Peter Lewisch*

Altruistic Punishment: The Golden Keystone of Human Cooperation and Social Stability?

https://doi.org/10.1515/auk-2020-0011

Abstract: 'Altruistic punishment' (i.e., costly punishment that serves no instrumental goal for the punisher) could serve, as suggested by the pertinent experimental literature, as a powerful enforcer of social norms. This paper discusses foundations, extensions, and, in particular, limits and open questions of this concept—and it does so mostly based on experimental evidence provided by the author. Inter alia, the paper relates the (standard) literature on negative emotions as a trigger of second party punishment to more recent experimental findings on the phenomenon of 'spontaneous cooperation' and 'spontaneous punishment' and demonstrates its (tight) emotional basis. Furthermore, the paper discusses the potential for free riding on altruistic punishment. While providing valuable insights into the understanding of social order, 'altruistic punishment' is thus not the golden keystone of social stability.

Keywords: public good games, altruistic punishment, intrinsic disutility of punishment, spontaneous punishment effect, review of punishment decisions

1 Introduction

The problem of social order is perhaps the paramount topic in the social sciences. How can social order emerge? How can it thrive? And, how can it protect itself against unravelling? To the extent that selective incentives are needed for compliance, what would the role of punishment be and who would carry it out? These are difficult questions—thus far, the social sciences have provided only partial answers.

This paper discusses specific aspects of punishment, namely of costly punishment with no instrumental purpose (so called 'altruistic punishment'). This kind of punishment (i.e., punishment for punishment's sake, however intrinsically

***Corresponding author: Peter Lewisch,** Department of Criminal Law and Center for the Economic Analysis of Law, University of Vienna, Vienna, Austria, e-mail: peter.lewisch@univie.ac.at

motivated) is carried out even when there is no future in which to invest as well as by individuals who are both affected and unaffected by the underlying harmful act. Why would such a strange type of punishment be of interest at all for the social sciences? Because it could deepen our understanding of a phenomenon that may provide a crucial missing element in explaining the astonishingly high degree of cooperation and compliance that we see around us. Punishment, embodying the paradigmatic 'negative incentive' for rule compliance, is, when it comes to dispersed harm, in itself a (second order) public good, which would not be provided in a sufficient degree on a voluntary basis—unless there is an additional driver for this punishment beyond an instrumental purpose, namely the aforementioned 'altruistic punishment' (covering a broad range of possible non-instrumental motivations, including concerns for justice or fairness). To put all this into perspective, the current paper briefly recasts the underlying discussion in terms of prisoners' dilemma interactions. In addition, it sketches out, in a nutshell, the reasons for costly punishment and the conceptual foundations of 'altruistic punishment', and discusses, in its main part, recent research that sheds some light on limits and open questions surrounding the phenomenon of 'altruistic punishment'.

2 Prisoners' Dilemmas and Public Good Games, Focussed and Dispersed Harm

In the (economically inspired) social sciences, the problem of social order has been mostly discussed through the lens of the 2 by 2 prisoners' dilemma game (played both as a 2-person game 'without environment' or embedded in larger groups) as well as its large-number-equivalent, the public good game. In a public good game (used on several occasions later in the text), players receive an initial endowment that they can either keep for themselves or invest in a 'group-project' for the (equal) benefit of all players involved (contributors and non-contributors alike). Whereas the overall (social) optimum would be reached if everybody invested their full endowment, it is individually rational to withhold one's own contribution and take a free ride on one's peers.[1]

1 Typically, this game is played with 4 players; and contributions are doubled by the experimenter and then equally split among group-members. Assuming an initial endowment of 10, full investments by everybody would yield a 'cake' of 80 (10 x 4 x 2) and an individual share of 20. In contrast, rationality would dictate withholding one's own 10 and taking a share of the cake—if all other players contribute fully, this would amount to an additional payoff of 15 (i.e., (10 x 3 x 2)/4), amounting to a total payoff of 25.

It is worth noting that the prisoners' dilemma (henceforth PD), in all of its variants, is confined to the categorical choice between cooperation and defection[2] and does not, in itself, provide for an independent ('external') punishment option: possible selective (positive and negative) incentives emerge as a by-product of choices ('cooperation' or 'defection') within a potentially productive interaction. In the 2 by 2 case, whether in isolated interaction or embedded in a larger environment, the positive and negative effects of any strategy focus on the other player; in public good games, the effects of cooperation and defection are dispersed across the entire group.

In a one-shot interaction, defection is the strictly dominant choice in PDs and public good games, regardless of what the others do. Since most individuals follow this reasoning, cooperation is doomed to failure. In finitely interated interaction, backward induction would likewise lead to overall defection. The experimental evidence (provided by Voluntary Contribution Mechanism experiments, reported later on in this paper) shows that, while people are initially willing to contribute voluntarily, these contributions decrease over time (without any internal mechanism available to bring cooperation back to earlier/higher levels).

As long as players in public good games can only choose between cooperation and defection (with the possible additional exit-option allowed in a certain variant of the game), the effects of either choice are indiscriminately spread amongst all group members[3] , affecting co-operators and defectors alike and limiting de facto the maintenance of cooperation.[4] What is needed to stabilize cooperation in public good games is a selective negative incentive that focusses only on defectors. One possible tool for this goal would be the 'exclusion' of defectors from the group (see e.g., Danneberg et al. 2018). Exclusion is similar to punishment, but, at least in the realm of the mobility of our modern times, it has lost most of its earlier scare, is difficult to handle in practice, and enormously costly in terms of monitoring efforts. The alternative tool in the 'institutional toolbox' is a selective disincentive targeted directly towards wrongdoers and imposing a disutilit on them as a response to their defection: punishment.

2 While still allowing for different degrees of cooperation, namely different amounts of contributions in public good games.

3 See Lewisch 1995, 38 and 59 and, in particular, Vanberg/Buchanan 1988.

4 Similarly, an exit-based strategy of conditional cooperation cannot work in public good games. If disappointed co-operators leave a public good game, the environment (in terms of a reduced level of cooperation) deteriorates for all remaining players—defectors and co-operators, alike.

3 Punishment: Rational and Altruistic

3.1 General

As discussed in the previous section, in the world of PDs and public good games, cooperation is fragile, whereas punishment, as a distinct disutility imposed on the wrongdoer ex post, may serve as a tool to avert the breakdown of cooperation and stabilize social order. Punishment may be needed all the more if we broaden the picture realistically beyond the narrow boundaries of PDs and public good games and allow potential wrongdoers to negligently and/or intentionally inflict harm on others (= impose on them a negative externality) outside of a potentially productive interaction on a mere 'hit-and-run-basis' (theft, assault, murder, etc). The imposition of a specific sanction, a penalty, on the wrongdoer after the deed would selectively address the violation and provide incentives ex ante to discourage the commitment of such acts.

Punishment, however, is costly (e.g., Lewisch 1995, 85). Punishment in itself does not make good on the externality imposed by the wrongdoer. Rather, it by definition exceeds compensation and imposes a certain disutility on the trespasser that, as such, does not translate into a material benefit for the punisher. Costs of punishment typically consist in the use of resources for purposes of searching out wrongdoers, bringing them to justice procedurally, and enforcing the sanction (in a developed legal system: 'the verdict') against them.

In a one-shot interaction, punishment would necessarily be a waste in terms of ex ante incentives because, with no future ahead, punishment cannot buy the punisher anything in terms of accomplishment of instrumental goals whatsoever. A possible wrongdoer can exploit the costliness of punishment because, if punishment is wasteful after commission of the deed, a respective threat would not be credible. Emotional reactions (e.g., punishment triggered by anger), though wasteful in their own right, can work as a 'precommitment device', such that in the light of the imminent risk of an emotionally driven response, the potential wrongdoer may refrain from the act.[5]

If we expand the time-horizon and allow for future encounters, punishment turns into a potentially rational investment in deterrence, provided that there is

5 The classic reference is Frank 1988, which, however, apart from the preface (ix and x) provides a limited treatment of genuine punishment issues. Still, the preface condenses the main message: 'If people expect us to respond irrationally to the theft of our property, we will seldom need to, because it will not be in their interests to steal it.' See also the pertinent discussion in Lewisch 1995, 92 and from a psychological viewpoint, e.g., Gollwitzer 2007.

a chance to thereby influence the behavior of the relevant actors (i.e., potential trespassers). The decision by potential punishers to actually carry out costly punishment depends on their individual cost-/benefit-calculus. The individual will carry out punishment (as a necessary but not sufficient condition) only to the extent that expected benefits are higher than expected costs. If the inverse is true, the violation will still remain rationally unenforced. So far—so good.

If the wrongdoer commits a deed with dispersed harm (as is, e.g., the case in public good games), the above calculus remains the same. Potential punishers' decisions to punish will depend only on their own costs and benefits, disregarding the benefits to others. The individual will, therefore, not carry out punishment even if it was beneficial on an aggregated (social) level, if it is individually not cost-justified, because costs would focus on the individual decision maker, whereas benefits are dispersed. Even if the benefits of individual punishment exceeded its cost, it would be preferable in a public good game to take a free ride on the punishment activities by the other victims. However, since most individuals follow this reasoning, punishment will not be meted out to an efficient degree. Generally speaking, in cases of dispersed harm, punishment by those affected by the deed (and all the more, by observers) is in itself a 'second order' public good. The problem of enforcement in these settings is, therefore, typically a problem of chronic underenforcement.[6]

It is here where the 'behavioral/experimental' contributions[7] come in. They start with the observation that, empirically, we observe a high degree of order and rule compliance around us, more than one would expect if punishment was only enforced on 'traditional rational choice grounds'. There are basically two possible explanations for this high degree of cooperation and compliance, namely either that 'voluntary cooperation/compliance' (i.e., 'unenforced compliance') is actually higher than rational choice analysis would suggest and/or that the high degree of cooperation/compliance is the (rational) response to high levels of (anticipated) punishment that include punishment that is not motivated by instrumental goals and that, overall, exceeds punishment levels, as predicted by standard rational

6 Typically, but not necessarily. If sufficient deterrence requires punishment only by a certain number, say, one of several (or of many) potential punishers, uncoordinated punishment efforts (even though each cost is justified for the individual punisher) may both be wasteful in terms of overall resources spent and excessive in terms of the punishment actually imposed (see Lewisch, 1995, 130).

7 The respective literature has, however, not claimed itself to be formally part of 'behavioral economics'; the term is used here just as an abbreviation for economic research that systematically investigates human behavior that seemingly goes beyond the predictions of classic 'homo oeconomicus analysis'.

choice analysis. The pertinent behavioral literature, and therefore also this article, focusses primarily on these punishment aspects.

3.2 'Altruistic' Second- and Third-Party Punishment

According to the standard terminology, punishment carried out by individuals affected personally by the negative external effects of the underlying action is called 'second-party' punishment (even if the second party is only one of several 'victims', as is the case in public good games, where punishment also benefits other unrelated victims). Conversely, punishment by an unrelated person (observer) is referred to as 'third-party' punishment.[8] In the literature, costly punishment is termed 'altruistic' if it does not yield any material gain to the punisher (e.g., Fehr/Gächter 2002, 137). Altruistic punishment, hence, does not serve any conceivable instrumental ('consequential') purpose for the punisher[9], as is most prominently the case in a one-shot interaction. The main interest is with 'third-party altruistic punishment', i.e., with punishment provided by observers/bystanders with no instrumental objective, because such punishment could conceivably assume a crucial role in providing enforcement in an otherwise fragile environment. Regarding its role in providing the 'cement of society', 'altruistic punishment' (i.e., punishment beyond what is individually rational) is of particular importance when it comes to enforcing 'solidarity rules' (with the negative effects of the violations being distributed across an entire group without discrimination), where punishment is a (second order) public good.

Whereas terminological questions are not of interest here, the substantive questions involved warrant attention—and also some reconsideration. If punishment is called 'altruistic' because it does not provide any good to the punisher (e.g., punishment in a one-shot interaction), it may likewise not be of any instrumental value for the punisher's peers (and, therefore, is actually not motivated by any 'concern for the other'). Infact, to the extent that the term 'altruistic' does not go beyond description of punishment as 'instrumentally useless', as is the case in a one-shot interaction, one may question this connotation altogether.[10] The (somewhat surprising) use of the term 'altruistic' in the literature results from the

8 In case of third party punishment, the positive effects of punishment are conceptually fully external.

9 For a lucid critical discussion of the concept of 'altruism' in this respect, see Leist 2005.

10 Such instrumentally useless punishment behavior may, however, be motivated by an act committed against others (so that the punisher is a third party) or against the punisher her-/himself (= second party).

repeated public good games by Fehr/Gächter (2002, 137): in order to avoid direct reciprocity or reputation effects, the group composition was changed after each encounter so that subjects would never interact twice. Since, however, future group members may (indirectly) benefit from punishment in previous rounds insofar as the punished subjects, having experienced the disutility of punishment, could possibly alter their behavior due to a certain learning effect, such punishment may—very indirectly[11]—be called 'altruistic',[12]

Be this as it may, why would someone carry out costly 'instrumentally useless' (i.e., in this sense, 'altruistic') punishment? Behavioral contributions share the view that traditional rational choice accounts of punishment have overlooked a relevant factor in the explanation of punishment, 'something' that contributes to the aforementioned, higher than predicted degrees of voluntarily provided punishment.

Perhaps the most straightforward way to define this extra-something is 'punishment for punishment's sake', i.e., punishment based not on consequential considerations but rather on considerations of fairness, justice, or deserts. Conceptually, punishment can thus be seen not (only) as an investment decision, but as a consumption activity, where people, when punishing despite an insufficient investment reason, are willing to 'buy justice' simply because they feel better with individually provided punishment than without. Their punishment may thus be grounded on 'revenge/just deserts/moral wrong'-considerations, both when they are directly affected as a (at least partial) immediate victim or when they are observers witnessing the deed against third parties.[13] In this light, 'altruistic punishment' does indeed increase the punisher's utility; thus, the punisher is willing to incur the respective cost for this purpose ('costly altruistic punishment').

11 Very indirectly indeed because, on rational grounds, each game is a fresh game with new players—if the rounds are separable, it would rationally not make sense to play the game based on previous outcomes. Note, moreover, that the term 'altruistic' in the Fehr/Gächter paper is not used in an instrumental sense (such that it is motivated by the desire to influence the future behavior of the punished individuals for encounters with other individuals), but that the authors themselves interpret their results such that negative emotions would trigger this—only in its effects 'altruistic' – punishment.

12 Along similar lines, Fehr/Fischbacher 2003 treat rejections in the ultimatum game as an altruistic act because, under the experimental design used, proposers would meet different responders in ten successive rounds—thus, the proposer could use his/her experience form previous encounters for the next round.

13 It seems straightforward to assume though that, other things equal, people would be more inclined to engage into such punishment for its own sake, if they themselves are directly affected by the deed than if they are only a bystander/observer. See, however, for fairness violations the inverse experimental finding reported in FeldmannHall et al. 2014.

The term 'altruistic' does not fit well in this respect. Rather, punishment is meant to convey an intrinsic utility to the punisher that is independent of any instrumental purpose. Altruistic punishment may thus be alternatively characterized as 'intrinsically motivated punishment'.

The question, however, then shifts to the underlying mechanism, whereby people derive utility from punishment. One can assume a wholly agnostic position, arguing that punishment in these situations is just a matter of 'tastes' (and also tastes for punishment). Going one step further, the desire for punishment could conceptually be the result of an immediate ('affective') impulse for revenge or retribution (see Mackie 1982), a spontaneous or also deliberate desire to correct perceived unfairness[14], concerns for justice, or, in a more pronounced way, also the result of some 'rule-internalization' in the sense that punishment ought to be done (and that the potential punisher would feel that s/he did not provide it). Again, 'altruism' appears to imply a certain mis-characterization regarding either of those driving forces.[15]

It seems fair to say that the phenomenon of 'altruistic punishment' is, at least partially, a cultural phenomenon[16] that is embedded in a certain (institutional or 'quasi-institutional') context[17] of social norms[18]. In light of that context, one may well acknowledge that humans "possess a deeply rooted social interest" (Leist 2005, 168) in terms of a need of social relations as well as that, regarding the phenomenon of 'altruistic punishment', the single victim or observer views the underlying harmful act as a 'social disturbance' (a violation of social norms) that calls for costly rectification and also that s/he perceives that (possibly costly) rectification as the right thing to do.[19] Recent contributions have increasingly

14 See generally, Fehr/Schmitt 1999.

15 The promotion of justice and fairness would typically not be considered an altruistic act, because it (at least, if the punisher cares about these concepts) fosters necessarily also the well-being of the punisher.

16 For the very broad discussion on genetically or culturally predisposed altruism, in particular with reference to anthropological insights, see as examples, Fehr/Fischbacher 2003, 788; Alexander 2005; Nowak/Sigmund 2005; Boehm 1999; 2008; 2014; Sterelny 1992; 1996; 2016; Stich 2007, and Warneken 2013, all with further references, also to the underlying socio-biological literature.

17 Regarding the much contested question of possibilities and limits of an evolutionary emergence of altruistic punishment, see Fowler 2005, with references, proposing an evolutionary model studying the dynamics of a population with punishers. He shows how altruistic punishment can emerge and persist in a world with incentives not to contribute and not to punish non-contributors, and how it would dominate cooperators, defectors, and non-participants.

18 See e.g., Kimbrough/Vostroknutov 2016 and 2018. For a commentary on cognitive and neural foundations of social norms and their enforcement, see Buckholtz/Marois 2012.

19 Irrespective of any terminological questions, this phenomenon is worth being studied.

emphasized the need for an overarching approach that integrates opportunity-seeking and rule following behavior;[20] an approach that could also be fruitfully applied, in future research, to such questions of punishment.

3.3 Experimental Findings and Their Limits

The proof is in the eating of the pudding. The proof for 'altruistic punishment' is empirical, to be provided not by paper & pencil-experiments with 'cheap talk', but, at least primarily, by incentivized experiments, where 'real money' is at stake and subjects have to purchase punishment ('punishment points') out of their own endowments. The pertinent literature has built its claims on such (also neuro-economic) experiments showing that people do indeed punish in settings where punishment cannot be explained on consequentialist/instrumental grounds.[21]

As for the possible drivers of such 'altruistic punishment', the relevant neuro-economic research suggests the 'sweetness of revenge' (as experienced in the brain reward center) serves as a crucial driver of such punishment.[22]

Fehr and colleagues have suggested that "individuals derive satisfaction from the punishment of norm violators" (Quervain et al. 2004, 1254), that they "seem to feel bad if they observe that norm violations are not punished, and they seem to feel relief and satisfaction if justice is established" (1254), and that neurologically the activation of the dorsal striatum reflects the anticipated satisfaction from punishing defectors, conferring a 'sweet taste of revenge' to the punisher.[23] On the sociological meta-level, they claim no less than 'altruistic punishment' to be "a key force in the establishment of human cooperation" (Fehr/Gächter 2002, 139) and as serving as a "decisive force in the evolution of human cooperation".

As for second-party (altruistic) punishment, the authors emphasize the role of emotions as a trigger for punishment even without instrumental purpose. "Taken together, these observations are consistent with the view that emotions are an important proximate factor behind altruistic punishment"(Fehr/Gächter 2002,

20 See e.g., Kliemt 2020 and Vostroknutov 2020 for further references.

21 Some contributions in the literature (Guala 2012) argue that experimental results are not sufficient to demonstrate that costly punishment would sustain social cooperation in the real world and emphasize the lack of pertinent field experiments. This lucid criticism is well taken but in the end appears to underestimate concerns for fairness, justice, or simple rule-obedience as drivers of individually provided (though possibly collectively required) punishment in the real world. See also Lewisch 1995, 136, for a discussion of anthropological evidence for collective enforcement.

22 For the 'emotional opposite'—he 'warm glow of giving'—see in particular Andreoni 1990.

23 See also Sanfey et al. 2003.

137).[24] In public good games, therefore, "most punishment acts would be expected to be executed by above-average contributors and imposed on below-average contributors".[25]

Moreover, the authors interpret altruistic, third-party punishment as the keystone of social stability. Whereas, at least in the light of a specific line of experiments, "the sanctions of a single third party were not" strong enough to make norm violations unprofitable so that "more than one third party is needed to enforce the norm … this condition is probably met frequently in real life" (Fehr/Fischbacher 2004, 85). Viewed as such, "altruistic third-party sanctions are likely to be powerful enforcers of social norms" (Fehr/Fischbacher 2004, 85).[26]

In the remainder of this article, I will argue that, whereas the 'phenomenon of altruistic (second- and third-party) punishment' provides important insights for an explanation of social order, its impact may be less straightforward than predicted by the aforementioned literature for three reasons; reasons that are not derived from limitations of real-world applications (see again Guala 2012) but which are already present conceptually in the experimental world itself. I will discuss these reasons in the context of experiments that I have conducted together with colleagues, while linking these experimental findings to other developments in the literature.

- First, the aforementioned neuro-economic view suggests that punishment is empirically more prevalent, because—contrary to traditional analysis—it is not only costly but also conveys 'enjoyment' to the punisher (namely, the 'sweet taste of revenge'). However enriching this view may be, it may not capture the full picture. On the 'intrinsic level', there may exist not only an 'intrinsic utility of punishment' but also an 'intrinsic disutility of punishment' that could ultimately constrain the amount of punishment actually meted out (and, therefore, also limit its incentives).
- Second, the pertinent literature views 'second-party punishment' as a widespread and powerful enforcement device that is, in public good games, emotionally triggered by disappointed contributors. Again, this view provides valuable insights. However, it also (partially) disregards the possibility (i) that the relationship between above average contribution and punishment may

24 "Our results suggest that free riding causes strong negative emotions and that most people expect these emotions." (Fehr/Fischbacher 2002, 137)

25 "Negative emotions are the proximate cause of the punishment … ." Sanctions "by second parties directly harmed were much stronger than third-party sanctions" (Fehr/Fischbacher 2002, 137).

26 This view is not universally shared. See Guala 2012 who calls for "a re-orientation away from its current obsession with costly punishment".

be less direct than assumed and (ii) that the described phenomenon may be limited only to certain sub-groups of people.
- Third, regarding third-party punishment, the literature assumes, on the one hand, that bystanders would be, in principle, willing to enforce 'distribution and cooperation norms' without individual instrumental benefit (driven strongly by the sweet taste of revenge) and also, on the other hand, that such third-party punishers provide something similar to the long searched or 'keystone for social stability'. This is because, typically, there are many observers/bystanders as possible candidates for such punishment present, so that the probability is high that a violation will not remain unenforced. However, this view implicitly assumes that the respective 'sweetness of revenge or satisfaction of a desire for just desert' would only be experienced by the 'third-party punisher' him-/herself. If, in turn, the situation was such that bystanders who do not punish themselves but only watch the punishment carried out by other third parties could also experience a similar satisfaction out of such punishment, then the opportunity to take a free ride on costly punishment, namely enjoying the satisfaction of punishment without meting it out oneself (= carrying its cost), would re-emerge.

In the subsequent section, I will discuss these points in sequence.

4 Open Questions

4.1 Intrinsic Disutility of Punishment

Traditional economic analysis of punishment has always emphasized the costliness of punishment. The higher the price of a commodity (and thus also: the price of punishment), the less of it will be purchased. This is also true for 'altruistic punishment'. In fact, 'altruistic punishment' provides a good illustration of the agnostic self-understanding of economics, holding that even though we do not know why people engage in a certain activity, its frequency will decline if the price of the activity increases. Experimental findings show exactly that.

The "intrinsic disutility of punishment" (Buchanan 1975, 130) describes a second category of punishment costs that goes beyond the 'regular' resource-components (for investigations, proceedings, etc.) required for penal enforcement. It captures the negative emotions by the punisher associated with the deliberate infliction of a bad onto someone else. 'Intrinsic disutility' accounts for the straightforward fact that people normally do not like to harm another being. Peo-

ple typically do not enjoy hitting someone else in her/his face, scratching her/his skin, or breaking her/his bones. Mothers dislike 'punishing' their misbehaving children ('Wait until daddy comes; he will punish you'). The hangman has always been the outcast of society. Some people do not even like to squash mosquitos but instead prefer to shy them away. As such, this category covers exactly the opposite of emotional enjoyment or satisfaction derived from the very act of punishment, as suggested by the behavioral literature. Where does the concept of the intrinsic disutility of punishment come from? As Buchanan (1975, 133) states:

> The basic costs of punishment are subjective, and these can best be conceived in a utility dimension. The imposition of penalties on living beings, whether or not these beings have violated law, causes pain, utility loss, to the normal person who must, directly or indirectly, choose these penalties. 'Punishing others' is a 'bad' in economic terms, an activity that is, in itself undesirable, an activity that normal persons will escape if possible or, failing this, will pay to reduce.

One immediate consequence of this intrinsic disutility of punishment, as just outlined, therefore, is that people have a direct and immediate incentive to avert the respective cost by either avoiding the infliction of punishment altogether or by punishing in an overly lenient manner (and, in fact, too leniently in comparison to what they themselves would consider adequate from an ex ante perspective).

The concept of the intrinsic disutility of punishment is indeed well taken and, for whatever reason, underrated—not to say: hardly noticed—in the literature. It is (at least as such) taken up neither by the classic economic approach nor—and quite surprisingly so—by the mainstream behavioral/experimental literature. In fact, to the extent that this emotional component increases the cost of punishment (either simply as an additional cost-component or as an emotional factor reducing the otherwise present enjoyment of punishment), the economist would predict less punishment to be carried out ('consumed').

The underlying phenomenon of something such as 'punishment of others as an emotional bad' has, however, not remained unnoticed. On the one hand, Carlsmith/Wilson/Gilbert (2008) have, under the telling title 'The Paradoxic Consequences of Revenge', reported the experimental finding that punishers in a standard punishment condition (following a public good game), while expecting emotional relief from the act of punishment, actually experience a deterioration of their emotional status.[27] As long as people are unaware of these negative emotional effects of punishment, the respective costs would not be decision-relevant. In itinerated interaction, one would, however, predict a learning effect regarding

27 This negative emotional status can be well seen as a manifestation of the aforementioned intrinsic disutility of punishment.

the disutility associated with the act of personal punishment and, thus, an induced reduction of punishment.[28]

On the other hand, the recent empathy-literature has, on a more general level, emphasized that people, as social beings, "learn and adapt … behavior to avoid harm to others" and that even, in contrast to what early experiments suggested (Milgram 1963, 371), "interpersonal harm aversion has been proposed to form the basis of prosocial behavior and morality" (Lengersdorff et al. 2020, 7286; Crockett 2013; Crockett et al. 2014). At least in certain settings, individuals are willing to spend more money to protect others from pain than themselves ('hyperaltruistic choices', Crockett et al. 2014) and people appear to be particularly good at learning for the benefit of others ('hyperaltruistic learning', see again Lengersdorff et al. 2020). Coming from the very different angle of experimental studies on utilitarianism, Capraro et al. (2019) have contributed the finding that the priming of intuition in experimental moral dilemma settings seems to favor a refusal by the subjects to inflict harm 'for the greater good'. In the light of this literature, it seems plausible that this concern for the other is still present when it comes to actually inflict pain and suffering onto an offender in the course of punishment.

This brings us to the point: Taking into account the possible intrinsic positive emotions of punishment is, by itself, not sufficient to predict higher levels of punishment if this emotional side is janus-headed, with a 'pleasant' joyful face of sweet revenge, on the one hand, and an 'uneasy' face of 'intrinsic disutility' on the other.[29] What matters would be the overall outcome of these two orthogonal (intrinsic) factors.

In light of these two conceptually distinct emotional ('intrinsic') components of punishment, the research question emerges under what conditions either emotional frame would possibly dominate the other. This question has not yet been answered and, as far as I can see, has not even been posed in the pertinent literature. In the realm of this analytical article, I can only provide a first sketch. It appears as if the overall personal 'emotional' experience of punishment, in a specific case, depends on the 'frame' through which the potential punisher perceives the trespasser and his/her deed. If this frame is guided more by concerns for deterrence, just desert, or even empathy with the victim, enjoyment-aspects of punishment are more likely to dominate. In contrast, if the frame is guided

28 This disutility of punishment is, therefore, not necessarily already taken into account ('priced in') in the experimental punishment decisions.

29 Moreover, since the 'sweetness of revenge' embodies a 'hot' emotional feeling and the unwillingness to inflict pain seems to be driven by intuition, neither emotional status is of a deliberative nature.

by concerns for the wrongdoer him/herself and perhaps even empathy towards him/her, the intrinsic disutility component will instead prevail.

Viewed as such, the question is of interest which factors determine the emergence of which frame. This is a complex question. There is no reason to assume the relevant frame to be stable/static or to emerge in a linear way. Rather, the relevant frame emerges out of a 'struggle' of both punisher-related factors and situation-/trespasser-dependent factors. Within this struggle, the psychological 'identification' (either with the victim or the wrongdoer) that the potential punisher develops and the degree of moral outrage caused by the underlying act apparently play a decisive role (see generally Lewisch 2004). Saying that the emergence of the relevant frame is situation-dependent, moreover, directs attention to the possibility of influencing certain factors that contribute to the concrete shaping of the frame, of emphasizing some factors and, at least to a certain degree, even of 'manipulating' them. And indeed, very much along these lines, penal proceedings in general and criminal trials in particular can be seen as the showdown of this struggle for influencing the decision-maker (= judge) to ultimately adopt the 'appropriate' frame. On a more general level, the major consequence of this discussion remains that identifying components of intrinsic utility in punishment is only 'half the story', if punishment may also involve an intrinsic disutility that can (partly) offset or even dominate the 'sweet taste of revenge' (and also the desire for just desert).

4.2 Second-Party Punishment as a Spontaneous Impulse?

This topic is of a very different kind. Second-party punishment concerns punishment by a victim (again, if there is no future in which to invest). The case of punishment as an emotional response to a wrong that the 'victim-punisher' has experienced on her- or himself has an immediate appeal. While some contributions in the literature address second-party punishment in the context of a 'one wrongdoer versus one victim' setting, most contributions discuss second-party punishment in terms of punishment by one (of several) victim(s) in a public good game. Thus, the setting is one of dispersed harm.

The interest in second-party punishment is twofold. It emphasizes the role of negative emotions for punishment and links those emotions to the frustration of being exploited in public good games as an above-average contributor. The aforementioned behavioral literature suggests a causal relationship between a certain preparedness towards (conditional) cooperation in the underlying public good game, leading to above-average contributions, which, in case of exploitation

by free riders, would trigger anger and punishment.[30] Although this literature concedes the heterogeneity of human actors, it still makes generalized predictions regarding punishment. In particular, the argument insinuates that individuals who would voluntarily cooperate in public good games would also punish even if there is no future to invest in.

There is nothing inherently wrong with such an approach. Quite to the contrary, emotional frustration and anger may well have their place in explaining second-party punishment, in particular in public good games, where the (prospective) punisher is among several 'victims' of the free-rider. However, the underlying mechanism of this punishment is likely to be less linear than assumed in the pertinent literature. On the one hand, the question is of interest whether the same factors that prompt a player to cooperate in the public good game also determine his/her punishment decisions. On the other hand, the question remains whether the aforementioned negative emotional status of the potential punisher follows, upon learning about the outcome of the public good game, from her/his dispositional ('pro-social') traits or rather simply 'situationally' from his/her (higher) investment in the underlying public good game. The differences to the more crude 'classical' behavioral approach are subtle but existent—and they shed some light on the concrete ingredients for relevant second party altruistic punishment to occur.

Let me discuss these questions with a small detour, namely by starting (briefly) with voluntary contributions in the public good game in terms of 'spontaneous cooperation' and then turning to the voluntary provision of 'altruistic' punishment likewise in terms of 'spontaneous punishment'. Measuring the relationship between the size of contributions in the public good game and the respective 'reaction time' for such a contribution, David Rand and colleagues (2012) have suggested a 'spontaneous cooperation effect': the quicker the decision, the higher the contribution and vice versa.

The aforementioned finding and even more so its interpretation by the authors have been intensively debated and also criticized in the literature. The pertinent discussion has, however, been begging the question, whether a comparable phenomenon would exist also for punishment: If players had the opportunity to impose different amounts of costly punishment on their peers after learning about the results of the public good game, would they (also) exhibit the said inverse relationship such that punishment amounts decline over time? The perhaps surprising

30 It is not entirely clear whether this literature assumes the preparedness to cooperate as given and conceives the propensity to punish as a consequence of the frustration with free riding or whether it assumes cooperators to be strong reciprocists in the sense that they cooperate in light of their punishment option.

answer is: yes, on average a negative correlation exists between measured decision time and contributions to punishment (Mischkowski et al. 2018). The shorter the reaction time, the higher, on average, the punishment ('spontaneous punishment effect').

On that 'macro-level' (in terms of aggregated average outcomes), both findings, spontaneous cooperation and the spontaneous punishment effect, are particularly impressive. On the 'micro-level' of individual motives[31], however, the question is of interest, which factor—or rather which 'dispositional' or 'emotional' status of the contributor/punisher—would be crucial in triggering voluntary contributions in the public good game and/or voluntary punishment, respectively. Is it really that people are spontaneous cooperators and spontaneous punishers, as these findings at first sight may suggest? Closer observation of the 'micro-motives' behind the macro-results warrants caution.

In the most recent literature[32], a critical view regarding the assumption of a general spontaneous cooperation effect prevails. This literature (e.g., Mischkowski 2020, 76, also for further references) emphasizes 'decision conflict' (rather than natural goodness of man or spontaneous cooperation) as the primary predictor of decision times and suggests to explain differences in decision times along a U-shaped pattern with low decision times for decision extremes on either side (pronounced social value orientation or pronounced self-interest).

Still, as shown by Mischkowski/Glöckner (2016), a pronounced social value orientation does trigger a spontaneous cooperation effect. The authors could demonstrate that the Rand-effect of spontaneous cooperation is not a universal phenomenon, but that it is driven by a particular sub-group only, namely the 'pro-socials' (as determined in the experiments applying the standard 'social value orientation' test). For all others, there is no comparable relationship between the amount of contributions and the reaction time. For the pro-socials, however, the effect is so pronounced that it drives, in average terms, also the overall aggregated result.

The question then is whether a comparable 'micro-motive' would also exist regarding punishment, namely whether a particular subgroup, identifiable possibly again by their value orientation (possibly again the pro-socials), would drive

31 See as the basic reference for analysis along the lines of 'micromotives and macrobehavior': Schelling 1978.

32 The literature on the subject is abundant, but see in particular Evans et al. 2015; Evans/Rand 2018; Krajbich et al. 2015; Rand 2016; Rand et al. 2016; Yamagashi et al. 2017; Capraro 2019, and Andrighetto et al. 2020.

the aforementioned spontaneous punishment effect. The respective experiments[33] have shown (Mischkowski et al. 2018) that, whereas again one particular sub-group drives the spontaneous punishment effect, it is not the same subgroup as in the public good game. Infact, there is no significant relationship between pro-sociality and reaction time regarding punishment.

The spontaneous punishment effect is driven by one particular sub-group—a sub-group, however, that is defined by twofold ('combined') characteristics, namely (i) above average contributors who (ii) are in a negative affect. Again, the effect for this specific subgroup (above-average, highly upset contributors) is strong enough to generate, on average, a similar pattern for the entire group on the aggregate level. When disaggregating the results, however, there is no direct effect of pro-sociality on amounts of punishment in time.[34] Moreover, not everybody (and even not every pro-social) is an above average contributor; and not every above-average contributor reacts with anger on the other less-than-average-contributors. And it is only the joint characteristics of highly upset, situationally above-average contributors that generate the spontaneous punishment effect.

Where does all this leave us with the more general question of altruistic second-party punishment? To the extent that the above experiments have revealed the phenomenon of decreasing punishment investments with increasing time, one could argue that 'spontaneous punishment' provides something like an immediate, natural, and powerful response to an observed defection. Closer observation, however, has revealed that this phenomenon applies only to the well-defined subgroup of the highly upset above-average contributors.

There are two points worth pondering.

First, regarding the question of 'pro-sociality' of the punishment: in the public good games considered in the aforementioned experiments, affective rather than pro-social[35] motives drive the decision to punish. On a more general basis, however, it appears debatable whether punishment would qualify as a pro-social act altogether (see also Mischkowski 2020, 83). After all, punishment consists in the deliberate infliction of harm onto another (i.e., an 'anti-social' rather than a 'social act'). In certain settings, however, punishment may derive its 'pro-sociality' from its context, say, by overcoming a second-order public good problem in cases of dispersed harm (namely, the shortfall in punishment due to its 'public good

33 Replicating, in passing, the spontaneous cooperation effect for pro-socials in the underlying public good game.

34 There is, however, an indirect effect via increased contributions in the public good game.

35 Pro-sociality is, in our experiments, not linked to spontaneous punishment; it is the negative affect following an above-average contribution (irrespective of motivation) that drives 'spontaneous punishment'.

nature'). In other settings, it may just appear as the expression of a self-righteous or even unduly vengeful attitude. It is also unclear whether, and possibly under what circumstances, a possible concern for retribution carries with it a social connotation. In that light, pro-sociality may well coincide with punishment in certain situations only.

Second. The results reported above caution an overly optimistic view regarding the overall effectiveness of punishment in such settings. If the main trigger of the spontaneous punishment effect is the simultaneous concurrence of higher than average contributions and of negative affect, the basis for such a punishment, as a general enforcement tool, is rather narrow. It relies on a sufficient number of contributors and, additionally, a sufficient percentage of co-operators with a sufficiently strong affective status. Whereas the phenomenon of punishment in general is more encompassing than the 'spontaneous punishment effect' and some incentives of punishment also stem from non-spontaneous (though smaller) punishment, the immediate impulse for punishment appears to be limited.[36]

Overall, the working mechanism underlying the spontaneous punishment effect is likely to be more fragile than assumed in the relevant literature.

4.3 Third-party Punishment, Free Riding and Institutional Reform

4.3.1 Free Riding in Altruistic Third-Party Punishment

The third, more fundamental point concerns 'third-party altruistic punishment'. Fehr and colleagues attribute to 'third-party punishment' a central role as an enforcement device to bring about rule compliance and social stability. In that respect, they appear to suggest implicitly that intrinsic enjoyment of punishment, in particular localized in the brains at the reward center, is something that only the punisher himself/herself can enjoy. In this light, there would indeed be no other way to enjoy this punishment than its actual in-person infliction onto another. But what if the punisher need not carry out punishment him-/herself to experience the positive (emotional) effects of punishment because those positive effects, at least

36 Regarding the timing of the punishment, the situation is complex. In general, an immediate punishment-response is likely to impress a wrongdoer more than a delayed reaction. In the aforementioned experiment, however, we talk about seconds not hours, days, or weeks. Further research would have to examine whether a comparable 'timing effect' is also present, if we extend the possible 'reaction period' considerably, say, as in real life settings, to weeks or months. For a general contribution of the timing of reactions, see Grimm/Mengel 2011.

partly, spill over to bystanders/observers[37] (say, when they are informed about such punishment or even witness it directly)? Under this condition, the punisher could enjoy the 'emotional' benefits of punishment without bearing the pertinent cost. This is, however, only another way to say that a potential to free-ride on altruistic punishment would re-emerge through the back-door, namely by being able to experience the satisfaction of punishment, while its cost is borne by others.[38] Such a re-emergence of the potential to free-riding on 'altruistic punishment' would conceptually undermine the enforcement power of third-party punishment. 'Altruistic third-party punishment' is then not the final keystone to overcome the underprovision of punishment as a second-order public good but, in itself, the object of free-riding (Lewisch et al. 2011).

The pertinent questions can be answered empirically. For this reason, we have conducted an experiment that followed exactly the set-up of the Fehr experiments on altruistic third-party punishment (= dictator game played between A and B with the potential punisher C having the option to impose costly punishment on A)[39] as the baseline scenario. This means that after the random assignment of roles, A may share his/her endowment with B (B remaining entirely passive, receiving whatever s/he receives). C is then given the opportunity to punish A, namely by spending 'tokens' from his own endowment to reduce A's payoff.[40] The new twist in this experiment was the introduction of a variant to this baseline-scenario ('in-group scenario') with a second potential punisher (C2) who on his/her turn could also punish A.[41] The expectations by the C1s about the punishment by C2s were collected on a non-incentivized basis. The question then is whether punishment

37 See the findings reported in Mendes et al. 2018 on the willingness of preschool children to watch the punishment of antisocial others.

38 Free-riding on punishment would not be feasible to the extent that the punisher derives some extra utility out of the personal act of punishment, say in terms of esteem or acclaim by his/her compatriots. This emotional extra bonus would, however, not fall under the category of intrinsic utility, but would be some kind of external (emotional) reward.

39 Since in this scenario, the underlying interaction between A and B is that of the split of A's endowment, one may well see C's behavior as a (costly) correction of an unfair behavior rather than „punishment' of some forbidden act. Still, the baseline experiment copied exactly this setting of the original Fehr-experiment. For punishment in a setting that may be interpreted as a 'stealing scenario', see the subsequent experiment reported under 4.3.2.

40 The experiment provided (not decisive for the points of interest here) two variants, one with a high cost of punishment (payment of 1 token leads to a deduction of 2 tokens from A's payoff) and one with low cost of punishment (where punishment is more effective, such that a payment of 1 token leads to a deduction of 3 tokens).

41 Technically, the experiment was performed under the strategy-method such that the Cs made their choices ex ante for each level of transfer from A to B. In the in-group scenario. The Cs were asked, being informed about a ½ possibility of being the only player C and a ½ possibility of

provided by punisher C1 declines, increases, or remains constant in light of the presence of this second (potential) punisher in comparison to the stand-alone case.

There are three possible ways in which the introduction of a second ('horizontal') punisher could affect the punishment of C1.

- There can be no effect. If punishment, in economic terms, was an 'emotional good', such that the enjoyment of punishment may only be consumed by the punisher her-/himself, the presence of a further ('second') punisher would not be of any relevance for the punishment choices by C1. C1 would simply disregard punishment by C2 because all that matters is the punisher's own 'in person' infliction of punishment.
- In contrast, a full 'crowding out' effect is conceivable: If punishment was an 'instrumental good' so that punisher C1 was interested only in A being punished, while not deriving any (additional) personal enjoyment out of the act of punishment, s/he would adapt his/her punishment behavior accordingly. (Expected) Punishment by C2 would fully crowd out the respective punishment by C1, such that C1 would reduce her/his punishment exactly according to the expected amount of punishment by C2.
- Finally, punishment can be a mixed good. A partial crowding-out would take place in the sense that C1 would somewhat, but not fully, account for the expected punishment by C2 and would partly reduce his/her own punishment.

There are two main findings to report.[42] The first finding is that free-riding on altruistic punishment occurs in a substantial, statistically significant manner. For all scenarios (and all types of punishment costs), average punishment per person is lower in the presence of a second potential punisher than in the stand-alone-case.[43]

The second main finding is that people in their role as potential punishers are very heterogenous regarding third-party punishment and possible free-riding (such

being one of two possible punishers, to declare their punishment for both cases, stand-alone and in-group, and for all transfer levels.

42 In passing, one may mention regarding the baseline-scenario that also in these experiments 'altruistic third-party punishment' occurred and that it was cost-sensitive (meaning that a higher cost for punishment decreased the amount of punishment actually meted out).

43 In these experiments, the comparison between the stand-alone and the in-group scenario was straightforward because, under the strategy method, potential punishers were asked to indicate their own (costly) punishment for either scenario ex ante (expectations regarding the punishment of the second punisher were elicited, on a non-incentivized basis, later on). Note that a separate problem of marginal costs and benefits exists in case of consecutive punishment because, after actual punishment by the first punisher, the second would have to decide whether marginal benefits from additional units of punishment are justified by their costs.

that, in addition to the no-punishment option, all three aforementioned cases of a possible influence of a second punisher actually exist). A considerable sub-group of around 45% followed the homo oeconomicus model; these individuals never punish, and the presence of a second punisher (one would say: naturally) does not affect this choice. The residual percentage covers those individuals who punish at least to some extent. Still, they show a very divergent punishment pattern. Only a (not negligible) minority of around 11% treats punishment as an 'emotional good' and punishes without consideration to the presence of a potential second punisher. For a group of comparable size (13%), a full crowding out takes place: that is to say, these individuals reduce their own punishment in exactly the same amount by which they expect punishment by the second new punisher. And for a third, larger group (24%), some crowding out occurs—they reduce their punishment somewhat but not to the full extent of the anticipated punishment by the C2s.[44]

Put differently, altruistic third-party punishment exists as a purely 'emotional good', but its incidence is quite limited (around 10%). Only for this key-group of 'altruistic punishers', the presence of a further punisher has no effect and 'free-riding on altruistic punishment' cannot occur.[45] As for the rest, either there is either no punishment at all or the amount of punishment provided depends on the presence of a potential second punisher (such that anticipated punishment by this second punisher crowds out, either fully or partially, punishment in a stand-alone case). In the overall, there is hence considerable potential for free-riding on 'altruistic punishment' by one's peers.

This potential may even increase, if we link these results with those obtained in the above reported experiment by Carlsmith et al. (2008). While ex ante both actual punishers and 'mere observers' believe that punishment of free riders in a public good game would meliorate their mood, the act of punishment worsens the mood of the punisher without generating a comparable negative effect on the observers. To the extent that people experience this effect, and hence the extra cost associated with personal punishment, the incentives to free-ride on the punishment by one's peers are exacerbated.[46]

44 The numbers do not add up to 100% because 7% were excluded for formal reasons.

45 It is, however, conceivable that some altruistic punishers, namely those whose enjoyment of punishment is linked to a certain retributive 'ideal', reduce their punishment because, for them, the prospect of a possible aggregate over-punishment on A is (emotionally) worse than the reduction in self-imposed punishment.

46 It is, however, also conceivable that at least some altruistic punishers may react to a shortfall in punishment by their peers due to free-riding and resume and reinforce personal punishment, as long as their net benefit of punishment (enjoyment minus worsening of mood) remains positive.

The findings of the aforementioned experiment are 'qualitative' in the sense that they identify different categories of punishers. However, they do not, and cannot, infer predictions as to the quantitive occurrence of punishment in society, unless one knows about the respective composition of the different types of punishers in a given population at a certain time. Moreover, we have not studied the effects of an enlargement of the respective group (and, hence, the number of potential punishers) on the willingness of the individual to carry out punishment herself/himself. Such enlargement would be irrelevant if 'altruistic punishment' was a genuinely 'emotional good' with the personal concern for retribution/justice being the only driver of this punishment. If, as is the case, the presence of a further punisher influences—namely reduces—the punishment behavior, then it also seems reasonable to assume that an enlargement of this group would reduce the incidence of individually imposed punishment (although generally the effects of group-size on cooperation are complex with even a possible curvilinear effect on cooperation in one-shot dilemmas, see Capraro/Barcelo 2015 and Barcelo/Capraro 2015).

Where do the aforementioned results on the potential for free-riding on altruistic punishment take us regarding a possible shortfall in overall enforcement (in terms of the negative incentives a potential wrongdoer faces) and, hence, with third-party punishment as a tool to provide rule compliance and stability in society? Incentives for potential wrongdoers depend not on the punishment of a single punisher but on aggregated punishment. This question is difficult to answer on the basis of the aforementioned experiment because—despite the decline in average individual punishment in the in-group setting—aggregated (expected) punishment increased as a consequence of two potential punishers being present: in all settings, on average, the sum of the punishment by C1 and his/her anticipated punishment by C2 exceeded punishment in the stand-alone case.

In principle, the amount of aggregated punishment depends on the number of punishers (i.e., group size) and the size of punishment per punisher. Therefore, even if individual punishment shrinked with group size, overall punishment could increase if the larger group size makes good for a decrease in individual punishment. Conversely, aggregated punishment will decrease with group size if individuals will reduce their individual punishment in larger groups more than what will be offset by the higher multiplier. Again, normative concerns for 'just desert' may limit amounts of individually imposed punishment if the punisher

otherwise presumes unduly harsh aggregated punishment (as the result of separate punishment decisions).[47] These complex questions require separate treatment.

Conceptually, however, the main result holds that free riding on altruistic punishment exists, that it exists to a considerable extent, and that, analytically, it carries with it the potential for a gross and systematic underenforcement of the relevant rules.

This is, however, not to say that individual punishment, namely even altruistic third-party punishment, could not be influenced, and reinforced, by appropriate institutions. It can. See the subsequent 'vertical' experiment in the next section.

4.3.2 Vertical ('Second-Instance') Punishment and Institutional Reform

The paper 'Third-party punishment under judicical review' (Lewisch et al. 2015) analyzes the effects that the introduction of an additional player (a second 'punisher') on a 'vertical level' ('appeals layer') generates on punishment. The role of this additional player is not simply to decide on the imposition of possible additional punishment as a linear add-on to the punishment already imposed by the first punisher but rather, similar to that of an appeals court, the new player has the last word on punishment of A. S/he enjoys full discretion and may modify the punishment decision of the first punisher in all directions. Modifications, however, are costly both for the second decision maker and the first punisher, whose decision is (in whatever direction) overruled.

The pertinent research question is threefold. The first interest is, whether people would at all engage in such costly corrections of other people's punishment in the absence of instrumental benefits ('altruistic checks on punishment'). Second, it addresses the question whether the introduction of a second 'vertical' punisher (the 'second instance') would influence the 'downstream' punishment (by the first punisher) and, if so, in what direction and to what extent. Third, it examines the effects of such institutional change on the underlying 'stealing' behavior by A (namely whether the As would, in any sense, anticipate changes in punishment by adaptions to the incidence of 'stealing').

The design of this experiment is similar to the aforementioned setting with the difference that, in the baseline scenario, the interaction between A and B is not a dictator game but a stealing scenario (with A having the option to take some of B's

47 Note, however, that at least in certain cases the 'social disturbance' caused by the same physical act may differ according to the number of the individuals affected so that higher aggregated punishment would not necessarily imply a unduly harsh penalty.

endowment). In this baseline 'Trial Treatment', the observer C again has the option to impose costly punishment (in two possible amounts, soft/harsh) on A. The 'Appeal Treatment' parallels the aforementioned set-up, but introduces a 'second instance' in which player D has the task to review C's punishment decision (in every direction). D can modify/confirm C's punishment decision as such (punishment/no punishment) and D may also change the harshness of punishment (high/low punishment). Note that D is free to change C's decision in every direction; s/he can punish when C has not punished and vice versa. Costs for D are such that they reflect the differences in efforts between confirming and overturning a decision: whereas confirmation is costless, overturning (in either direction) is costly. Moreover, still with the task of reflecting actual incentives in such a setting, concurrent decisions by D have no effect on C's payoff, while an overturning leads to a (modest) 'cost of reversal' to be borne by C (reflecting the 'reputational harm' judges may suffer in case of a reversal).

On a pure 'rational choice' basis, the introduction of a second instance should not matter because if reversal is costly without any instrumental value, nobody would overrule the first instance. In a behavioral perspective, in turn, one would predict a certain amount of punishment in the first instance and also some costly reversals in the second. And that is, indeed, what could be observed in the respective laboratory experiments. People care about the right amount of punishment and they are willing to incur costs to override a decision that, for them, does not seem appropriate. This correction is altruistic because it does not provide the second instance with an instrumental benefit but is a costly decision that imposes, for any modification of the punishment, a disutility on the first punisher.

In substance, the main results are the following. 'Altruistic' corrections occur; and they are not infrequent. The introduction of a second instance increases the level of 'first-instance punishment': whereas, in this experiment, it does not affect the incidence of punishment by the Cs, it strongly influences its harshness.[48] A further result concerns the 'harshness in the instance': 78% of the appeals' decisions are confirmatory. Still, out of the remaining 22% reversals, 80% concern a switch from soft to harsh punishment, which is all the more interesting, as punishment is already harsh in the first instance in the shadow of the newly introduced second instance. Third, interestingly enough, the potential thieves obviously anticipate this increase in punishment and adapt their behavior accordingly. The 'incidence of theft' drops in the Appeals Treatment from 42% to 19%.

48 Whereas in the baseline Trial Treatment only 48% of those Cs who decide to punish pick the harsh option, this number increases to 88% in the Appeals Treatment.

The results, moreover, provide for certain indirect additional twists. One of these twists concerns the 'overall actual punishment'. The above mentioned percentages are based on the strategy method and, hence, on ex ante responses. Because of the (anticipated) deterrent effect of the introduction of the appeals level and the reduced incidence of thefts, actual punishment (punishment carried out) declines considerably, because there were less thefts to punish. In fact, the aggregated number of the total 'punishment points' (= punishment actually actually meted) decreases by a switch to the Appeals Treatment by one half. More than that, a second twist, if one takes the average overall payoff across all players as an indicator of 'welfare', this welfare measure increases in a statistically significant manner.

Interesting results, perhaps—but what is the overall point for enforcement? While the phenomenon of what is called 'altruistic punishment' has been mostly discussed in an institution-less environment, the experiment shows that one can encourage and direct such a punishment, by straightforward institutional means, in different ways and also in a welfare increasing manner. In the above reported experiments, the introduction of a 'second instance' increased harshness of punishment in 'the first instance' even though the decision by the second instance could go in either direction and any change of punishment was costly. Note (once again) that potential wrongdoers were able to anticipate these subtle effects, reduce their behavior accordingly, and allow for an outcome in which—due to the reduced incidence of 'stealing'—less punishment has to be meted out.

The idea that selective incentives on the potential punisher to actually carry out punishment (or to increase its level) can effectively boost punishment levels and stabilize rule compliance is not at all new. These incentives may be of the blunt 'punish the punisher for not punishing' type but they may be—at least in an institutionalized environment—more subtle, e.g., by courts checking on prosecutorial behavior or higher courts checking on the appropriate punishment levels of lower courts. In the experiment at hand, punishment became harsher by the 'first instance', even though the second instance could have also decreased punishment.

One of the pertinent tools in the constitutional economist's/lawyer's toolbox would be the introduction of a second punishment layer, of 'judicial review'. Many more institutional devices are conceivable.

5 Outlook

Big questions at the beginning but only fragmented answers, derived from specific constellations, at the end? Yes and no. Experimental economics is always limited to the study of only one effect at a time. The experiments on 'altruistic punishment' by Fehr & colleagues (and followers), while as such necessarily performed in small steps, have addressed crucial questions; and the results obtained have prompted the authors to suggest a far reaching interpretation[49] regarding the role of 'altruistic punishment' for social stability. The experiments reported here address those questions, again in small (marginal) steps, but from a different (experimental and, regarding the concept of the 'intrinsic disutility of punishment', also conceptual) angle.[50]

This brings us back to the general point. Has experimental economics solved the puzzle of social order by solving possible drawbacks with individually provided punishment? Is the concept of altruistic (third- and second-party) punishment the ultimate keystone in the edifice of social stability? This paper submits that this is not the case. While grossly contributing to our understanding of social order, it is not the final keystone in our understanding of social stability. It is 'another'—though important—'brick in the wall'. A large brick in a small wall? Well, no. The problem of social order is and remains the Great Wall of the social sciences.

Acknowledgment: I am grateful to the editors for inviting me to contribute and for generous advice. Moreover, I wish to thank three referees for their highly valuable suggestions.

49 This 'overarching' interpretation, going beyond the often only very fragmented answers that experimental economics provides, has greatly enriched the discussion.

50 While the phenomena studied in these experiments again allow for an 'overarching interpretation', the results indicate a need for caution regarding the possibility of 'uniform explanations'. Rather, these experiments emphasize the heterogeneity of human behavior and also the dependence of at least some behavioral patterns from underlying psychological categories (see, e.g., the direct influence of 'pro-sociality' on voluntary contributions but not on voluntarily provided punishment). This is not to say that we have to give up attempts for understanding the 'big picture'. However, it suggests (in addition to the obvious interest in the emergence of these different categories) that one should focus more, and also for discussions on the possibilities and limits of social order, on the interplay of different actors and groups of actors within society rather than on 'society as such'.

References

Alexander, J. (2005), The Evolutionary Foundations of Strong Reciprocity, in: *Analyse & Kritik* 27, 106–112

Andreoni, J. (1990), Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving, in: *The Economic Journal* 100, 464–477

Axelrod, R. (1984), *The Evolution of Cooperation*, New York

— (1986), *An Evolutionary Approach to Norms, in: American Political Science Review* 80, 1095–1111

Barcelo, H./V. Capraro (2015), Group Size Effect on Cooperation in One-Shot Dilemmas, in: *Scientific Reports* 5, 1–8

Boehm, C. (1999), The Natural Selection of Altruistic Traits, in: *Human Nature* 10, 205–252

— (2008), Purposive Social Selection and the Evolution of Human Altruism, in: *Cross-Cultural Research* 42, 319–352

— (2014), The Moral Consequences of Social Selection, in: *Behaviour* 151, Special Issue: Evolved Morality – The Biology and Philosophy of Human Conscience, 167–183

Buchanan, J. (1975), *The Limits of Liberty: Between Anarchy and Leviathan*, Chicago

Buckholtz, J./R. Marois (2012), The Roots of Modern Justice: Cognitive and Neural Foundations of Social Norms and their Enforcement, in: *Nature Neuroscience* 15, 655–661

Capraro, V. (2019), *The Dual-Process Approach to Human Sociality: A Review*, Available at SSRN 3409136

— /H. Barcelo (2015), Group Size Effect on Cooperation in One-Shot Social Dilemmas II: Curvilinear Effect, in *PLoS ONE* 10(7): e0131419

— /J. Everett/B. Earp (2019), Priming Intuition Disfavors Instrumental Harm But Not Impartial Beneficence, in: *Journal of Experimental Social Psychology* 83, 142–149

Carlsmith, K./T. Wilson/D. Gilbert (2008), The Paradoxical Consequences of Revenge, in: *Journal of Personality and Social Psychology* 95, 1316–1324

Crockett, M. (2013), Models of Morality, in: *Trends in Cognitive Sciences* 17, 363–366

— /Kurth-Nelson Z./J. Siegel/P. Dayan/R. Dolan (2014), Harm to Others Outweighs Harm to Self in Moral Decision Making, in: *PNAS* 111, 17320–17325

Croson, R./E. Fatas/T. Neugebauer, A. Morales (2015), Excludability and Contribution: A Laboratory Study on Forced Ranking in Team Production, in: *Journal of Economic Behavior and Organization* 114, 13–26

Dannenberg, A./C. Haita-Falah/S. Zitzelsberger (2020), Voting on the threat of exclusion in a public good experiment, in: *Experimental Economics* 23, 84–109

De Quervain, D./U. Fischbacher/V. Treyer/M. Schellhammer/U. Schnyder/A. Buck/E. Fehr (2004), The Neural Basis of Altruistic Punishment, in: *Science* 305, 1254–1258

Egas, M./A. Riedl (2008), *The Economics of Altruistic Punishment and the Maintenance of Cooperation, in Proceedings of the Royal Society - Biological Sciences (Series B)*, 871–878

Elster, J. (1989), *The Cement of Society*, New York

Evans, A./K. Dillon/D. Rand (2015), Fast But Not Intuitive, Slow But Not Reflective: Decision Conflict Drives Reaction Times in Social Dilemmas, in: *Journal of Experimental Psychology: General* 144, 951–966

Evans, A./D. Rand (2018), Cooperation and Decision Time, in: *Current Opinion in Psychology* 27, 67–71

Fehr, E./U. Fischbacher (2004), Third-party Punishment and Social Norms, in: *Evolution and Human Behavior* 25, 63–87

— /— (2003), The Nature of Human Altruism, in: *Nature* 425, 785–791
— /S. Gächter (2002), Altruistic Punishment in Humans, in: *Nature* 13, 1–25
— /S. Gächter (2000), Cooperation and Punishment in Public Good Experiments, in: *American Economic Review* 90, 980–994
— /K. Schmidt (1999), A Theory of Fairness, Competition, and Cooperation, in: *Quarterly Journal of Economics* 114, 817–868
FeldmanHall, O./P. Sokol-Hessner/J. Van Bavel/E. Phelps (2014), Fairness Violations Elicit Greater Punishment on Behalf of Another Than Oneself, in: *Nature Communication* 5, 5306
Fowler, J. (2005), Altruistic Punishment and the Origin of Cooperation, in: *PNAS* 102, 7047–7049
Frank, R. (1983), *Passion within Reason*, New York–London
Gollwitzer, M. (2007), How Affective is Revenge? In: G. Steffgen/M. Gollwitzer (eds.), *Emotions and Aggressive Behavior*, 115–129, Göttingen
Guala, F. (2012), Reciprocity: Weak or Strong? What Punishment Experiments Do (and Do Not) Demonstrate, in: *Behavioral Brain Sciences* 2012, 1–15
Grimm, V./F. Mengel, Let Me Sleep on It: Delay Reduces Rejection Rates in Ultimatum Games, in: *Economic Letters* 111, 113–115
Kimbrough, E./A. Vostroknutov (2016), Norms Make Preferences Social, in: *European Journal of Political Economy* 14, 608–638
— /— (2019), A Portable Method of Eliciting Respect for Social Norms, in: *Economics Letters* 168, 147–150
Kliemt, H. (2020), Economic and Sociological Accounts of Social Norms, in *Analyse & Kritik* 42(1), 41–95
Leist, A. (2005), Social Relations Instead of Altruistic Punishment, in: *Analyse & Kritik* 27(1), 158–171
Lengersdorff, L./C. Wagner/P. Lockwood/C. Lamm (2020), When Implicit Prosociality Trumps Selfishness: the Neural Valuation System Underpins More Optimal Choices When Learning to Avoid Harm to Others than to Oneself, in: *Journal of Neuroscience* 40, 7286–7299
Lewisch, P. (1995), *Punishment, Public Law Enforcement, and the Protective State*, Vienna, New York
— (2004), A Theory of Identification, in: *International Review of Law and Economics* 23, 439–451
– /S. Ottone/F. Ponzano (2011), Free-Riding in Altruistic Punishment: An Experimental Comparison of Third-Party Punishment in a Stand-Alone and in an In-Group-Environment, in: *Review of Law and Economics* 7, 161–190
— / — /— (2015), Third-Party Punishment under Judicial Review: An Economic Experiment on the Effects of a Two-Tier Punishment System, in: *Review of Law and Economics* 11, 209–230
Krajbich, I./B. Bartling/T. Hare/E. Fehr (2015), Rethinking Fast and Slow Based on a Critique of Reaction-Time Reverse Inferences, in: *Nature Communications* 6, 7455
Mackie, J. (1982), Morality and the Retributive Emotions, in: *Criminal Justice Ethics* 1, 3–10
Mendes, N./N. Steinbeis/N. Bueno-Guerra/J. Call/T. Singer (2018), Preschool Children and Chimpanzees Incur Costs to Watch Punishment of Antisocial Others, in: *Nature Human Behavior* 2, 45–51
Milgram, S. (1963), Behavioral Study of Obedience, in: *Journal of Abnormal and Social Psychology* 67, 371–378
Mischkowski, D. (2020), *Decision Time in Social Dilemmas – Personality and Situational Factors Moderating Spontaneous Behavior in First and Second Order Public Good Games, Doctoral Thesis published at Göttingen State and University Library*, University of Göttingen

— /A. Glöckner (2016), Spontaneous Cooperation for Prosocials, But Not for Proselfs: Social Value Orientation Moderates Spontaneous Cooperation Behaviour, in: *Scientific Reports* 6, 1–5
— /— /P. Lewisch (2018), From Spontaneous Cooperation to Spontaneous Punishment: Distinguishing the Underlying Motives Driving Spontaneous Behaviour in First and Second Order Public Good Games, in: *Organizational Behavior and Human Decision Processes* 149, 59–72
Nelissen, R./M. Zeelenberg (2009), Moral Emotions as Determinants of Third-Party Punishment: Anger, Guilt, and the Functions of Altruistic Sanctions, in: *Judgement and Decision Making* 4, 543–553
Nowak, M./K. Sigmund (2005), Evolution of Indirect Reciprocity, in: *Nature* 437, 1291–1298
Rand, D. (2016), Cooperation, Fast and Slow, in: *Psychological Science* 29, 1192–1206
— /J. Greene/M. Nowak (2012), Spontaneous Giving and Calculated Greed, in: *Nature* 489, 427–430.
— /V. Brescoll/J. Everett/V. Capraro/H. Barcelo (2016), Social Heuristics and Social Roles: Intuition Favors Altruism for Women But Not for Men, in: *Journal of Experimental Psychology: General* 145, 389–396
Rabellino, D./R. Morese/A. Ciaramidoro/B. Bara/F. Bosco (2016), Third-Party Punishment: Altruistic and Anti-Social Behaviors in In-Group and Out-Group Settings, in: *Journal of Cognitive Psychology* 28, 486–495.
Sanfey, A./J. Rilling/J. Aronson/L. Nystrom/J. Cohen, The Neural Basis of Economic Decision-Making in the Ultimatum Game (2003), in: *Science* 300, 1755–1758
Schelling, T. (1978), *Micromotives and Macrobehavior*, New York, London
Sterelny, K. (1992), Evolutionary Explanations of Human Behaviour, in: *Australasian Journal of Philosophy* 70, 156–173
— (1996), The Return of the Group, in: *Philosophy of Science* 63, 562–584
— (2016), Cooperation, Culture, and Conflict, in *British Journal for the Philosophy of Science* 67, 31–58
Stich, S. (2007), Evolution, Altruism and Cognitive Architecture: A Critique of Sober and Wilson´s Argument for Psychological Altruism, in: Biology and Philosophy 22, 267–281
Vanberg, V./J. Buchanan (1988), Rational Choice and Moral Order, in: *Analyse & Kritik* 10, 138–160
— /R. Congleton (1992), Rationality, Morality, and Exit, in: *American Political Science Review* 86, 418–431
Vostroknutov, A. (2020), Social Norms in Experimental Economics: Towards a Unified Theory of Normative Decision Making, in: *Analyse & Kritik* 42(1), 3–39
Warneken, F. (2013), Altruistic Behaviors from a Developmental and Comparative Perspective, in: K. Sterelny/R. Joyce/B. Calcott/B. Fraser (eds.), *Cooperation and its Evolution*, Cambridge, London
Yamagashi, T./Y. Matsumoto/T. Kiyonari/H. Takagishi, Y. Li/R. Kanai/M. Sakagami (2017), Response Time in Economic Games Reflects Different Types of Decision Conflict for Prosocial and Proself Individuals, in: *Proceedings of the National Academy of Science* 114, 6394–6399