of the sanctioning system at work and feedback over peers' behavior that plays a critical role in determining final contributions and earnings levels.

The remainder of the paper is structured as follows. Section 2 reviews the related literature. Section 3 illustrates the experimental design. Section 4 reports our main results and Section 5 concludes the paper.

2. Background

In a laboratory public goods game or voluntary contribution mechanism (*VCM*), there is a group of subjects who, as the game starts, receive an individual monetary endowment, from which they may contribute any amount to a public good that returns a payoff to each of them. The structure of monetary payoffs in the *VCM* makes it a classical 'social dilemma', as each agent has a dominant strategy to free ride, while at the social optimum each individual contributes his entire endowment. Therefore, the straightforward prediction based on the material self-interest hypothesis (and common knowledge of it) is that everyone should free ride, both in the one-shot and in the finitely repeated version of the game. By contrast, in the latter setting, initial cooperation turns out to be relatively high and, as time unfolds, the well-known 'decay' pattern typically occurs (Ledyard, 1995). In the last decades, an increasing number of *VCM* experiments have been investigating the role that institutions can play in the enforcement of cooperation. In particular, decentralized sanctioning systems are among the most widely studied in the experimental literature (Ostrom et al., 1992). In their pathbreaking studies, Fehr and Gächter (2000, 2002) demonstrate that while in non-punishment treatments cooperation rates indeed tend to fall over time, decay does not occur insofar as players are allowed to incur a cost to decrease others' monetary payoffs. This is due to the fact that many participants are willing to engage in individual punishment of inappropriate behavior (despite the net monetary costs associated with its usage) both with anonymous random matching and with fixed groups playing a finite number of times. However, recent work convincingly indicates that a long overlooked 'dark side' of arbitrary costly punishment needs to be seriously taken into account.

First, this institution in many cases significantly undermines the scope for self-governance, as, since everyone is free to punish everyone else, sanctioning may take the form of misdirected, 'antisocial' punishment – that is, low contributors punishing high contributors. Available evidence reveals that antisocial punishment is quantitatively significant (Anderson and Putterman, 2006) and substantially reduces contribution rates (Cinyabuguma et al., 2006), to the point that in some subject pools cooperation in the presence of sanctioning can be even *lower* than in its absence (Gächter and Herrmann, 2011). The negative effect of antisocial punishment on contribution levels is larger as long as it is targeted at outgroup members when competition between groups is created (Goette et al., 2012) and when it occurs within less industrialized societies (Herrmann et al., 2008). Second, when multiple stages of punishment are allowed, so that immunity of sanctioners from reprisals is removed, counterpunishment and feuds are likely to be triggered, limiting, once again, successful self-governance and leading, eventually, to a demise of cooperation (Denant-Boemont et al., 2007, Nikiforakis, 2008 and

[2] In some environments, the locally prevailing social norms prescribe that not only cheaters but also non-punishers have to be sanctioned. See, on this, Kandori's (1992) classical paper.

[3] Neighborhood crime watches and 'naming and shaming' practices provide further examples along these lines, though within socio-economic contexts where social norms and formal institutions appear to be complements, rather than substitutes.