Why do we punish?

On retribution, deterrence, and the moderating role of punishment system

Ethics statement: The study procedure was approved by the university's ethics committee (application number 71/20). Participants will give informed consent before starting the study protocol. We will comply with all relevant regulations provided by the American Psychological Association (APA).

Abstract

We investigate whether individuals' punishment behavior aims at compensating for inflicted harm (i.e., retribution) or at deterring the offender from committing the offense again (i.e., deterrence), and whether punishment motives depend on the punishment system. Implementing a strategy method, participants (N = 150) can assign punishment for each possible decision of an allocator in a group resource allocation task under three conditions: *Open punishment* (the allocator knows about the punishment, allowing for retribution and deterrence); *hidden punishment* (the allocator does not know about the punishment, precluding deterrence); and *unintentional offense* (decision is made by the computer not the allocator, precluding retribution and deterrence). Contrasting punishment in the hidden punishment and unintentional offense condition reveals retribution, whereas contrasting punishment in the open and hidden punishment condition reveals deterrence. We further examine whether punishment motives depend on whether individuals punish in a decentralized or centralized punishment system.

Keywords: retribution; deterrence; punishment motives

Introduction

The phenomenon that individuals invest resources to punish others for their unethical or uncooperative behavior is well established (e.g., Fehr & Gächter, 2002). It seems that society agrees that criminals, or more generally, people who violate social norms, should be punished. However, more difficult questions remain: For *what purpose*, *by whom* and under *what circumstances* should offenders be punished? What does society or an individual hope to achieve by punishing others? *Why* do people punish?

The present investigation examines two motives that individuals often claim influence their punishment decisions (Anderson & MacCoun, 1999): *Retribution*, which means punishment is assigned to compensate for the harm inflicted, and *deterrence*, which means punishment is assigned to deter the offender or others from committing the offense again. When examining actual punishing behavior, findings point to retribution as a driving factor for punishment, but less so to deterrence (Carlsmith et al., 2002; Carlsmith, 2008). Here, we aim to disentangle retributive and deterrence motives by implementing experimental conditions that disable the attainment of one motive and enable the attainment of the other. Specifically, we will implement a group resource allocation task in which participants act in the role of the recipient and can assign punishment (i.e., second-party punishment), with and without the possibility to influence behavior of the offender (cf. Crockett et al., 2014; Nadelhoffer et al., 2013; Molnar et al., 2020). Comparison of punishment under the different punishment conditions allows inference of both retributive and deterrence motives. Ultimately, this contribution aims to replicate and extend previous findings on the "intuitive retributivism hypothesis." In addition, we examine whether individuals punish in a centralized or decentralized punishment system as a potential boundary condition of punishment motives, as we propose that individuals may not feel inclined to punish in a deterrent way when acting as coequal group members but may do so in the role of the central punishment institution.

Retribution and deterrence

Retribution means to punish "because the offender deserves it." Retributive punishment aims at compensating for inflicted harm without necessarily aiming to achieve any future beneficial consequences (e.g., preventing future transgressions). Punishment should reflect the severity of inflicted harm as well as any altering circumstances, such as whether the norm violation was done intentionally (Carlsmith et al., 2002; Carlsmith, 2008).

Contrary to retribution, punishment driven by a *deterrence* motive considers the consequences of punishment for future interactions, aiming to deter the offender or others from committing the offense again (Carlsmith, 2008). Thus, deterrence-driven punishment aims to regulate the behavior of others and communicate what behavior is (not) acceptable. The future-oriented nature of deterrence implies that individuals expect that there will be future interactions that can be altered by imposing punishment. Apart from the expectancy of repeated interaction, three main factors need to be considered for deterrence: the frequency of the offense, the detection rate, and the publicity of offense and punishment (Carlsmith et al., 2002; Carlsmith, 2008). First, if an offense occurs often, that implies that the current punishment policy is not deterrent. Second, if the detection rate of the offense is low, the imminent punishment should be high to prevent transgressions. Third, punishment can only be deterrent if the punished individual and/or potential copycats know about it. Therefore, punishment should only occur if the offender and/or others learn about the punishment (Carlsmith et al., 2002; Carlsmith, 2008).

By varying factors that should be of relevance for one motive but not for the other (e.g., severity of the offense for retribution or publicity for deterrence), Carlsmith and colleagues (2002) found that individuals tend to be sensitive to factors associated with retribution but not to factors associated with deterrence when sentencing an offender (Carlsmith et al., 2002; Carlsmith, 2008). Carlsmith (2006) further showed that individuals rated information related to retribution as most relevant for the punishment decision; they also

requested that information sooner and more frequently (Carlsmith, 2006). Additionally, individuals were also more confident in their punishment decision when they processed retribution related information compared to deterrence related information.

While the studies described above used hypothetical scenarios, Crockett and colleagues (2014) used an economic game (i.e., a three-player trust game) to investigate punishment motives with real monetary consequences. They found that individuals tended to punish an unfair trustee even if the punished individual would never learn about the punishment. As this condition excludes an important deterrence factor (i.e., being informed about punishment), the observed punishment is interpreted to be motivated by retribution. However, they also observed that punishment occurred to a larger extent (i.e., higher amount and more often) when the punished individual was informed about the punishment compared to when the individual was not informed about the punishment. That shows that individuals also endorse the communicative function of punishment, which could indicate deterrence motives. The idea that individuals appreciate when punishment fosters deterrence is further supported by the finding that victims seeking revenge reported higher satisfaction when the offender understood revenge as a punishment compared to seeing the offender suffer from fate (Gollwitzer et al., 2011). Satisfaction was even higher when the offender not only understood the intention behind the punishment but also showed a positive moral change (Funk et al., 2014). In sum, there is evidence that individuals appreciate when punishment fosters behavior change (possibly indicating deterrence motives; Carlsmith, 2008; Gollwitzer et al., 2011), but they largely behave in a way that is more consistent with retribution (i.e., when the offender is uninformed about the punishment; Carlsmith et al., 2002; Carlsmith, 2008; Crockett et al., 2014; Gollwitzer & Bushman, 2012).

Centralized versus decentralized punishment

While individuals tend to punish in a manner consistent with retribution (Carlsmith et al., 2002; Carlsmith, 2008), they also report support for deterrence when they judge general

punishment rules and name deterrence as one motive for punishment (Carlsmith, 2006, 2008; Gollwitzer et al., 2011). It has often been argued that the mismatch between self-report and behavior stems from a lack of introspection that prohibits individuals from accurately reporting their own motives (e.g., Nisbett & Wilson, 1977). With the present investigation we propose, and empirically examine, the possibility that ordinary people may not feel inclined to punish fellow human beings in a deterrent way (cf. Guala, 2010, for a discussion of this idea). That is, even if individuals pursue deterrence motives (and thus report support for deterrence), they might not assign deterrent punishment because they feel it is not their place to regulate others' behavior. Indeed, punishment is often determined and enforced by central authorities (i.e., the law, the police) rather than by individuals (cf. Baldassari & Grossman, 2011).

It is assumed that there are several advantages of centralized punishment compared to decentralized punishment. First, centralizing punishment overcomes coordination problems. Some individuals might prefer others to be punished, but are unwilling to pay the cost and hope that—in a decentralized peer punishment system—others will punish instead (Casari & Luini, 2012; Elster, 1989; Fehr & Gächter, 2002). Second, centralizing punishment is more efficient. If all individuals separately punish one perpetrator, the costs of punishment often outweigh its benefits, leading to an overall lower payoff for the group in a decentralized system (Nosenzo & Sefton, 2014). Third, acting as central punishment institution leads to more respect for, or legitimacy of, the punishment (Baldassarri & Grossman, 2011; Gross et al., 2016). In addition, individuals preferred a centralized punishment system over a decentralized system or over no system at all (Baldassarri & Grossman, 2011; Kosfeld & Riedl, 2004).

Evidence from the social dilemma literature comparing centralized punishment (i.e., punishment can only be executed by one person or an institution) to decentralized peer punishment (i.e., each individual can punish) shows that in a public goods game, individuals punish more often and to a greater extent as a centralized punishment institution compared to

decentralized peer punishment (O'Gorman et al., 2009). We will examine whether the increase in punishment is explained by retributive or by deterrence concerns. On the one hand, it is possible that individuals in the centralized punishment system may feel inclined to retaliate on behalf of others. However, we propose that centralizing punishment could also reveal deterrence motives. Specifically, we expect that individuals who report that deterrence is a central punishment goal for them may not punish in a deterrent way in a decentralized punishment system (in which they act as one coequal group member), but they do punish in a deterrent way in a centralized punishment system (in which they act as the central punishment authority). To test this idea, our proposed study includes a manipulation of the punishment system—that is, whether all group members can punish the offender in their respective interaction (decentralized punishment), or whether there is only one group member who can punish the offender in their respective interaction (centralized punishment).

The present investigation

The present investigation aims to advance our understanding of two primary motives underlying punishment (i.e., retribution and deterrence). Our design allows a rigorous test for retribution- and deterrence-driven punishment by experimentally creating conditions that enable one motive but preclude the other (cf., Crockett et al., 2014). Further, we include centralized versus decentralized punishment as a potential moderating factor, thus extending the current literature by examining a potential boundary condition of retribution- and deterrence-driven punishment.

To infer whether individuals pursue retributive punishment motives, we implement a condition in which the punished individual is not informed if s/he is being punished (*hidden punishment condition*). As mentioned above, one important factor for the effectiveness of deterrent punishment is that an offender is informed about the punishment, therefore this condition precludes deterrence motives. However, this condition allows for retribution, as punishment can still compensate for the offense that occurred. To distinguish retributive

punishment from mere payoff-based motives like inequality aversion, we additionally include a condition in which the unfair behavior is caused by the computer instead of by the other participant (*unintentional offense condition*, cf. Crockett et al., 2014). Individuals, who are motivated to simply avoid unequal payoffs would punish in this condition, although the other person did not cause the harm and therefore did not "deserve" to be punished. The behavior of individuals with pure inequality aversion should not differ between the punishment conditions, however, individuals with a pure retribution motive should not punish in the unintentional offense condition but rather in the hidden punishment condition. Increased punishment in the hidden punishment condition compared to the unintentional offense condition therefore indicates retributive motives. We expect to find evidence for retributive motives by comparing punishment in the hidden punishment condition with punishment in the unintentional offense condition:

H1: Participants will assign a greater amount of punishment in the hidden punishment condition compared to the unintentional offense condition.

We will also implement a condition in which punishment is open—that is, where the punished individual is informed if s/he is being punished (*open punishment condition*). This condition allows to test for deterrence motives, as it includes a communicative function of punishment. The behavior of individuals with a pure retributive motive should not differ between the open and hidden punishment conditions, however, individuals with a pure deterrence motive should not punish in the hidden condition but should do so in the open condition. We therefore expect to find evidence for deterrence motives by comparing punishment in the open punishment condition with punishment in the hidden punishment condition:

H2: Participants will assign a greater amount of punishment in the open punishment condition compared to the hidden punishment condition.

Participants will have the opportunity to assign punishment under all three punishment conditions (within-subjects factor). Implementing the punishment condition as a within-subjects factor allows us to estimate whether individuals pursue only one motive or both. Specifically, for individuals who punish only to compensate, punishment should differ between the hidden punishment and unintentional offense condition but not between the open and hidden punishment condition; for individuals who punish only to deter the offender from committing the offense again, punishment should differ between open and hidden punishment condition but not between hidden punishment and unintentional offense condition. Not punishing differently under all three punishment conditions implies no evidence for retribution nor deterrence, and differentiating between hidden and unintentional punishment conditions, as well as open and hidden punishment conditions, indicates mixed motives.

To examine the moderating role of the punishment system, we will vary whether all members of the group can punish the allocator in their respective interaction or whether only one group member can punish the allocator (between-subject factor). If individuals aim to retaliate in behalf of others, there should be more punishment in the hidden punishment condition than in the unintentional offense condition under centralized punishment compared to decentralized punishment.

H3: We expect a two-way interaction effect between (a) hidden punishment versus unintentional offense and (b) punishment system, such that the difference between the assigned amount of punishment in the hidden punishment compared to the unintentional offense condition is larger in the centralized compared to the decentralized punishment system.

We further examine whether a potential increase in punishment in the centralized punishment system reflects deterrence motives, which would be indicated by the interaction between punishment condition (open vs. hidden) and punishment system (centralized vs. decentralized):

H4: We expect a two-way interaction effect between (a) open punishment versus hidden punishment and (b) punishment system, such that the difference between the assigned amount of punishment in the open compared to the hidden punishment condition is larger in the centralized compared to the decentralized punishment system.

As additional analyses, we will correlate self-reported motives with actual punishing behavior to answer the following questions: Does self-reported retribution correlate with the difference in punishment between the hidden punishment condition and the unintentional offense condition? Does self-reported deterrence correlate with the difference in punishment between the open and the hidden punishment condition?

Materials and Methods

Procedure

Each participant will first be informed about the general study procedure and sign an informed consent. Then participants will be randomly assigned to one of two punishment systems (centralized vs. decentralized) and the order of punishment conditions as well as the role as "allocator" or as "recipient." Participants will complete a series of allocation tasks in groups of four. First, participants will read detailed instructions for the allocation task and have the opportunity to ask questions. Before starting the allocation task, participants will have to pass a short quiz regarding their understanding of the instructions.

Participants will complete three sessions (one under each punishment condition), including two rounds of the allocation task. For each session, participants will be divided into groups of four (one allocator will be paired with three recipients). Participants will be told that the group composition changes between sessions, but the role as allocator or recipient remains fixed throughout the experiment. Then, participants will be informed that the allocator will sequentially interact with each of the other group members as recipient and that each interaction with the allocator consists of two rounds of the allocation task. Recipients will have the opportunity to reduce the income of the allocator after the first round.

At the beginning of each round, the allocator and the respective group member will each be endowed with 70 monetary units. In addition, the allocator can choose between two options: to assign an additional number of monetary units to him/herself and none to the recipient (i.e., Option A: 70/0) or to assign slightly less monetary units to him/herself but equally as many to the recipient (i.e., Option B: 60/60). Depending on the punishment condition, participants will be informed that the allocator made the decision or that the decision was made by the computer (unintentional offense condition). In addition, participants will be told whether the allocator will be informed about the punishment (open punishment condition) or if s/he will not be informed (hidden punishment condition; for more details see supplementary "Instructions for Participants"). Recipients can use their endowment of 70 MU to assign a punishment to the allocator (i.e., costly punishment). For every monetary unit invested in punishment, the allocator's income will be reduced by two monetary units. Decisions in the resource allocation task will be incentivized, as the monetary units earned during the interactions will be transformed into real money and one interaction will be randomly chosen and paid out to participants at the end of the experiment.

Of interest for the present study is the punishment decision when faced with an allocator who chooses the selfish Option A. However, participants in the role of the allocator are not likely to choose Option A, especially if they fear punishment (Engel, 2011; Baldassari & Grossman, 2011). To be able to capture reactions to an offense, we will implement the strategy method (Fischbacher et al., 2012); that is, participants in the role of the recipient provide their response pattern regarding punishment for each possible decision that can be made by the allocator before any interactions are executed. Specifically, participants in the role of the recipient will be asked to indicate with how many monetary units of their endowment they want to punish the allocator (a) in case s/he chooses Option A, and (b) in case s/he chooses Option B. Participants will then be paired with an allocator and depending on his/her decisions, the specified punishment will be executed.

After completing all rounds under each punishment condition, participants will fill out the Sentencing Goals Inventory (Clements et al., 1998), answer demographic questions (i.e., age and gender), and process manipulation checks. Participants will then be fully debriefed about the paradigm and the background of the study. Finally, participants will receive the fixed payment for their participation and receive the bonus payment according to their decisions in the allocation task.

Conditions and design

We implement a 3 (punishment condition: unintentional offense, open punishment, hidden punishment; within-subjects) x 2 (punishment system: decentralized vs. centralized; between-subjects) mixed design. Implementing punishment system as a within-subjects factor allows for direct comparison of the two punishment motives within the person. Implementing punishment system as a between-subjects factor reduces the number of interactions for each participant and in turn reduces unwanted influences like order effects.

Punishment conditions. In the *hidden punishment condition*, the allocator will not be informed between rounds if s/he was punished by the recipient or not. To disable the allocator from inferring whether s/he was punished when receiving the final income, a random number between -140 and +140 will be added to the total income. In addition, the assigned punishment will be subtracted. If the total amount is negative, the income from that interaction will be displayed as 0. Therefore, the allocator cannot know if a potential reduction is due to punishment or to the random subtraction. In the *open punishment condition*, participants will be told that the allocator will see the amount of punishment between rounds. In the *unintentional offense condition*, in which punishment is hidden, the allocation decision will be made by the computer and not by the allocator. Every participant will complete one session under each of the three punishment conditions in a randomized order.

Punishment systems. Participants in a decentralized punishment system will be told that each group member can punish the allocator in their respective interaction. Participants in the centralized punishment system will be told that they are the only group member who can punish the allocator. They will also be told that they will be the first recipient to interact with the allocator.

Additional measures. After all rounds are finished, participants will be asked to indicate their general punishment motives (Sentencing Goals Inventory; Clements et al., 1998) and to provide demographic information. A sample item for retribution reads: "The correctional system should punish offenders in proportion to the seriousness of their crimes." A sample item for deterrence reads: "Criminals should be harshly punished as examples to others."

Manipulation checks. To ensure participants' understanding of the punishment conditions and the experimental procedure, we will present exemplary decision screens and ask (a) who decided over the resource allocation (the allocator or the computer) and (b) whether or not the allocator will learn about the punishment. We will also ask who can assign punishment to the allocator (all recipients in their respective interaction or only one recipient). If participants answer any one of the questions incorrectly, they will be asked to re-read the instructions then answer the test questions again. After all interactions are finished, we will ask participants if they were always aware of the respective punishment condition when making their decisions during the experiment.

Data collection

Participant characteristics. Participation requirements are: (1) high English language skills (native speaker or C2 according to Common European Framework of Reference for Languages) to ensure the instructions are fully understood, (2) legal age (in this case 18 years) and (3) explicitly agreeing with the terms and conditions of the study (informed consent). We will recruit a sample of UK citizens through the crowdsourcing platform Prolific. Through the

platform we can reach participants with diverse demographic characteristics regarding age, education, or employment status. The procedure does not allow for calculation of the percentage of the sample approached who actually participated. Self-selection cannot be completely excluded.

Location and dates of data collection. The experiment will be conducted online. All interactions will be computer-mediated using oTree (Chen et al., 2016). Data collection will presumably take place in August 2020.

Agreements and payments made to participants. Prior to starting the experiment, participants will be informed about the (a) purpose, (b) procedure, (c) duration, (d) expense allowance, (e) potential benefits, and (f) potential risks of the study. Additionally, participants will be informed that their participation is voluntary and that they can end their participation at any time without giving a reason. Participants will be further informed that their data will be treated confidentially. Participants will receive a fixed payment of £1.30 for their participation and an additional bonus payment depending on the decisions during the experiment. The bonus payment can be between 0 and £1.40.

Institutional Review Board agreements, ethical standards met, and safety monitoring. The procedure is in line with the ethical principles stated in the Declaration of Helsinki and APA guidelines. The study procedure was approved by the university's ethics committee (application number 71/20). During data collection, one experimenter will be available via email at all times to address participants' questions and needs.

Sample size calculation. Crockett et al. (2014) found a medium-sized effect of Cohen's $d > 0.60$ for the contrast between hidden punishment and unintentional offense. To be able to detect even smaller effects of $f = 0.15$, we conducted an a priori sample size calculation using G*Power (Faul et al., 2009). We computed the required sample size assuming no correlations among repeated measures. (We actually expect a positive correlation among repeated measures as some individuals are going to punish more while others will

punish less across conditions, but not knowing the size of the correlation, we compute the sample size for the extreme of zero correlation among repeated measures; the required sample size decreases with a positive correlation among repeated measures.) The analysis revealed that a sample of 146 participants is required to detect differences between the punishment conditions (within-factor, H1 and H2) as well as the within-between interaction between punishment conditions and punishment system (H3 and H4) with a power of .80 at an alpha level of .05. As participants will be in groups of three recipients, we will collect data from 150 participants in the role of the recipient and match these with 50 participants in the role of the allocator. We will stop data collection after the session in which the number of participants acting in the role of recipient—after applying the exclusion criteria—reaches 150. To fill the group under centralized punishment, we will additionally recruit 150 participants who cannot punish the allocator in their respective interaction and are therefore not of interest for the current investigation.

Analysis Plan

Data exclusion criteria. We will exclude participants who (a) repeatedly (that is, after reading the instructions a second time) fail to answer correctly all questions regarding understanding of the instructions; (b) report that during the interactions they were not always sure about the respective punishment condition (for manipulation check items see Electronic Supplementary Material 1); or (c) fail to answer correctly one of two instructed response items (e.g., "This is an attention check. Please answer with 'strongly agree'.")

Coding. Punishment system will be coded with -1 = decentral punishment and 1 = central punishment. For the punishment conditions, two dummy variables will be created:

		Dummy1: Retribution	**Dummy2: Deterrence**
Punishment condition	Unintentional offense	1	0
	Hidden	0	0
	Open	0	1

Punishing behavior. To confirm that participants perceived the allocator choosing the selfish distribution of monetary units as an offense, we will compare punishment amount for intentional offenses (allocator chose distribution) and unintentional offenses (computer chose distribution) with a paired-samples t-test. For trials where the allocator chose the distribution (hidden and open condition), we will also compare punishment for the decision to share and the decision not to share with a paired-samples t-test.

Hypothesis testing. Our study design yields nested data with repeated measures of punishment under the three punishment conditions nested in participants. Therefore, we will conduct a hierarchical mixed regression analysis to test our hypotheses. If assumptions of normality are not met, we will use a non-parametric test (Wilcoxon signed rank test). We will calculate means and standard deviations of punishment amount for all experimental conditions. All hypotheses can be tested using one regression model. The central model will regress punishment amount assigned, provided the allocator chooses the selfish option, on the punishment condition (dummy coded), punishment system (effect coded), and their interaction terms as fixed effects (see Table 1). The model will include random intercepts and slopes for the dummy variables for punishment condition depending on person. We restrict the residual level-1 variance to zero, as all random variance is accounted for by the experimental conditions (cf., Lischetzke et al., 2015). We estimate the correlation between random slopes and intercepts.

Table 1. Central regression model and hypotheses tested.

Dependent variable	Predictors	Hypothesis tested
Punishment		
	Retribution	H1
	Deterrence	H2
	Punishment system	
	Retribution × Punishment system	H3
	Deterrence × Punishment system	H4

Note. Retribution and Deterrence are dummy coded (Retribution: hidden = 0, open = 0, unintentional offense = 1, Deterrence: hidden = 0, open = 1, unintentional offense = 0); Punishment is effect coded with decentral = -1 and central = 1.

H1: Simple effect of Retribution (Dummy1).

H2: Simple effect of Deterrence (Dummy2).

H3: Comparison of the simple slope for Retribution (Dummy1) and punishment system = -1 (decentral) with simple slope for Retribution (Dummy1) and punishment system = 1 (central). Significance of difference is indicated by the interaction between Retribution and punishment system.

H4: Comparison of the simple slope for Deterrence (Dummy2) and punishment system = -1 (decentral) with simple slope for Deterrence (Dummy2) and punishment system = 1 (central). Significance of difference is indicated by the interaction between Deterrence and punishment system.

We will also run the regression model including order of punishment condition, age, and sex as covariates to examine if the direction or significance of the effects reported above depend on specific characteristics of these variables.

Self-reported punishment motives. With the exploration of the relation between punishment behavior and self-reported punishment motives, we will obtain a better understanding of how the perception and attitudes towards punishment are consistent with punishment behavior. We will check for internal consistency of the Sentencing Goals Inventory Scale (Clements et al., 1998) measuring self-reported retribution and deterrence using Cronbach's α. If $\alpha > .70$, we will calculate a mean score for retribution and deterrence as the average of the items measuring each motive. We will report means and standard deviations for both motives. To examine the association between self-reported punishment motives and punishment behavior, we will correlate self-reported retribution with the difference in punishment in the hidden punishment and unintentional offense condition and correlate self-reported deterrence with the difference in punishment in the open and hidden punishment condition across and within punishment systems.

References

Anderson, M. C., & MacCoun, R. J. (1999). Goal conflict in juror assessments of compensatory and punitive damages. *Law and Human Behavior*, *23*(3), 313-330. https://doi.org/10.1023/A:1022308515445

Baldassarri, D., & Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences*, *108*(27), 11023-11027. https://doi.org/10.1073/pnas.1105456108

Balliet, D., Mulder, L. B., & Van Lange, P. A. (2011). Reward, punishment, and cooperation: a meta-analysis. *Psychological bulletin*, *137*(4), 594. https://doi.org/10.1037/a0023489

Carlsmith, K. M. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology*, *42*(4), 437-451. https://doi.org/10.1016/j.jesp.2005.06.007

Carlsmith, K. M. (2008). On justifying punishment: The discrepancy between words and actions. *Social Justice Research*, *21*(2), 119-137. https://doi.org/10.1007/s11211-008-0068-x

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*(2), 284-299. https://doi.org/10.1037//0022-3514.83.2.284

Casari, M., & Luini, L. (2012). Peer punishment in teams: expressive or instrumental choice? *Experimental Economics*, *15*(2), 241-259. https://doi.org/10.1007/s10683-011-9292-6

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88-97. https://doi.org/10.1016/j.jbef.2015.12.001

Clements, C., Wasieleski, D. T., Chaplin, W. F., Kruh, I. P. & Brown, K. P. (1998). The sentencing goals inventory: Development and validation. *Poster session presented at the biennial meeting of the American Psychology-Law Society, Redondo Beach, CA.*

Crockett, M. J., Özdemir, Y., & Fehr, E. (2014). The value of vengeance and the demand for deterrence. *Journal of Experimental Psychology: General*, *143*(6), 2279-2286. http://dx.doi.org/10.1037/xge0000018

Elster, J. (1989). Social norms and economic theory. *Journal of economic perspectives*, *3*(4), 99-117. https://doi.org/10.1257/jep.3.4.99

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, *14*(4), 583-610. https://doi.org/10.1007/s10683-011-9283-7

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63-87. https://doi.org/10.1016/S1090-5138(04)00005-4

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137-140. https://doi.org/10.1038/415137a

Fischbacher, U., Gächter, S., & Quercia, S. (2012). The behavioral validity of the strategy method in public good experiments. *Journal of Economic Psychology*, *33*(4), 897-913. https://doi.org/10.1016/j.joep.2012.04.002

Funk, F., McGeer, V., & Gollwitzer, M. (2014). Get the message: Punishment is satisfying if the transgressor responds to its communicative intent. *Personality and Social Psychology Bulletin*, *40*(8), 986-997. https://doi.org/10.1177/0146167214533130

Guala, F. (2010). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *University of Milan Department of Economics, Business and Statistics Working Paper*, (2010-23). http://dx.doi.org/10.2139/ssrn.1640616

Gollwitzer, M., & Bushman, B. J. (2012). Do victims of injustice punish to improve their mood? *Social Psychological and Personality Science*, *3*(5), 572-580. https://doi.org/10.1177/1948550611430552

Gollwitzer, M., Meder, M., & Schmitt, M. (2011). What gives victims satisfaction when they seek revenge? *European Journal of Social Psychology*, *41*(3), 364-374. https://doi.org/10.1002/ejsp.782

Gross, J., Méder, Z. Z., Okamoto-Barth, S., & Riedl, A. (2016). Building the Leviathan–Voluntary centralisation of punishment power sustains cooperation in humans. *Scientific Reports*, *6*(1), 1-9. https://doi.org/10.1038/srep20767

Kosfeld, M., & Riedl, A. (2004). The design of (de)centralized punishment institutions for sustaining cooperation. *Tinbergen Institute Discussion Paper No. TI 2004-025/1.* http://dx.doi.org/10.2139/ssrn.514182

Lischetzke, T., Reis, D., & Arndt, C. (2015). Data-analytic strategies for examining the effectiveness of daily interventions. *Journal of Occupational and Organizational Psychology*, *88*(3), 587-622. https://doi.org/10.1111/joop.12104

Molnar, A., Chaudhry, S., & Loewenstein, G. (2020). 'It's Not About the Money. It's About Sending a Message!': Unpacking the Components of Revenge. *(July 9, 2020). Available at SSRN:* https://ssrn.com/abstract=3524910

Nadelhoffer, T., Heshmati, S., Kaplan, D., & Nichols, S. (2013). Folk retributivism and the communication confound. Economics and Philosophy, 29(2), 235-261.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231. https://doi.org/10.1037/0033-295X.84.3.231

Nosenzo, D., & Sefton, M. (2014). Promoting cooperation: the distribution of reward and punishment power. In B. Rockenbach, P. A. Van Lange, & T. Yamagishi (Eds.), *Reward and punishment in social dilemmas.* Oxford University Press.

O'Gorman, R., Henrich, J., & Van Vugt, M. (2009). Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1655), 323-329. https://doi.org/10.1098/rspb.2008.1082

Electronic Supplementary Material

ESM1. Manipulation check items. (ESM_Manipulationcheck_Items.pdf)
This file contains the items used as manipulation checks.