Legitimate Punishment, Feedback, and the Enforcement of Cooperation

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

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. While such 'vigilante justice' turns out to enhance cooperation when retaliation is prohibited, this comes at a substantial cost, as welfare levels are usually low. 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 in the framework of a public goods game, by comparing it with a discretionary punishment mechanism. Our findings show that, despite the lack of additional monetary incentives and the risk to produce motivation crowding-out effects on subjects' propensity to cooperate, the introduction of legitimate punishment leads to substantial efficiency gains. Further, players' earnings are significantly higher. We also focus on the role of feedback and we interestingly find that removing the information over high contributors' choices only leads to 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 an institution based on legitimate punishment effective.

JEL Classification: C73;C91; D02 ; D63.

Keywords: Public Goods Games; Peer Punishment; Cooperation; Legitimacy; Feedback.

[*]**Corresponding author** :

Luca Zarri

Email: luca.zarri@univr.it

Address: Department of Economics, University of Verona

Viale Università 4, 37121 Verona, Italy.

Tel. +390461282280; Fax +30461282222.

[a]marco.faillo@unitn.it, Department of Economics, University of Trento.

[b]daniela.grieco@univr.it, Department of Economics, University of Verona.

[c]luca.zarri@univr.it, Department of Economics, University of Verona.

1. Introduction

In naturally occurring environments, punishment is a widespread phenomenon. The ubiquity of sanctioning is probably due to a significant degree to its importance for the proper functioning of society at large, the maintenance of social order as well as the efficiency of small-scale organizations and communities. However, since in real-life environments meting out punishments implies inflicting (sometimes extremely severe) costs on the punishees, a typical feature of sanctioning mechanisms is that their usage is far from being discretionary and 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: typically, A is allowed to castigate B if A is entitled to do so (entitlement); B can be punished if he deserves to be sanctioned due to his misconduct (desert)[1].

Everyday life abounds in situations where punishment systems have to be legitimate in order to be implemented. In the case of centralized sanctioning, only some institutions (for instance, police and courts) are entitled to impose sanctions on wrongdoers[2]. Trade unions and employers' associations usually set up arbitration boards entitled to monitor and enforce the compliance of their members. Also in the international arena we can find examples of centralized institutional arrangements based on the legitimate recourse to sanctioning mechanisms against violators: the EU Stability and Growth Pact aims to enforce budgetary discipline among EU member states and the goal of the United Nations Framework Convention on Climate Change (the so called Kyoto Protocol) is to reduce global greenhouse gas emissions by implementing legally binding agreements.

But legitimate punishment often takes the form of *decentralized*, peer-to-peer sanctioning. Here, entitlement derives from one's virtuous behavior in a given context. In some situations, legitimate peer punishment characterizes the functioning of formal institutions. In this regard, a world well-known example is provided by the peer review system[3] prevailing in the academic community at large as well

[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.

[2] On philosophical grounds it can be plausibly maintained that the very existence of the modern state itself rests upon a fundamental legitimacy argument: in a democracy, citizens delegate the power to the state and, due to its being the legitimate representative of the people, the government has access to coercive power. Within their geographical boundaries, states are sovereign and allowed to sanction citizens adopting wrongful behavior right because society as a whole conferred to them the legitimacy to do so. For a recent experiment on the legitimacy of control, see Schendler and Vadovic (2011).

[3] Even though the implementation of the peer review system often relies on central coordination devices (e.g. scientific journals' editorial boards), such evaluation mechanism is in its essence decentralized: peers' judgments typically play a crucial role in determining final decisions. Beyond academia, other forms of peer review (such as clinical peer review,

as within single university departments and research centres all over the world. The recourse to scholarly peer review is frequent as it occurs in decisions related to faculty advancement and tenure and whenever private or public funds need to be allocated or research projects need to be evaluated for possible publication on journals or books. The underlying logic is that not everyone can punish everyone else; by contrast, only those who mostly proved to be able to significantly contribute to the advancement of science (i.e. the most productive academics) are entitled to judge and evaluate their peers' research papers and projects.

In other cases, legitimate peer punishment operates *informally*, in the sense that there are situations where this principle holds and is associated with the usage of sanctioning among peers even though it lacks formalization. At the international level, in the current political debate on the hot topic of nuclear weapons development, a forcefully repeated claim is that while democratic countries (e.g. Israel) are entitled to produce nuclear weapons, non-democratic regimes (e.g. Iran and North Korea) are not. In principle, having a world authority entitled to use nuclear power in defence of humanity (or prohibiting its development altogether to everyone) would be safer for many, but since we know that so far the creation of a world government turned out to be extremely difficult to implement, the above argument seems to provide a more convincing second-best solution compared to arbitrary peer punishment (i.e. allowing every country to develop nuclear weapons regardless of the nature of its political regime).

Next, in many parts of the world it is often the case that serious difficulties in establishing a central authority endowed with the power to sanction wrongdoing arise due to lack or weakness of the rule of law. Under these circumstances, social norms typically play a crucial role acting as *substitutes* of formal institutions and decentralized legitimate punishment seems to be a better informal institution, as a norm enforcement device, than arbitrary peer punishment[4]. As an example of this, we might consider the case of Turkana, a large-scale, politically uncentralized, egalitarian, nomadic pastoral community engaged in warfare in East Africa (Mathew and Boyd, 2011). Here, antisocial behaviors like cowardice and desertions are severely but informally punished by community-imposed sanctions. Neighborhood crime watches – i.e. organized groups of citizens voluntarily devoted to crime and

physician peer review and software peer review) share these key features.

[4] Social norms are widely shared views about acceptable behaviors within a given social context (see e.g. Herrmann et al., 2008). In some environments, the locally prevailing social norms also prescribe that *sanctions themselves*, despite being informally administered, have to be legitimate in order to be implemented. In other words, there are social norms implying that punishment can be an appropriate tool to enforce, say, a given norm of fairness, but they also prescribe that punishment has to be legitimate for this to be the case. The presence of legitimacy makes it easier to make a sanctioning institution socially accepted. This is in line with Kandori's (1992) approach, which includes informal punishment of deviants (i.e. the presence of peer-to-peer legitimate sanctions) in the very definition of social norms.

vandalism prevention within neighborhoods – as well as citizens signalling their disapproval to littering by directly collecting items discarded by others provide further examples along these lines, though within socio-economic contexts where social norms and formal institutions appear to be *complements*, rather than substitutes.

At a lower scale, in small groups where problems due to principal-agent relationships frequently arise, self-monitoring via a fair rule appear to be a promising route to informally solve the free rider problem in the absence of monitoring opportunities on the part of the principal. In horizontal relationships in the workplace, if a worker makes very little effort compared to her partner, she may induce resentment or face legitimate sanctions from her. Due to this, Kandel and Lazear (1992) show theoretically that it may be optimal for this worker to do her 'fair share' and work more productively: hence, peer effects can countervail free riding in partnerships. Mas and Moretti (2009) address this question empirically in the context of a retail firm and interestingly find evidence of strong peer effects associated with the introduction of high-productivity workers into work groups: a 10 percent increase in coworker productivity results in a 1.5 percent increase in individual productivity. As the two authors point out, "When workers hold themselves accountable to their peers, workplaces have the potential to be cooperative environments". Dal Bo' (2007) observes that in these environments the notion of social norm coincides with the concept of 'corporate culture' (see Baker et al., 2002), that is the unwritten codes of behavior that shape interactions among members of the firm affecting their behavior and the performance of the firm[5].

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. In this article, we investigate this legitimacy-punishment nexus experimentally within a finitely repeated public goods game framework. We compare two decentralized sanctioning mechanisms such as a classic, arbitrary punishment institution and a legitimate punishment one. Our data provide clear evidence that legitimacy yields substantial benefits to cooperation, as we show that, unlike arbitrary sanctioning, legitimate punishment turns out to be a powerful device in both enforcing cooperation and raising individual earnings. These results have important efficiency implications for the design of mechanisms intended to deter misconduct. We also focus on the role of information within a legitimate punishment regime and find that restrictions on the punishment activity are effective only when

[5] Management theorists long studied the concept of 'corporate culture', viewed as the set of habits and rules followed by employees within the firm. This set of rules and habits is stable in the long run as these rules include punishment of deviators (Cremer, 1986).

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 the amount of information over peers' contribution behavior provided to the subjects that plays a critical role in determining final contribution 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 discusses our findings and concludes the paper.

2. Related literature

In a 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, in contrast, at the social optimum each individual contributes his entire endowment. Therefore, the straightforward prediction based on so called *Homo Oeconomicus* is that everyone should free ride, both in the one-shot and in the finitely repeated game. However, in the finitely repeated version of this game, the following pattern typically occurs: initially, average contributions are relatively high, whereas, as the game unfolds, they gradually decline and cooperation converges to a near-negligible level (Ledyard, 1995). In the last years, an increasing number of *VCM* experiments have been investigating the role that institutions can play in the enforcement of cooperation. Some influential papers focused on centralized mechanisms (see Yamagishi, 1986; Chen and Plott, 1996; Falkinger et al., 2000), whereas others explored decentralized systems (Ostrom et al., 1992; Fehr and Gächter, 2000; 2002; Casari and Plott, 2003; Fudenberg and Pathak, 2010). Among institutional arrangements, punishment systems are among the most widely studied in the experimental literature. Since in social dilemmas the maximization of social welfare conflicts with individual payoff maximization, the implementation of a sanctioning institution aimed at castigating deviant individual behavior is an extensively used solution. So far, laboratory studies have concentrated on two broad classes of punitive mechanisms, tackling the social dilemma problem from two different angles: voluntary decentralized and centralized punishment.

2.1. The dark side of arbitrary peer punishment

As far as decentralized punishment is concerned, in their pathbreaking study, Fehr and Gächter (2000; 2002) demonstrate that while in non-punishment treatments (*VCM* without punishment opportunities) cooperation rates indeed tend to fall over time (round after round), this 'decay phenomenon' does not occur insofar as players are allowed to incur a cost to decrease others' monetary payoffs (*VCM* with punishment opportunities). Insofar as we suppose that in the laboratory subjects act selfishly in order to systematically maximize their monetary gains, costly punishment is a puzzle: in a finitely repeated *VCM* with punishment options, subjects should not use such options, due to the net monetary costs associated with their usage. By contrast, peer punishment of free riders has been shown to represent a powerful device to foster and successfully sustain cooperation in social dilemmas both with anonymous random matching and with fixed groups playing a finite number of times. However, recent work convincingly reveals that a long overlooked 'dark side' of arbitrary punishment exists[6]. In particular, the following important five drawbacks of this punishment regime have been identified in the last years: (1) the quantitative relevance of antisocial punishment in many subject pools; (2) the lack of robustness to institutional changes such as the possibility of retaliation; (3) the risk of motivation crowding-out; (4) the low level of average earnings and (5) the relative infrequency of arbitrary sanctioning institutions in real-life environments. Let us shortly discuss each of these (partially intertwined) limitations of arbitrary peer punishment.

Firstly, 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. Recent evidence indicates that antisocial punishment is quantitatively significant (Anderson and Putterman, 2006) and substantially reduces contribution rates (Cyniabuguma et al., 2006), to the point that in some subject pools cooperation in the presence of punishment 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., 2011) and when it occurs within less industrialized societies (Herrmann et al., 2008)[7]. Second, when multiple

[6] It is worth specifying that *any* decentralized punishment institution implies some restrictions over punishment opportunities. Also (what we refer to here as to) arbitrary punishment is restricted as retaliation is prohibited and first-order punishment only is allowed. Therefore, strictly speaking, what differentiates legitimate punishment from other decentralized punishment mechanisms (and arbitrary punishment in particular) is not the presence of restrictions *per se*, but the nature of imposed restrictions. However, for expositional ease, in the following sections we will use expressions like arbitrary (resp., legitimate) punishment and unrestricted (resp., restricted) punishment interchangeably.

[7] As Gächter and Herrmann (2011) correctly point out, "Punishment of cooperators has been largely neglected in previous research on social preferences because it was negligible compared to the punishment of free riders. Our results show that this neglect is not warranted because punishment of cooperators can be very significant in some subject pools".

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 Nikiforakis and Engelmann, 2010). Since the opportunity to retaliate punishments exists in many real-life decentralized interactions (Nikiforakis, 2008), these negative results show that an arbitrary punishment mechanism is not robust to realistic institutional changes. Thirdly, a further problem is that since this form of punishment exclusively relies on deterrence, that is on extrinsic motives to cooperate, the risk is either not to elicit people's intrinsic motivations to comply (if any) or even to crowd them out, especially when incentives are weak[8]. Fourth, it is important to note that solving the free rider problem is only one part of the problem as a whole. Recent papers indicate that, even in the presence of a single stage of sanctioning, the success of discretionary punishment in enforcing cooperation comes at a substantial cost. Botelho et al. (2005) analyze Fehr and Gächter's (2000; 2002) data and find *lower earnings* when punishment is allowed than under no punishment (see also Bochet et al., 2006 and Cyniabuguma et al., 2006 for similar results)[9]. This evidence shows that 'vigilante justice' is a double-edged sword (Goette et al., 2011), as it raises cooperation levels but, unless we consider a significantly longer time horizon (Gächter et al., 2008), leads to average earnings which are lower than in the absence of sanctioning opportunities[10]. From an economic perspective, this is a serious shortcoming of unrestricted punishment, showing that such a sanctioning system risks to determine efficiency losses and, therefore, to turn into a wasteful activity for those societies or organizations that adopt it. Finally, a fifth relevant problem with this cooperation enforcement device has been recently raised by Guala (2011), who questions its quantitative relevance outside the laboratory. As he points out, ethnographic evidence from tribal societies or the historical evidence on common pool resource usage does not provide a lot of support for either the use or the efficacy of this form of costly sanctioning[11].

[8] Fehr and Rockenbach (2003) provide experimental evidence that sanctions underlying selfish or greedy intentions – unlike sanctions perceived as fair – produce extremely negative effects on cooperation. For two recent studies on the importance of the interaction between social preferences and formal institutions, see Gneezy et al. (2011) and Bowles and Polania-Reyes (2011).

[9] The same occurs to average payoffs in 13 out of 16 participant pools of Herrmann et al. (2008). Similarly, Dreber et al. (2008) show that when in a repeated prisoner's dilemma players can choose between cooperation, defection and costly punishment, average group payoffs are not higher than when the sanctioning option is not available. Further, since punishing is costly not only for the punishees but also for the punishers, the 'winners' (i.e. those who get the highest earnings) in their experiment are the individuals who abstain from sanctioning.

[10] Denant-Boemont et al.'s (2007) study also finds that the existence of additional rounds of sanctions has a significantly negative effect not only on the level of contributions but also on welfare levels, which turn out to be lower than when no sanctioning mechanism exists.

[11] As shown by Nikiforakis (2008), a relevant problem with arbitrary punishment in real-life situations (but not in the lab, insofar as only first-order sanctioning is permitted) is the high risk of retaliation by the punishees.

2.2. The weaknesses of centralized punishing mechanisms

On the whole, the arguments reported above strongly question the belief that individuals are able to successfully govern themselves through discretionary peer punishment (Nikiforakis, 2008). Thus, it is natural to think that an alternative, viable solution could be to delegate the power to sanction non-cooperators to an external enforcement agent, i.e. a Hobbesian 'Leviathan' entitled to watch individuals' behavior and punish free riders[12]. However, a centralized solution appears to be largely unsatisfactory under many respects. The major reason is fourfold. Firstly, in many social environments we cannot take for granted that creating an Orwellian institution monitoring citizens' behavior is socially *desirable*, especially in an era when citizens are increasingly concerned about privacy protection. Another reason why a 'Big Brother society' may turn out to be an inadequate solution has to do with the informational dimension. As noted by Mas and Moretti (2009), in many jobs employers cannot observe the exact contribution provided by each worker to the production of total output: this feature of the workplace is common in most clerical occupations, many manufacturing jobs, construction, agriculture, and retail, especially when the number of employees working on a task is large. As we pointed out in the Introduction, the overwhelming acceptance in science of an evaluation system such as the peer review process provides us with a clear example of a decentralized mechanism which is considered better than a centralized one. The rationale for this can be identified in the belief that, in order to properly evaluate projects and research products, relying on each field's main experts is the wisest choice. The underlying argument, which has a clearly Hayekian flavor, is that the relevant knowledge is dispersed and a decentralized system is better able to detect it and fulfil its potential, compared to a centralized one. Analogously, the provision of a local public good like urban security often make citizens' collaboration with the central authority crucial for a policy to be successful. Despite the recent improvements in technological surveillance techniques, it would be impossible even for the most efficient 'hired gun' to monitor all urban areas all the time and rapidly intervene against criminals. The dispersed nature of information is the main reason why informing on other citizens has often been invoked and incentivized to complement the Leviathan's efforts and foster public goods provision in many domains (e.g. tax compliance)[13].

[12] Andreoni and Gee (2011) take this road to deal with the issue of cooperation within small-scale organizations and propose a method which they term the 'hired gun' mechanism. They experimentally investigate whether a delegated authority (which, therefore, ceases to be a peer in the group) that punishes only the largest deviators can effectively enforce socially desirable outcomes. They also compare the effectiveness of this hired gun to discretionary peer punishment.

[13] A movie like 'The Lives of the Others' has dramatically unveiled the importance of this tool even within a powerful centralized system like the communist East Germany before the fall of the Berlin Wall: if the totalitarian regime had held all the relevant information about its citizens, it would not have been necessary to systematically rely on informing on the part

Thirdly, even apart from informational problems, monitoring individuals can be extremely costly. In particular, recent studies emphasize the importance of potentially significant 'hidden costs of control' (Falk and Kosfeld, 2006; Schnedler and Vadovic, 2011). In the standard analysis of the principal-agent relationship, principals hire agents due to the efficiency gains conferred by delegation. However, principal-agent relationships are typically characterized by a conflict of interest and asymmetric information. Falk and Kosfeld's (2006) laboratory results indicate that the decision to control significantly reduces the agents' willingness to act in the principal's interest: explicit incentives backfire and performance is lower if the principal controls, compared to if he trusts[14]. As the two authors point out, "Elements in the labor contract that can be perceived as signals of distrust and control, such as minimum performance requirements, may harm more than help. Similarly, characteristics of the workplace environment that limit freedom of choice and signal low expectations, such as high levels of monitoring and surveillance, may be equally counterproductive" (p. 1612).

Finally, there are relevant implementation difficulties to be seriously taken into account. The major one can be identified in what Kosfeld et al. (2009) term the 'dilemma of endogenous institution formation': jointly, everyone profits if a sanctioning institution is formed, but each individual profits more if only the others form the institution[15]. The presence of some individuals trying to free ride on the formation process by refraining from participating risks to dramatically weaken the newborn institution from the outset. As noted by the authors, the United States' withdrawal from the Kyoto Protocol had a strong impact on other nations' (e.g. Netherlands and Australia) motivation to fulfil the agreement[16]. More generally, we see today that reaching binding agreements to create new political and economic institutions at the supranational level to effectively deal with new, global-scale problems, turns out to be a rather ambitious goal, as it typically implies a long-term delegation of power.

Hence, for these four reasons, centralized punishing mechanisms do not seem to represent a satisfactory solution to overcome social dilemmas. Moreover, as we noted in the introductory section, in many parts of the world it is often the case that serious obstacles in establishing a central authority endowed with the power to sanction wrongdoing arise due to lack or weakness of the rule of law.

of their relatives, friends, colleagues and neighbors.

[14] These findings confirm that it is always extremely important to carefully consider the (sometimes subtle and counterintuitive) interplays taking place between institutional arrangements and individual preferences (including social ones; see Gneezy et al., 2011 and Bowles and Polania-Reyes, 2011).

[15] As the authors point out, this is a particular type of the well-known 'second-order free riding problem'.

[16] This is in line with one of the main results of Kosfeld et al.'s (2009) experimental analysis, as they show that only full-size organizations, in which all players participate, have a reasonable chance of being implemented. As far as laboratory studies dealing with centralized punishing systems are concerned, Chen and Plott (1996) find that the classic Groves-Ledyard mechanism calls for high punishment levels in order to lead to higher provision of the public good and higher efficiency.

Under these circumstances, informally enforced social norms typically play a crucial role acting as substitutes of formal institutions.

3. Experimental setup

3.1. Legitimate peer punishment

In the previous section we have shown that, as far as punitive institutions aimed at solving free rider problems are concerned, relevant downsides emerge with regard to both arbitrary peer punishment and centralized sanctioning mechanisms. Therefore, in this paper we decided to take a different road. In particular, like the studies referred to in Section 2.1., we focus on a decentralized institutional arrangement based on sanctioning opportunities. However, we depart from the above cited papers by investigating 'legitimate peer punishment': only 'virtuous' participants (i.e. relatively high contributors) are entitled to sanction and only free riders (i.e. relatively low contributors) can be sanctioned[17]. We claim that this key feature of our experimental design captures the legitimacy element which characterizes many real-life situations, including the ones reported in the Introduction. While some recent studies have concentrated on institutions which derive their legitimacy from a *process* of endogenous choice (Gürerk et al. 2006, Ertan et al., 2008, Kosfeld et al., 2009 and Sutter et al., 2010), we analyze an enforcement device which is exogenously imposed (like in Fehr and Gächter, 2000) but at the same time legitimate due to its inner, structural features, i.e. due to its conditioning the possibility to punish on the adoption of cooperative behavior in the first place[18]. Hence, in our experiment the legitimate punishment institution allows for the endogenous formation of subjects who are 'first among equals', who may change from round to round. While under legitimate centralized punishment it's one's belonging to the institution itself that entitles one to sanction others[19], under legitimate peer punishment it's one's *behavior* that entitles her to punish her peers. As far as formal institutional arrangements are concerned, this mechanism is aimed at capturing the essential features of institutions in which, for the reasons discussed in the previous section, a central authority (e.g. a principal) delegates the monitoring activity to the individuals who actually work for the pursuit of a common objective (e.g. a team of agents), but at the same time regulates this activity by introducing some

[17] The specific details of our experimental setup will be described in subections 3.3 and 3.4.

[18] Related papers where punishment is not unrestricted include Ertan et al. (2009), Xiao and Kunreuther (2010) and Casari and Luini (2009). In the latter, sanctioning is permitted only insofar as it is requested by a coalition of at least two subjects.

[19] In turn, the possibility for a person to be hired by a legitimate centralized punishing institution often depends on one's previous conduct. For example, in many countries, you need to have a clear criminal record to apply for jobs such as police officer or judge, where you will need to punish violators on a daily basis.

general procedural principles defining who is entitled to punish whom. The central authority – for example, the editor of a journal or the head of a funding institution in the case of the peer review process – might directly intervene when these general procedural principles are violated – for example when an unqualified person is appointed as referee –, but it does not question the decentralized logic of the enforcement device (the peer-to-peer evaluation process, in this case). In other words, the sustainability of cooperation under legitimate punishment rests ultimately on the power of peers, i.e. on individuals' autonomous decision to punish the free riders.

In our sanctioning institution, some key restrictions are imposed over both *who* is allowed to punish and *whom* punishers can punish[20]. These assumptions are in line with what happens within several naturally occurring environments like the ones recalled in the Introduction, where it is often the case that the social acceptance of punishment is conditional on (i) the punisher being entitled to punish (*entitlement*) and (ii) the punishee being a wrongdoer and, therefore, deserving to be punished (*desert*). When the two requirements of entitlement and desert are met, we say that punishment is legitimate (i.e. a principle of legitimacy holds)[21].

Since we investigate a finitely repeated *VCM* with punishment options, a two-stage game gets played in every period: at stage 1, players simultaneously choose how much to contribute to the public good (contribution stage) and at stage 2 they have access to punishment options (punishment stage). However, the principle of legitimacy requires that a single individual acts as a 'high contributor' at stage 1 in order to earn the right to be a punisher at stage 2[22]. More specifically, we assume that a subject is entitled to punish another subject at stage 2 only if her contribution at stage 1 has been *strictly higher* than the contribution of the peer she wants to punish[23]. As a consequence, high

[20] Therefore, our design also differs from recent experimental *VCM* protocols where norms prescribing who can punish and/or who can be punished emerge endogenously within a group (see e.g. Casari and Luini, 2009; Kosfeld et al., 2009). Casari and Plott (2003) is an example of an experimental paper where, like in the present setup, 'virtuous' restrictions on punishment are exogenously imposed. Xiao and Houser (2011) assume that when a round is monitored, then that round's lowest contributor will incur a small sanction. However, they suppose that punishment is not peer-to-peer but exogenous, that is under the experimenters' control.

[21] As we observed in the introductory section, this may occur either formally (when formal institutions entitled to sanction wrongdoing exist) or informally (e.g. when social norms supported by informal (legitimate) sanctions are in force), depending on the social context.

[22] As far as immediate monetary consequences of subjects' sanctioning decisions are concerned, it is worth noting that while in Casari and Plott (2003) the subjects who find and sanction free riders are monetarily rewarded, in our design legitimacy, by allowing cooperators to have access to punishment options, only makes them entitled to costly punish wrongdoers. Xiao and Kunreuther (2010) compare deterministic vs. stochastic punishment in the framework of a prisoner's dilemma game and, in two out of six treatments, introduce a rule such that, like in the present paper, only cooperators are allowed to punish non-cooperators. However, studying the impact of legitimate punishment in a two-player game like the prisoner's dilemma, where each player always knows who punished whom, significantly differs from investigating the effectiveness of legitimacy in a multi-player environment like the *VCM*.

[23] This implementation of the principle of legitimacy differs from the prevailing form of restricted punishment endogenously emerging in Ertan et al. (2008). In their public goods game experiment, subjects vote on whether to allow sanctioning of group members whose contributions are (a) below-average, (b) above-average and (c) equal to the average: it

contributors are (partially) immune from punishment, in the sense that they cannot be sanctioned by players who contributed less or the same amount as them. Like in a standard, finitely repeated *VCM*, insofar as all the subjects are supposed to be driven by material self-interest only and this information is common knowledge, the unique subgame perfect equilibrium is for all agents to *never punish* and *never contribute*.

3.2. The role of information over virtuous peers' contribution behavior

It is reasonable to believe that in this context the impact of punishment on cooperation could also depend on the amount of information about others' behavior, a variable which has surprisingly received scant attention among experimental economists within the rich punishment literature[24], and which could have a significant influence on the perception of the legitimacy of the sanction. We believe that within an environment in which the right to sanction is awarded on a meritocratic basis, feedback over how the most virtuous members of the group behave might play an important twofold role in promoting cooperation. First, when this information is provided, a member who has been punished is not only aware of the fact that her contribution to the public good is lower than the contribution of the member who has sanctioned her, but she also knows the exact level of contributions of those who have gained the right to punish. In this sense, the provision of information on the most virtuous members' choices contributes to shed light on the degree of entitlement of the punishment activity. Second, this kind of feedback could also serve a pure cognitive function, as an individual who knows how the virtuous members of her group behave also knows what she ought to do to avoid undergoing punishment in the next future and what is the level of contributions expected by the other group members.

In our study, we address this problem by comparing the case in which subjects have information on every other coplayer's contributions with the case in which each member is informed only about the average contribution of her group and on the contribution of the members who have contributed strictly less than herself. In the latter case, members whose contribution is not the highest do not know what is the highest level of contribution in their group.

turns out that eventually the majority of groups opt for prohibiting punishment of higher-than-average contributors. Noussair and Tan (2009) investigate whether this ability of a voting process to converge to the optimal institutional structure is robust to a specific change in the environment, that is the existence of heterogeneity in the value to the group of subjects' contributions. While their results extend the findings of Ertan et al. (2008), the two authors also find that agents fail to converge (through voting) to the efficient punishment regime.

[24] For recent exceptions, see Nikiforakis (2010), Fudenberg and Pathak (2010), Grechenig et al. (2010) and Xiao and Houser (2011). As Nikiforakis (2010) points out, institutional details such as the format in which feedback about the actions of others is given can affect the efficacy of peer punishment in promoting cooperation.

3.3. Procedures

A total of 168 subjects participated voluntarily in the experiment at the CEEL Lab of the University of Trento. A total of 9 sessions were conducted, between December 2009 and November 2010. Six sessions had 20 participants and the other three had 16 participants. The experiment was programmed by using the z-tree platform (Fischbacher, 2007). The subjects, were undergraduate students (64.3% from Economics, 49.5 % females, 80.3 % Italian). We employed a between subjects design: no individual participated in more than one session. In each session, the participants were paid a 5 euro show up fee, plus their earnings from the experiment. The average payment per participant was 15.70 euros (including the show-up fee) and the sessions averaged approximately 1 hour and 30 minutes. At the beginning of each session, participants were welcomed and asked to draw lots, so that they were randomly assigned to terminals. Once all of them were seated, the instructions[25] were handed to them in written form before being read aloud by the experimenter. We took great care to ensure that the participants understood both the rules of the game and the incentives. They had to answer several control questions and we did not proceed with the actual experiment until all participants had answered all questions correctly.

In each session, there were 20 periods of interaction that proceeded under identical rules. The participants in a session were randomly assigned to groups of size four, so that they did not know the identities of the other members of their group. Like other experimental studies (see e.g. Cinyabuguma et al., 2006; Denant-Boemont et al., 2007), we used a partner protocol that kept the composition of each group constant over rounds, so that, at the end of each period, individuals remained in the same group. We did this as repeated interaction is a typical feature of several naturally occurring environments (e.g., businesses or collectives) where collective action problems arise and peer punishment occurs (Mas and Moretti, 2009; Mathew and Boyd, 2011 and Xiao and Houser, 2011). However, individuals' labels were reassigned on a random basis in each period. For example, the same player could be designated as player *45* in period t, as player *6* in period $t+1$, and as player *38* in period $t+2$. Therefore, our partner protocol was also characterized by anonymity of the components of the group and change of participants' labels across rounds[26]. The design and the parametric structure of the experiment are based on those of Fehr and Gächter (2000).

[25] A translation of the instruction sheet is provided in Appendix A. Original instructions were written in Italian. They are available upon request from the authors.

[26] Although a stranger protocol with random re-matching allows ruling out strategic punishment and reputation motives altogether, a partner protocol seems to work as well as a stranger protocol. Nikiforakis (2008), based on Botelho et al. (2005), addresses this issue by comparing results from a stranger protocol and a partner protocol and finds that differences in punishment decisions are not significant (whereas differences in punishment levels are).

3.4. Treatments

As anticipated above, our experimental design focuses on the role of both the institutional and the informational dimension. In order to do this, we implemented the following three treatments: (1) a baseline, unrestricted punishment and full information (Baseline) treatment, (2) a restricted punishment with full information (Full R.) treatment and (3) a restricted punishment with partial information (Partial R.) treatment.

There were 3 sessions (20 subjects in two sessions and 16 in the other) for the Baseline, 3 sessions (with 20 subjects in two sessions and 16 in the other) for the Full R. and 3 sessions (with 20 subjects in two sessions and 16 in the other) for the Partial R. For each treatment, in each session the subjects were divided in groups of N=4 (as in standard *VCM* experiments) subjects, who played a two-stage finitely repeated public goods game with punishment options for T=20 periods. Participants were aware of the number of rounds they were going to play and of the number of stages: information on the following rounds allows one to evaluate the effect of the threat of being punished in stage 2 and on contribution decisions in stage 1.

Overall, the three treatments differ along two dimensions (see Table 1): *nature of peer punishment* (unrestricted vs. restricted) and *feedback about others' contribution levels* (full vs. partial) in the group.

[TABLE 1]

3.4.1. Baseline

In the Baseline treatment, punishment is unrestricted and subjects are provided with full information, that is there is feedback about *all* their group co-players' individual contributions. This is a replication of the standard *VCM* with punishment and partner protocol (Fehr and Gächter, 2000), where everyone can freely punish everyone else in the group. Each of the 20 rounds consists of two stages.

In stage 1, each participant receives a fixed amount e =20 of tokens and has to decide whether she wants to invest or not an amount $g_i \leq e$ into a public project. Decisions are made simultaneously and with no information about peers' choices. At the end of stage 1, each participant is informed about her current earnings, which consist of two elements:

a. The amount of her initial 20 tokens that she has kept for herself (i.e. 20 tokens – her contribution to the project);

b. Her income from the project. The income to her is equal to 40% of the total of the four individual contributions to the project.

Therefore, her earnings at the end of stage 1 are calculated by the computer in the following way:

$$\pi_i = (20 - g_i) + 0.4\sum_{j=1}^{4} g_j$$

In stage 2 subjects are informed about the contribution the other members of their group and can decide to assign between 0 and 10 punishment points to any them. Points assignment is costly and costs are charged according to a standard cost function as in Fehr and Gächter (2000) (Table 2).

[TABLE 2]

Each point that a subject receives reduces her earnings at stage 1 by 10%.
Punishment is anonymous: subjects do not know the identity of the peer who has punished them.
Each participant's earnings at the end of stage 2 are calculated by the computer in the following way:

Each participant's earnings after stage 2 = earnings at the end of stage 1- cost of points she assigned at stage 2 - 10%* number of points received*earnings at the end of stage 1

3.4.2. Legitimacy-based treatments

The other two treatments are characterized by the presence of legitimacy (i.e. entitlement and desert): both in the Full R. and the Partial R. treatment, a subject is entitled to sanction another subject in stage 2 only if her contribution at stage 1 has been *strictly higher* than the contribution of the peer she wants to punish. The difference between the two treatments regards the feedback that subjects receive at the end of stage 1, in each period: while in Full R. subjects are informed about the full vector of others' contributions (like in the Baseline), in Partial R. subjects are informed only about the *average* contribution level and the specific contribution levels of their group co-players who contributed *strictly less* than them. Therefore, no specific information about more virtuous peers is provided to them in this treatment. As in the Baseline, the punisher's identity is unknown to the punishee.
In these two treatments stage 1 is exactly the same as in the Baseline, but now participants know that they can go on with stage 2 in the experiment only if they contribute more than their peers, that is, as

we explained above, only if they are entitled to do so[27]. Specifically, player *i* will be entitled to sanction player *j* in stage 2 only if $g_i > g_j$. In stage 2, subjects are given the opportunity to simultaneously punish those who contributed less than them by assigning a certain amount of points. This implies that the highest contributor in a group is fully immune from punishment.

4. Results

4.1. Contribution levels

Figure 1 displays the time pattern of average contributions by period in the three treatments.

[FIGURE 1]

In all the treatments contribution levels do not decline over time.

Result 1. Punishment prevents the decline of cooperation over time in all the treatments.

[TABLE 3]

Besides this well-known general positive effect of punishment, our data (Table 3) show that, given the same type of restrictions over the punishment activity, subjects who are informed about the contributions of all the other members of their group (Full R. treatment) contribute significantly more than subjects who are informed only about the average contribution of their group and on less virtuous peers' contributions (Partial R. treatment) (Wilcoxon Rank-sum Test with group averages as observation: z=2.43; p-value: 0.014). At the same time, given the same level of information, contributions in the Full R. treatment are on average significantly higher than contributions in the baseline treatment (z=2.61; p-value: 0.08). The introduction of restrictions on the punishment activity has a positive effect on the level of contributions. These differences characterize also the distribution of contributions in the final period of the game (Figure 2). Result 2 follows.

[FIGURE 2]

[27] It is important to make clear that we never used loaded terms such as 'legitimacy', 'entitlement', 'desert', 'punishment', 'free riding' and 'immunity' during the experiment.

Result 2. The introduction of restrictions significantly increases the level of cooperation. In the presence of restrictions, when partial information over other contributions is provided, a significant decrease in cooperation occurs[28].

This result is supported by the regression analysis[29] reported in Table 4, which takes into account the effect of a set of control variables and sheds further light on the role of restrictions and information in shaping the contribution levels.

[TABLE 4]

Besides the treatment effect, contributions in each period are positively (and significantly) affected by the average contribution in the group in the first period (variable AV_first). Therefore, each group's behavior in the first period represents a key determinant of subsequent contribution choices in the group: cooperation seems to be sustained also by idiosyncratic features of the specific group.
Higher contributions in the Full R. treatment also result in a higher level of efficiency (figure 3). Taking group average earnings as independent observations, we observe that average earnings in the Full R. treatment are significantly higher than average earnings both in the Baseline (Wilcoxon Rank- sum Test: z=2.52; p-value: 0.011) and in the Partial R. treatment (Wilcoxon Rank- sum Test: z=2.89; p-value: 0.003)[30].

[FIGURE 3]

Result 3. Average earnings are significantly higher when punishment activity is restricted and subjects have information on the contributions of all the other members of their group.

[28] The levels of contribution observed in the Partial.R and in the Baseline are not significantly different (Wilcoxon Rank-sum Test with: z=-0.046; p-value: 0.96). Note however that a direct comparison between the Baseline and the Partial.R treatments is not particularly useful, since Partial.R differs from the Baseline both for the presence of restrictions and for the quantity of information provided to the subjects.

[29] All the estimations have been carried out with STATA 11.

[30] The result is robust to controls for average contribution in the first period, quantity of assigned points, quantity of received points, gender, age, nationality, major and number of past experiments.

4.2. Punishment behavior

As Result 2 shows, the introduction of restrictions in the aim of preventing the assignment of punishment points to virtuous subjects results in higher contribution levels. In order to account for this evidence we shall give a closer look at the punishment activity in the three treatments and assess the impact of antisocial punishment in the Baseline treatment.

[FIGURE 4]

With regard to the distribution of punishment points, in all the treatments we observe the typical decreasing pattern, which is faster in the Full R. treatment (Figure 4). The difference between the average quantity of points assigned in the three treatments is not statistically significant (Table 5) (Wilcoxon rank sum Full R. vs Partial R.: z=-1.19; p-value= 0.23; Wilcoxon rank sum Full R. vs Baseline: z=-0.87; p-value= 0.38).

[TABLE 5]

However, it is worth noting that in the Baseline treatment a non-negligible percentage of punishment points are assigned to virtuous subjects. Table 6 reports the absolute quantities (column 2) and the percentage (column 3) of punishment points assigned in the Baseline treatment by a subject *i* to a subjects *j* when the contribution of *i* is smaller than the contribution of *j*. We define this type of behavior as "weak antisocial punishment", as distinguished from "strong antisocial punishment". The latter is observed when *i* punishes another subject *j* whose contribution is greater than both the contribution of *i* and the average contribution of the group (columns 4 and 5). In our sample 19.5% of the overall punishment activity (number of punishment points assigned in all periods) can be classified as weak-antisocial, while 12.2% is strongly antisocial. On average 14.4% of group' s punishment points assigned is weakly antisocial and 9% is strongly antisocial.

[TABLE 6]

[FIGURE 5]

The presence of a strong form of punishment of virtuous subjects (strong antisocial punishment) in the Baseline treatment emerges also in Figure 5, which displays the relationship between the distance from the average contribution of the group and the average quantity of points received. In the Baseline treatment, in some cases strong positive deviations are still punished. This evidence is supported by the results of the following regression analysis (results in table 7):

$$\textit{punishment points received}_{igt} = \beta_0 + \beta_1\, pos\_dist\_\, av_{igt} + \beta_2\, neg\_dist\_\, av_{igt} \qquad (Eq.\ 1)$$

where $pos\_dist\_\ av_{igt}$ is the positive distance from the group's average contribution, i.e. the difference between the subject's contribution and the group average contribution; this variable is equal to zero when the subject's contribution is below the average. The variable $neg\_dist\_\ av_{igt}$ is the absolute negative distance from the average of the groups, i.e. the absolute value of the difference between the group's average contribution and the subject's contribution; it is equal to zero when the subject's contribution is above the average.

[TABLE 7]

While in all the treatments the quantity of punishment points received decreases as the negative distance from the average increases, positive distance from the average has a significant effect on the quantity of points received only in the two treatments with restrictions.

Result 4. When the punishment activity is unrestricted, a non-negligible percentage of points are assigned also to subjects who contribute more than the punisher (weakly antisocial punishment) and in some cases also to the most virtuous subjects (strongly antisocial punishment).

Result 4 is compatible with the higher level of contributions observed in the Full R. treatment, where both weakly antisocial and strongly antisocial punishment are ruled out.

4.3. Determinants of changes in individual contribution levels

As we have shown in the previous subsections, the three treatments are significantly different in terms of contributions levels, but not in terms of punishment points assigned. Hence, an analysis of the effects of punishment in altering contribution levels is in order. In particular, we test whether high contributors' and low contributors' reactions to punishment are different. Having observed that a non-negligible share of punishment activity in the treatment without restrictions (Baseline) can be classified as antisocial, we shall investigate whether this punishment has also a perverse effect on the contribution level of the most virtuous members of the group[31] – i.e. whether it weakens their willingness to cooperate. In order to do this, the following equation is estimated for each treatment, distinguishing between subjects whose contribution is *below* the average contribution of the group and subjects whose contribution is *equal or greater than* the average of the group:

$$contribution_{igt} - contribution_{igt-1} = \beta_0 + \beta_1\, received\_punishment_{\,igt-1} + \beta_2\, dist\_av_{igt-1} \qquad (Eq.\ 2)$$

where *received_punishment* $_{igt-1}$ represents the number of punishment points that the subject has received in the previous period, whereas $dist\_av_{igt-1}$ is the distance between the subject's contribution and the average contribution in the group in the previous period. Results of the estimation are reported in Table 8, which shows a regression to the mean in all the treatments observed also by Denant-Boemont (2007): the higher the distance from the average in the previous period, the higher is the absolute increase of the contribution level in the current period.

With regard to the effect of punishment, we detect a positive and significant effect on low contributors' change in levels of contribution in the two treatments with restrictions (Full R. and Partial R.). The same effect is not observed for low contributors of the Baseline. Moving to high contributors, in the Baseline treatment we observe a negative reaction to punishment. The opposite effect is observed in the treatment with full information and restrictions (Full.R), while high contributors in the Partial.R do not show any significant change in the level of contributions as a consequence of punishment. This evidence confirms the presence of a significant perverse effect of antisocial punishment that can explain the low level of contributions observed in the Baseline. The introduction of restrictions prevents the occurrence of this effect because high contributors know that punishment points come from the most virtuous members of their group.

[TABLE 8]

Result 5. In all the treatments, regardless of the presence of restrictions, the increase in contribution levels is stronger the higher the distance from the average in the previous period.

Result 6a. Punishment has a positive effect on low contributors' willingness to cooperate only in the presence of restrictions.

[31] For a detailed analysis of this effect see Ones and Putterman (2007).

Result 6b. Punishment exerts a negative effect on high contributors' willingness to cooperate in the Baseline treatment, while it has a positive effect in the case of high contributors in the treatment with full information and restrictions.

Result 6a interestingly shows that free riders who are punished in a given period increase their contribution in the subsequent period only insofar as punishment is legitimate. A possible interpretation is that under this institution free riders feel guilty for not adopting the situationally appropriate behavior (Ross and Nisbet, 1991) and positively react to what they perceive as a signal of disapproval on the part of their virtuous peers. This would be in line with the 'incentives-as-signals' hypothesis suggested by Bowles and Reyes (2011) as well as with Masclet et al. (2003), who analyze the effect of non-monetary sanctions and find that free riders who receive more disapproval points significantly increase their contributions in the next period.

Finally, in the aim of exploring the role of information about others' behavior in shaping contribution reactions to punishment, we estimate the following equation for each treatment, by considering only the subsample of subjects whose contribution in the previous period was not the highest:

$$\mathit{contribution}_{igt} - \mathit{contribution}_{igt-1} = \beta_0 + \beta_1\, \mathit{received\_punishment}_{igt-1} + \beta_2\, \mathit{dist\_av}_{igt-1} + \beta_3\, \mathit{dist\_highest}_{igt-1} \qquad \text{(Eq. 3)}$$

Where $\mathit{dist\_highest}_{igt-1}$ is the distance between subject's own contribution in period t-1 and the highest contribution in the group in period t-1. We run two separate estimations, by distinguishing between subjects whose contribution level in the previous period is *below* the average of the group and subjects whose contribution level in the previous period is *above* the average of the group (Table 9). In the case of subjects with contribution levels below the average, information on the most virtuous peers *per se* does not affect the increase in contributions in the Full R. treatment, i.e. in the treatment where the full vector of peers' contribution is available, antisocial punishment is ruled out and subjects have the possibility to use virtuous peers' behavior as a reference point. The evidence on the Baseline is particularly interesting. In this case, for subjects who contribute below the average, the distance from the virtuous subjects exerts a significant and negative effect on the change in contribution levels: the lower the subject's contribution at t-1 with respect to the highest contribution of her group at t-1, the lower the increase in her contribution moving from t-1 to t. These subjects seem to use the information on most virtuous peers to infer the extent to which they can behave as free riders: the more altruistic their peers are, the more profitable the choice of behaving as a free rider. In other words, in the Baseline, the information about the highest contributors is (opportunistically) interpreted by the less virtuous subjects as the assurance that someone else is carrying the burden of the public project, so that there is no need to do the same.

With regard to subjects who contribute above the average, in the baseline treatment the information about the highest level of contribution in the group exerts neither a positive nor a negative significant effect on the increase in contributions. In the Full R. treatment, the highest level of cooperation is taken as a reference point.

[TABLE 9]

Result 7. In the full information treatment with restrictions, the highest contribution level in the group is used as a reference point only by subjects who contribute above the average of the group. In the full information treatment without restrictions, subjects who contribute below the average of the group use the information on the most virtuous members strategically to infer the potential gain from free riding.

5. Discussion and conclusion

Legitimate peer punishment is an ubiquitous phenomenon within several real-life domains, from teamwork and scientific research evaluation to neighboorhoods and international relations. Further, it plays a crucial role in the informal enforcement of social norms in communities where the rule of law is either weak or absent (e.g. warfare communities). Yet, so far there was no experimental evidence concerning the effects of legitimacy-based sanctioning institutions and feedback on cooperation. Our work contributes to shed light on the issue by means of a specially designed public goods game where punishment is permitted but high contributors only can punish and low contributors only can be punished[32].

We first wondered whether our legitimacy-based institution would be conducive to significantly higher cooperation levels, compared to the *VCM* with unrestricted punishment opportunities, despite the lack of additional monetary incentives to cooperate and the risk to generate motivation crowding-out. Our results confirm that this is the case, providing clear evidence that legitimate punishment yields

[32] As Fudenberg and Pathak (2010) point out, understanding *when* and *why* costly punishment actually facilitates cooperation in public goods games is important for the design of economic institutions.

substantial benefits to cooperation[33]. Thus, the effectiveness of using sanctioning mechanisms to encourage contributions to public goods appears to crucially depend on the *nature* of the punitive institution.

One reason why we decided to investigate legitimate punishment is that we expected such an institution not to suffer from the limitations which the recent studies cited in the previous sections have found with regard to unrestricted punishment. However, our first major finding was not an obvious one, as in previous public goods game experiments with punishment options it was the case that introducing institutional changes led to lower cooperation rates and earnings levels. For instance, Fuster and Meier's (2010) study indicates that monetarily incentivizing contributions may backfire. By contrast, we showed that, ceteris paribus, passing from arbitrary to legitimate peer punishment neither causes crowding-out nor is neutral (or slightly cooperation-enhancing): it significantly deter misconduct and raises cooperation rates. Further, it also increases earnings levels[34].

In our experiment, antisocial punishment is documented to play a relevant role when available: if the punishment activity is unrestricted, a non-negligible percentage of points are assigned also to subjects who contribute more than the punisher (weakly antisocial punishment) and in some cases also to the most virtuous subjects (strongly antisocial punishment). Under unrestricted sanctioning, the possibility that antisocial punishment occurs may also generate a 'motivational crowding-out' effect on virtuous players, as knowing that a significant probability to be castigated exists even for high contributors may weaken their willingness to cooperate. By contrast, insofar as sanctioning is legitimate, this effect can be ruled out. More generally, a critical condition for a punishment institution to be successful is that "the incentives provided by punishment do not crowd out pre-existing social preferences that might have induced contributions in the absence of punishment, as is observed in a large number of public goods and principal agent experiments surveyed in Bowles (2008) and Bowles and Hwang (2008). The counterproductive effects of explicit incentives in the experiments they survey appear to arise when the punishment or fines fail to evoke shame in the shirker, but rather convey negative information about the individual imposing the incentive" (Carpenter et al., 2009). Our results

[33] Also Ertan et al. (2008) find that an institution based on prohibiting punishment of high contributors is effective in raising cooperation levels and earnings. However, unlike the present study, they (1) focus on an average-based rule (the one we considered in an extension which is available upon request), rather than on a peer-to-peer rule, and (2) investigate the dynamics of its endogenous emergence (through voting) when several institutional options are available ex ante.

[34] In light of these results, we view our findings as supportive of evolutionary models based on group selection such as Boyd et al. (2003), where the possibility that punishment not only fosters cooperation but also raises group average payoffs plays a critical role.

suggest that, unlike under unrestricted punishment, the incentives provided by an institution based on legitimate punishment do not appear to crowd out pre-existing social preferences.

A possible explanation of these findings has to do with the idea that people can be strongly influenced by the cues of appropriate behavior offered by the situation in which an action is taken (Ross and Nisbet, 1991). As pointed out by Bowles and Reyes (2011), there are situations where an *incentives-as-signals mechanism* is at work: incentives are implemented for a purpose, and since the purpose is often evident to the target of the incentives, the target may also infer information about the person who designed the incentive and the nature of the task to be done (see on this also Falk and Kosfeld, 2006). In our setup, we cannot rule out that this kind of mechanism underlies the behavioral differences observed passing from unrestricted sanctioning to a legitimate punishment institution.

Therefore, one of the key messages conveyed by our work is that punishment need not be assigned to central authorities to effectively work, as a cooperation enforcement device. Decentralized punishment can be successful. The key requirement to be met for this to occur is that we go beyond vigilante justice and allow for legitimate peer punishment. Legitimate peer punishment represents an enforcement device which is at the same time *decentralized* – because the enforcement of cooperation is delegated to the members of the group –, *legitimate* – because a member of a group can punish another member only if her contribution is higher than the contribution of the member she wants to punish –, *feasible* – because it does not imply the delegation of power to a central authority –, and *efficient* – because it leads to higher levels of cooperation and earnings. As far as the principal-agent relationship is concerned, our findings suggest that a principal, though he cannot monitor his team of agents, may delegate a project and ask them to rely on legitimate peer punishment in order to accomplish the task.

We also find that the amount of information available to group members plays a critical role in determining the success of a legitimate punishment institution. In the first sentence of his survey on mechanism design, Myerson (1988) defines a *mechanism* as "a specification of how economic decisions are determined as a function of the *information* that is known by the individuals in the economy" (italics added). Hence, information is part of the very notion of what a mechanism is, from an economic viewpoint. We interpret our result over the comparison of the two legitimacy-based treatments as confirming that this is the case: the fact that removing part of the information provided to the people determines a radically different aggregate outcome, in terms of both cooperation rates and earnings levels, shows that information is a crucial component of the institution under study. The degree of information available to our players crucially affects the effectiveness of a legitimate peer punishment regime.

On the whole, our three-treatment design reveals that it is the *interaction* between behavioral restrictions and amount of information that crucially affects aggregate cooperation levels and earnings. The significant difference between contribution levels in the Full R. and the Partial R. treatment (Result 2) indicates that providing subjects with explicit information about higher contributors' choices, that is 'virtuous' subjects' behavior, plays a key role in the enforcement of cooperation. In this regard, it is natural to refer to an interesting series of recent experimental articles investigating the role of 'leadership' in social dilemma games and finding that leadership significantly raises average contribution levels. In these studies, leadership is typically implemented by letting an appointed leader influence others 'by example': she decides and announces her contribution *before* the other group members (simultaneously) make their contribution decisions (Güth et al., 2007). In contrast, in our work we impose all subjects' contribution and sanctioning decisions to occur simultaneously in every period. Further, higher contributors' choices are never made salient throughout the experiment. However, the significant difference in contribution levels observed between our Full R. and Partial R. treatment suggests the following interpretation: subjects behave as if they perceived the legitimacy principle as endogenously conferring a leadership to high contributors, by making them (and only them) entitled to sanction lower contributors and (at least partially) immune from sanctioning. Under full information, this form of endogenous leadership (through legitimacy) leads to a significant increase in average contribution levels, in line with the leadership papers[35]. An even more specific analogy connects our paper to Güth et al.'s (2007) experiment, where, in one of the implemented treatments, they suppose that full information holds and leaders can punish others through exclusion, i.e. veto power. Interestingly, it is right in this case of an 'empowered leader' – the closest to our Full R. treatment – that they obtain the strongest result in terms of contribution levels, also compared to cooperation rates observed under pure leadership by example[36].

Our study also leaves interesting avenues for further research, including the relative effectiveness of other legitimacy-based enforcement devices (e.g. based on positive incentives to cooperate, such as legitimate rewarding), the robustness of our major findings across alternative designs (e.g. ultimatum games, allowing for rejection only to responders who receive unfair offers) as

[35] As to empirical work, the effects of 'leading by example' have been analyzed with regard to charitable fundraising: a well-known result from these studies is that if renowned philanthropists donate to a specific project and this is publicly announced, others often tend to follow (Vesterlund, 2003). Further, so called 'seed money' typically generates a similar effect. We find that in our legitimacy-based framework a somewhat similar effect holds even though we refer to a simultaneous-move, rather than a sequential, game.

[36] As far as psychological experiments on leadership are concerned, it is interesting to note that several studies converge in finding a positive effect on contributions when the leader adheres to the principles of procedural fairness (see e.g. De Cremer et al., 2005).

well as the performance of the investigated mechanism across different cultural contexts. In this regard, we speculatively argue that legitimate punishment institutions might turn out to be even more effective, compared to vigilante justice, within developing societies, as recent research on cross-cultural differences (Herrmann et al., 2008; Gächter and Herrmann, 2011) indicates that the level of antisocial punishment here is far higher than in Western societies.

ACKNOWLEDGEMENTS: We gratefully acknowledge the Italian Ministry of Education, University and Research (2008 Prin project on "Social Capital, Corporate Social Responsibility and Performance" – Padova unit) and the Departments of Economics of the Universities of Trento and Verona for financial support. Our gratitude also goes to Alessandro Bucciol, Jeffrey Carpenter, Gary Charness, Luca Corazzini, Guido De Blasio, Roberto Galbiati, Laura Gee, George Loewenstein, Luigi Luini, Antonio Nicita, Stefania Ottone, Louis Putterman, Marco Rossi, Robert Sudgen, Lise Vesterlund, Erte Xiao and to the participants in the ESA World Meeting in Copenhagen, the VII LabSI Workshop on Experimental and Behavioral Economics in Siena, the 2nd Maastricht Behavioral and Experimental Economics Symposium, the IAREP/SABE 2010 Conference in Koln, the 2010 European Association of Law and Economics (EALE) Conference in Paris, the 2011 University of Southampton Workshop on the Determinants and Implications of Prosocial Behaviour, the XIV Summer School on Economics and Philosophy in San Sebastian, the European ESA Conference 2011 in Luxembourg, and the colleagues of the Department of Economics of the University of Verona, Siena, Padua and Florence for useful suggestions and comments. The usual caveats apply.

References

Alchian, A.A., Demsetz, H. (1972). Production, information costs, and economic organization, American Economic Review, 62 (5), 777–795.

Anderson, C., Putterman, L. (2006). Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism, Games and Economic Behavior, 54, 1-24.

Andreoni, J. (1993). An experimental test of the public-goods crowding-out hypothesis, American Economic Review, 83, 1317-1327.

Andreoni, J., Gee, L. (2011). The hired gun mechanism, mimeo, University of California, San Diego.

Baker, G., Gibbons, R., Murphy, K.J. (2002). Relational contracts and the theory of the firm, Quarterly Journal of Economics, 117, 39-84.

Becker, G. (1968). Crime and punishment, Journal of Political Economy, 76 (2), 169-217.

Bochet, O., Page, T., Putterman, L. (2006). Communication and punishment in voluntary contribution experiments, Journal of Economic Behavior and Organization, 60 (1), 11-26.

Botelho, A., Harrison, G.W., Pinto, L., Rutstrom, E.E. (2005). Social norms and social choice, Working Paper 30, Núcleo de Investigação em Microeconomia Aplicada (NIMA), Universidade do Minho.

Bowles, S., Polania-Reyes, S. (2011). Economic incentives and social preferences: substitutes or complements?, Journal of Economic Literature, forthcoming.

Boyd, R., Gintis, H., Bowles, S., Richerson, P.J. (2003). The evolution of altruistic punishment, Proceedings of the National Academy of Science of the United States of America, 100, 3532-3535.

Carpenter, J., Bowles, S., Gintis, H., Hwang, S. (2009). Strong reciprocity and team production, Journal of Economic Behavior and Organization, 71 (2), 221-232.

Casari, M., Luini, L. (2009). Cooperation under alternative punishment institutions: an experiment, Journal of Economic Behavior and Organization, 71(2), 273-282.

Casari, M., Plott, C. (2003). Decentralized management of common property resources: experiments with a centuries-old institution, Journal of Economic Behavior and Organization, 51, 217-247.

Chen, Y., Plott, C.R. (1996). The Groves-Ledyard mechanism: An experimental study of institutional design, Journal of Public Economics, 59, 335-364.

Cinyabuguma, M., Page, T., Putterman, L. (2006). Can second-order punishment deter perverse punishment?, Experimental Economics, 9 (3), 265-279.

Cremer, J. (1986). Cooperation in ongoing organizations, Quarterly Journal of Economics, 101, 33-49.

Dal Bo', P. (2007). Social norms, cooperation and inequality, Economic Theory, 30, 89-105.

De Cremer, D., van Knippenberg, D., van Knippenberg, B., Mullender, D., Stinglhamber, F. (2005). Rewarding leadership and fair procedures as determinant of self-esteem, Journal of Applied Psychology, 90 (1), 3-12.

Denant-Boemont, L., Masclet, D., Noussair, C.N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment, Economic Theory, 33, 145-167.

Dreber, A., Rand, D.G., Fudenberg, D., Nowak, M.A. (2008). Winners don't punish, Nature, 452, 348-351.

Ertan, A., Page, T., and Putterman, L. (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem, European Economic Review, 53, 495-511.

Falk, A., Kosfeld, M. (2006). The hidden costs of control, American Economic Review, 96 (5), 1611-1630.

Falkinger, J., Fehr, E., Gächter, S., Winter-Ebmer, R. (2000). A simple mechanism for the efficient provision of public goods: Experimental evidence, American Economic Review, 90, 247-264.

Fehr, E., Gächter, S. (2000). Cooperation and punishment in public goods experiments, American Economic Review, 90 (4), 980-994.

Fehr, E., Gächter, S. (2002). Altruistic punishment in humans, Nature, 415, 137-140**.**

Fehr, E., Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism, Nature, 422, 137-140.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments, Experimental Economics, 10, 171-178.

Fudenberg, D., Pathak, P.A. (2010). Unobserved punishment supports cooperation, Journal of Public Economics, 94 (1-2), 78-86.

Fuster, A., Meier, S. (2010). Another hidden cost of incentives: the detrimental effect on norm enforcement, Management Science, 56 (1), 57-70.

Gächter, S., Herrmann, B. (2011). The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia, European Economic Review, 55 (2), 193-210.

Gächter, S., Renner, E., Sefton, M. (2008). The long-run benefits of punishment, Science, 322, 5907, 1510.

Gneezy, U., Meier, S., Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior, Journal of Economic Perspectives, 25 (4), 1-21.

Goette, L., Huffman, D., Meier, S., Sutter, M. (2011). Competition between organizational groups: its impact on altruistic and anti-social motivations, Management Science, forthcoming.

Grechenig, K.R., Nicklisch, A., Thöni, C. (2010). Punishment despite reasonable doubt. A public goods experiment with uncertainty over contributions, Journal of Empirical Legal Studies, 7, 847-867.

Guala, F. (2011). Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. Behavioral and Brain Sciences, forthcoming.

Gürerk, O., Irlenbusch, B., Rockenbach, B. (2006). The competitive advantage of sanctioning institutions, Science, 312, 108-111.

Güth, W., Levati, M.V., Sutter, S., van der Heijden, E. (2007). Leading by example with and without exclusion power, Journal of Public Economics, 91, 1023-1042.

Herrmann, B., Thoeni, C., Gächter, S. (2008). Antisocial punishment across societies, Science, 319, 1362-1367.

Kandel, E., Lazear, E.P. (1992). Peer pressure and partnerships, Journal of Political Economy, 100 (4), 801-817.

Kandori, M. (1992). Social norms and community enforcement, Review of Economic Studies, 59, 63-80.

Kosfeld, M., Okada, A., Riedl, A. (2009). Institution formation in public goods games, American Economic Review, 99 (4), 1335-1355.

Ledyard, J. (1995). Public goods: a survey of experimental research, in Kagel, J., Roth, A. (eds.), Handbook of Experimental Economics, Princeton, Princeton University Press.

Masclet, D., Noussair, C., Tucker, S., Villeval, M.C. (2003). Monetary and non-monetary punishment in the voluntary contributions mechanism, American Economic Review, 93 (1), 366-380.

Mas, A., Moretti, E. (2009). Peers at work, American Economic Review, 99 (1), 112-145.

Mathew, S., Boyd, R. (2011). Punishment sustains large-scale cooperation in prestate warfare, Proceedings of the National Academy of Sciences of the United States of America, 108, 11375-11380.

Myerson, R.B. (2008). Mechanism design, The New Palgrave Dictionary of Economics.

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

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

Nikiforakis, N. (2010). Feedback, punishment and cooperation in public goods experiments, Games and Economic Behavior, 68, 689 -702.

Noussair, C., Tan, F. (2009). Voting on punishment systems within a heterogeneous group, CentER Discussion Paper N. 19, Tilburg University.

Ones, U., Putterman, L. (2007). The ecology of collective action: A public goods and sanctions experiment with controlled group formation, Journal of Economic Behavior and Organization, 62(2), 465-521.

Ostrom, E., Walker, J., Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible, American Political Science Review, 86, 404-417.

Ross, L., Nisbett, R.E. (1991). The Person and the Situation: Perspectives of Social Psychology, Philadelphia, Temple University Press.

Schnedler, W., Vadovic, R. (2011). Legitimacy of control, Journal of Economics and Management Strategy, 20 (4), 985-1009.

Sutter, M., Haigner, S., Kocher, M. (2010). The carrot or the stick? Endogenous institutional choice in social dilemma situations, Review of Economic Studies, 77 (4), 1540-1566.

Vesterlund, L. (2003). Informational value of sequential fundraising, Journal of Public Economics, 87, 627-657.

Xiao, E., Houser, D. (2011). Punish in public, Journal of Public Economics, 95, 1006-1017.

Xiao, E., Kunreuther, H. (2010). Punishment and cooperation in stochastic social dilemmas, mimeo, Carnegie Mellon University.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good, Journal of Personality and Social Psychology, 51, 110-116.