

What Motivates Direct and Indirect Punishment?
Extending the ‘Intuitive Retributivism’ Hypothesis

Catherine Molho^{a*}, Mathias Twardawski^b, & Lei Fan^c

^a *Institute for Advanced Study in Toulouse*

^b *Ludwig-Maximilians Universität München*

^c *Vrije Universiteit Amsterdam*

* Corresponding author | catherine.molho@iast.fr

Institute for Advanced Study in Toulouse

Université Toulouse 1 Capitole

1, Esplanade de l'Université

31080 Toulouse, France

Ethics statement: The study described herein will be conducted in accordance with ethical guidelines for conducting research with human subjects. Prior to data collection, ethics approval will be requested from the institutional review board of the first author's institution. Informed consent will be obtained from all subjects prior to their participation. The pre-registration, materials, and data for the study will be made publicly available.

Abstract

Punishment represents a key mechanism to promote cooperation and deter norm violations. Individuals engaging in informal punishment often evoke retribution motives – i.e., wanting to repay the harm done – and/or general deterrence motives – i.e., wanting to prevent onlookers from committing similar offenses in the future. Punishment motivated by retribution is tailored to the severity of offenses, with more severe offenses deserving stricter punishments. Punishment motivated by general deterrence is instead tailored to different factors, such as the observability of punishment, with more widely observed penalties being more effective at deterring similar offenses by onlookers. While the relative importance of these motives is debated, experiments that vary both retribution-relevant and deterrence-relevant factors find that the former are more crucial in determining penalties. Here, we aim to replicate and extend prior work by (a) testing the role that the severity of offenses and the observability of punishment play in motivating (b) *distinct* ways of punishing offenders via high-cost, overt means (i.e., *direct* punishment) versus lower-cost, covert means (i.e., *indirect* punishment). We hypothesize that direct punishment is better suited to serve retribution motives, as it can be more readily adjusted in proportion to the severity of offenses. In contrast, we hypothesize that indirect punishment is better suited to serve general deterrence motives, as it can effectively broadcast condemnation and communicate norms of acceptable behavior to an audience. To test these hypotheses, we aim to recruit 345 participants for an online experiment. Participants will read one out of four vignettes describing an offense and, in a 2×2 design, we will manipulate the severity of the offense (*high* versus *low*) and the observability of punishment (*high* versus *low*). We will use self-reports to assess participants' desires to punish offenders directly *and* indirectly, their endorsement of retribution and deterrence motives, their emotional responses, basic personality traits and demographic information.

Introduction

Punishment represents a key mechanism to promote cooperation and deter norm violations (Balliet et al., 2011; Boyd & Richerson, 1992; Gintis et al., 2008). In modern societies, formal institutions, such as the judicial and prison systems, have a monopoly on imposing penalties for offenses that violate the law. At the same time, individuals and communities regularly face norm-violating behaviors that are not subject to the law (e.g., free-riding, lying, cheating; Hofmann et al., 2014, Molho et al., 2020), but nevertheless have detrimental consequences for cooperation and public goods provision. Experimental research suggests that people are willing to punish non-cooperators, both when they have been personally victimized by an offense (Fehr & Gächter, 2002; Henrich et al., 2006) and when they are merely third-party observers (Fehr & Fischbacher, 2003; Marlowe et al., 2008, 2011). However, the extent to which people impose ‘costly punishment’ outside the laboratory, in naturally occurring interactions, remains contested (Baumard, 2010; Guala, 2012; Pedersen et al., 2019). Moreover, research conducted in field settings suggests that individuals are more inclined to punish offenders using lower-cost, covert tactics (e.g., gossip and benefit withdrawal; Balafoutas et al., 2014, 2016; Molho et al., 2020), rather than bear the high costs of direct confrontation.

To date, there remains considerable debate regarding the motives underlying punishment. Most previous work has focused on people’s attitudes toward formal punishment, such as prison sentencing by the judicial system (Carlsmith, 2006; Carlsmith et al., 2002), but much less attention has been devoted to the broad range of informal punishment responses that people can employ in daily life settings. To punish offenders, individuals can use various tactics, including physical and verbal aggression, communication, reputation manipulation, and benefit withdrawal (Boehm, 1993; Cushman et al., 2019; Raihani & Bshary, 2019). The present work aims to improve our

understanding of informal punishment by (a) examining the relative contribution of retribution-versus deterrence-relevant factors in determining (b) desires to punish offenders through various means, including direct confrontation and more indirect reputation manipulation. A better understanding of people's tendencies to engage in informal punishment is key to solving contemporary challenges, from promoting large-scale cooperation, to battling human-driven climate change, and to deterring free-riding in the face of current global epidemics.

Motives Underlying Punishment

Moral philosophical theories and empirical research have distinguished between two broad classes of motives underlying punishment of offenders: retribution and deterrence. According to a retribution perspective, punishment is motivated by the desire to balance or repay the harm caused by an offense (Carlsmith et al., 2002). Punishment motivated by retribution (and concerns about 'just deserts') is thus sensitive to the severity of offenses, such that more severe offenses deserve harsher penalties. In empirical support of this view, norm violations are met with more punishment when they are perceived as more morally wrong (Hofmann et al., 2018), and as deviating more from cooperation levels in one's group (Fehr & Fischbacher, 2003). While retribution-motivated punishment is typically adjusted to fit the severity of the crime, it is considered less sensitive to the observability of punishment. That is, punishment motivated by a desire to repay harm should be less affected by the presence of an audience. Indeed, decision-making experiments suggest that people engage in punishment, even in one-shot interactions with strangers, which allow no opportunities to induce future cooperation and involve no onlookers (Crockett et al., 2014).

In contrast, according to a deterrence perspective, punishment is primarily motivated by the desire to prevent future norm violations, from the same offender (i.e., *special* deterrence) or other individuals more broadly (i.e., *general* deterrence). Punishment motivated by deterrence (i.e.,

concerns about limiting future offenses) may be sensitive to distinct factors from those influencing retributive punishment. Specifically, punishment aiming at *general* deterrence (Twardawski et al., 2020) should depend on the observability of punishment, because more widely observed penalties can be more effective at deterring onlookers from engaging in similar offenses in the future (Carlsmith et al., 2002). Broadcasted condemnation can communicate norms of acceptable behavior and coordinate punishment of future instances of unacceptable behavior (DeScioli & Kurzban, 2013). Consistent with the idea that punishment functions to deter future offenses, research suggests that people preferentially punish those with whom they expect to interact and cooperate with in the future (Krasnow et al., 2012, 2016), and engage in more punishment in the presence of observers (Kurzban et al., 2007). Importantly, while deterrence-motivated punishment is typically upregulated when there are more onlookers, it is considered less sensitive to the severity of offenses. Strictly speaking, deterrence-focused systems and actors aim to make an example out of even small-time offenses. Thus, when the goal is to limit re-offending, imposing high punishments and maximizing their publicity should be most effective (Carlsmith et al., 2002).

In sum, there is empirical support for the role of both retribution and deterrence concerns in motivating informal punishment of offenders. Some experimental studies have taken a step further in assessing the relative importance of these motives, by varying both retribution-relevant and deterrence-relevant factors and measuring their impact on prison sentencing decisions (Carlsmith, 2006; Carlsmith et al., 2002). Their findings suggest that, although people might report being motivated by deterrence concerns, their decisions are primarily influenced by retribution-related factors. A key goal of the present research is to attempt to replicate these findings by experimentally manipulating the severity of offenses—which should be more relevant when punishment is guided by retribution, but not deterrence, motives—and the observability of

punishment—which should be more relevant if punishment is guided by deterrence, but not retribution, motives. In doing so, it will test two *alternative* hypotheses:

H1: Desires to punish offenders will be stronger when the severity of an offense is high (versus low), irrespective of the observability of punishment. [retribution perspective]

H2: Desires to punish offenders will be stronger when observability of punishment is high (versus low), irrespective of the severity of offenses. [general deterrence perspective]

Moreover, the present research will examine two versions of the ‘intuitive retributivism’ perspective, which suggests that retribution-relevant concerns have primacy over deterrence-relevant concerns in influencing people’s desires to punish offenders. First, aiming to replicate findings by Carlsmith and colleagues (2002), we will test a ‘strong’ version of this perspective, suggesting that *only* retribution-relevant factors will shift individuals’ desires to punish offenders, whereas deterrence-relevant factors will not (**H1_a**). Second, we will test a ‘weak’ version of intuitive retributivism as an alternative hypothesis, suggesting that both retribution *and* deterrence-relevant factors will influence desires to punish offenders, but that the former will have stronger effects on punishment tendencies than the latter (**H1_b**).

Direct and Indirect Punishment Tendencies

Prior empirical work investigating the motives that underlie punishment has typically treated various means of punishment as equivalent, either subsuming them under the umbrella of costly punishment (e.g., Fehr & Gächter, 2002; Henrich et al., 2006) or focusing on punishment imposed by the judicial system (i.e., sentences in terms of years in prison; Carlsmith et al., 2002). However, in response to norm violations that occur in daily life, people can use multiple means of punishment, which can be either overt and costly—what we will refer to as *direct* punishment—or covert and less costly—what we will refer to as *indirect* punishment. Considering a broad range

of direct and indirect punishment responses to norm violations can substantially increase the ecological validity of findings (Molho et al., 2020) and elucidate differential links between motives and distinct forms of punishment.

Direct and indirect means of punishment are characterized by different benefits and costs, and they might be differentially suited to serve retributive versus deterrent goals. To illustrate, direct punishment, which involves overtly confronting offenders via physical or verbal means, can be very costly because it exposes punishers to risks of retaliation from offenders (Campbell, 1999; Guala, 2012; Nikiforakis, 2008). At the same time, confrontational punishment may be better suited to serve retribution motives. This is because punishing offenders directly, be it via physical aggression or verbal reprimanding, can be more straightforwardly adjusted and scaled in proportion to the severity of offenses.

In contrast, indirect punishment, which includes covert strategies of reputation manipulation (e.g., gossip and social exclusion; Feinberg et al., 2014; Wu et al., 2016), is typically less costly than direct confrontation, because it doesn't expose the punisher's identity to the offender (Archer & Coyne, 2005). At the same time, indirect punishment, such as via gossip, may be less suitable to serve retribution motives. As mentioned earlier, one of the key elements of retribution involves administering punishment that fits the crime—i.e., punishment that is neither too harsh nor too lenient. While a punisher can conceivably adjust the negativity of shared information according to the seriousness of an offense, it is much more difficult to control the spread of such information. Gossip can easily get out of hand and its effects are beyond the gossiper's control. Instead, indirect means of punishing offenders—and especially gossip—may be better suited to serve general deterrence goals. For example, by gossiping about offenders, individuals can communicate accepted norms of behavior (Beersma & Van Kleef, 2011; Foster,

2004) and broadcast their condemnation of offenses, in ways that deter *any* other individual from committing the same wrongdoings in the future (DeScioli & Kurzban, 2009, 2013). In line with these ideas, we will test the following hypotheses:

H3: Direct, but not indirect, punishment tendencies will be stronger when the severity of an offense is high (versus low).

H4: Indirect, but not direct, punishment tendencies will be stronger when the observability of punishment is high (versus low).

General Deterrence Versus Reputation Accounts

Importantly, there are two accounts of why the observability of punishment may influence people's desires to punish offenders. As we have posited above, a general deterrence account suggests that tendencies to punish offenders will be stronger when punishment can be observed, because observability increases the potential to broadcast norms of acceptable behavior in a way that limits re-offending. A reputation account also suggests that tendencies to punish offenders will be stronger when punishment can be observed, but for different reasons. According to this account, people may upregulate their punishment in the presence of an audience to reap reputational benefits, in terms of being perceived as a cooperative or trustworthy partner (Barclay, 2006; Jordan & Rand, 2019; Raihany & Bshary, 2015).

Our design allows us to disentangle whether the observability of punishment influences desires to punish via a general deterrence versus a reputation mechanism. Specifically, if observability influences punishment mainly because people want to build or maintain a good reputation, we will see a similar effect of observability on desires to punish directly *and* indirectly (i.e., we will not find support for H4). In contrast, if observability influences punishment mainly because people take up opportunities to broadcast condemnation and communicate moral norms,

we will see observability specifically upregulating desires to punish others indirectly, e.g., via gossip and ostracism (i.e., we will find support for H4).

Another way to test these two alternative explanations is by assessing individual differences in general deterrence versus reputation concerns. According to a general deterrence account, we would expect the effect of observability on desires to punish offenders to be stronger among individuals with higher self-reported deterrence motives. In contrast, according to a reputation account, we would expect the effect of observability on desires to punish offenders to be stronger among individuals with higher self-reported reputational concerns. Our study will explore these alternative possibilities in auxiliary analyses (see Analysis Plan).

Study Overview

In sum, our study aims to test and extend an intuitive retributivism account of the motives underlying punishment tendencies (Carlsmith et al., 2002). To do so, it employs a vignette design which is similar to that used in Carlsmith and colleagues' seminal studies, but uses different vignettes that describe daily life offenses (adapted from prior work; Fan et al., 2020; Molho et al., 2017) to improve ecological validity. Importantly, the study focuses on self-reported desires to punish offenders, rather than actual punishment decisions, and there are multiple reasons why the two may diverge (Baumert, Halmburger, & Schmitt, 2013). In response to hypothetical offenses, people may experience strong urges to punish, that they would not necessarily implement in real life—e.g., due to power and physical strength differentials (Molho et al., 2020; Sell et al., 2009) or emotion regulation processes (Gross, 1998; Gross & John, 2003). Nevertheless, studying the factors driving tendencies to punish offenders can offer important insights into punishers' underlying motives. Here, we extend previous accounts of the motives underlying punishment, by

examining how retribution-relevant versus deterrence-relevant factors influence desires to punish *directly*—using overt, high-cost means—versus *indirectly*—using covert, less costly punishment.

Methods

Sample and Data Collection

Prior to data collection, ethics approval will be requested from the Institute for Advanced Study in Toulouse / Toulouse School of Economics institutional review board. Before participation, we will provide participants with information about the procedure and goals of the study and request informed consent.

We aim to collect data online via ZPID's PsychLab. In order to determine our targeted sample size, we have conducted an a priori power analysis for the 2×2 between-subjects design described below (see 'Design and Measures'). This power analysis suggests that we need an $N = 327$ participants, in order to have 80% statistical power to detect moderate effects (i.e., $f = 0.20$) of the severity of offenses and the observability of punishment on individuals' ratings of appropriate punishment severity, pertaining the conceptual replication (H1/H2) as the central focus of the present research (for more details see 'Analysis Plan'), with an $\alpha = 0.05$. Because we expect to exclude ~5% of participants based on inattentiveness (see 'Exclusion Criteria'), we will aim to recruit a sample of $N = 345$ participants. No analyses will be conducted before data collection is complete.

Exclusion criteria. Survey completion time is one of the best identifiers of inattentive responding (Leiner, 2019) and will be used as an exclusion criterion in our study. Specifically, we will calculate the median completion time of our survey and then exclude participants who spend half of the time or less in completing it.

Design and Measures

After providing informed consent, participants will read one out of four vignettes describing offenses occurring in a daily life setting (i.e., a party). In a 2×2 between-subjects design, we will manipulate the severity of the offense (retribution-relevant factor: high versus low) and the observability of punishment (deterrence-relevant factor: high versus low). Vignettes have been adapted from previous studies (Molho et al., 2017; Fan et al., 2020), to represent the *same* offenses as either causing severe or slight damage and to represent a potential punishment response as being highly observable or not.

After participants read the vignette, they will answer manipulation check questions. To assess participants' perceptions of the severity of the offense, we will ask them two questions assessing how morally wrong and how harmful they think the offender's behavior was, on 7-point Likert scales (1 = *not at all*; 7 = *extremely*). To assess participants' recollection of the observability of punishment, we will ask them two questions assessing how likely they think it is for other guests to know their reaction to the offense (1 = *not at all*; 7 = *extremely*). The four manipulation check questions will be presented in randomized order.

Then, we will measure participants' desires to punish the offender in the vignettes. Specifically, we will ask participants to indicate the extent to which they think the offender should be punished (1 = *not at all*; 7 = *very much*; Hofmann et al., 2018). Further, we will measure participants' tendencies to engage in direct, confrontational punishment via physical or verbal means (e.g., '*I would insult the offender to his face.*') versus indirect, covert punishment via gossip and social exclusion (e.g., '*I would mention something bad I've heard about the offender to other guests who know him.*'). We will use five items for direct punishment and five items for indirect punishment, which will be scored on 7-point Likert scales (1 = *not at all*; 7 = *very much*; adapted

from Molho et al., 2017; Fan et al., 2020). Punishment items will be presented in randomized order.

For exploratory purposes, we will also assess participants' emotional responses to the offense using two methods. First, we will ask participants to indicate their endorsement of arrays of faces expressing negative emotions: sadness, fear, anger, and disgust (faces retrieved from the Radboud Faces Database, Langner et al., 2010; see Molho et al., 2017). The use of facial arrays to measure emotions has been shown to better differentiate experienced anger and disgust toward moral offenses. Second, we will use a more traditional approach to measuring emotions. That is, following Lopez and colleagues (2019), we will also use lexical terms to measure emotions, asking participants to indicate the extent to which they felt *sad*, *fearful*, *angry*, and *grossed out/disgusted*, when reading the vignettes. Facial arrays of emotions and lexical term items will be presented in randomized order.

To perform auxiliary analyses on the associations between individuals' self-reported motives and their punishment tendencies, we will assess endorsement of retribution and deterrence motives, using items adapted from previous research (McKee & Feather, 2008). Additionally, we will measure general reputational concern (Jordan & Rand, 2019) to disentangle general deterrence from reputation mechanisms. Finally, we will measure individual differences (SVO, Murphy et al., 2011; trait aggression, Webster et al., 2014; justice sensitivity, Baumert et al., 2014; and disgust sensitivity, Tybur et al., 2009) and basic demographic information (e.g., gender, age, level of education, socioeconomic status).

Analysis Plan

Reliability and Factor Analyses

Before creating aggregates of the manipulation check items for the severity of offenses and observability of punishment, as well as direct and indirect aggression items, we will assess reliability and perform factor analyses. For the manipulation check items, we will perform correlation analyses and we will consider large correlations ($r > 0.4$) as sufficient evidence that items tap the same construct. For the direct and indirect punishment scales, we will calculate Cronbach's alpha as an indicator of internal consistency and use the traditional cut-off of $\alpha > 0.7$ as evidence of satisfactory reliability. If we find evidence of low reliabilities in the case of scales measuring punishment, we will proceed to conduct factor analyses (using principal components analysis with oblimin rotation). We will check for items with low loadings on their respective factor, and items that exhibit low item-total correlations. If removing items with low item-total correlations improves reliability, we will consider using aggregates of fewer items. If so, we will report results both for the original scales and for aggregates based on factor analytic findings.

Manipulation Checks

We will conduct 2×2 ANOVAs testing the effects of the severity (*high* versus *low*) and the observability (*high* versus *low*) manipulations on the manipulation checks. We will use an aggregate of the moral wrongness and harmfulness items as an index of perceived severity, and an aggregate of the other two items as an index of perceived observability. We expect the severity manipulation, but not the observability manipulation, to have a main effect on perceived severity. Likewise, we expect the observability manipulation, but not the severity manipulation, to have a main effect on perceived observability.

Main Analyses

To test **H1** and **H2**, we will conduct a 2×2 ANCOVA testing the effects of the severity manipulation (*high* versus *low*), the observability manipulation (*high* versus *low*), and their

interaction on individuals' overall rating of appropriate punishment (i.e., how much the offender should be punished). In our analyses, we will include participant gender as a covariate, to account for well-documented sex differences in aggressive tendencies (Archer, 2004). In line with previous work, we expect men to have higher overall ratings of appropriate punishment compared to women.

According to a retribution perspective, we would expect to observe no severity \times observability interaction, but a main effect of the severity manipulation on punishment, such that more punishment is deemed appropriate when the offense severity is high as compared to low (**H1**). Moreover, based on a 'strong' version of the intuitive retributivism perspective, we would expect to observe no significant main effect of the observability manipulation on punishment (**H1a**). Based on a 'weak' version of the intuitive retributivism perspective, we might observe a main effect of the observability manipulation on punishment, but we would expect the severity manipulation to have a stronger effect than the observability manipulation. To compare the effects of the two manipulations on punishment, we will use standardized Cohen's d as a measure of effect size. Further, we will transform Cohen's d s to r s (i.e., to correlation coefficients), and statistically compare the effect sizes of severity and observability manipulations using the procedure described by Eid, Gollwitzer and Schmitt (2011; see also Lenhard & Lenhard, 2014).

To test **H3** and **H4**, we will conduct a mixed 2 (between-subjects severity: *high* versus *low*) \times 2 (between-subjects observability: *high* versus *low*) \times 2 (within-subjects punishment type: *direct* versus *indirect*) ANCOVA. The focus of these analyses will be on the severity \times punishment type and the observability \times punishment type interactions. However, we will test for main effects of the severity and observability manipulations and include the three-way

interaction between severity \times observability \times punishment type in our model (for the sake of completeness). We will use the aggregates of direct punishment items and indirect punishment items as two levels of the within-subjects punishment type factor. Again, we will include participant gender as a covariate, and test for previously documented sex differences in aggressive tendencies. Specifically, we will test for a main effect of participant gender, as well as the gender \times punishment type interaction. In accordance with prior work, we expect men to report stronger desires to engage in direct punishment compared to women, and we will also test whether, reversely, women report stronger desires to engage in indirect punishment compared to men (though evidence for this latter difference is weaker; see Archer, 2004; Molho et al., 2017).

According to **H3**, we expect to observe a severity \times punishment type interaction, such that the severity of the offense will have a positive effect on direct punishment tendencies (with direct punishment being higher when severity is high rather than low), but no effect on indirect punishment tendencies. Reversely, according to **H4**, we expect to see an observability \times punishment type interaction, such that the observability of punishment will have a positive effect on indirect punishment tendencies (with indirect punishment being higher when observability is high rather than low), but no effect on direct punishment tendencies.

Auxiliary Analyses

We will conduct secondary auxiliary analyses to examine the relations of self-reported retribution and deterrence motives with overall punishment, as well as the direct and indirect punishment aggregates. We will further conduct exploratory analyses examining how the severity and observability manipulations influence participants' negative emotions, and how self-reported retribution and deterrence motives relate to experienced emotions.

General deterrence versus reputation accounts. To disentangle two potential mechanisms by which observability of punishment may influence individuals' desires to punish, we will conduct auxiliary analyses including measures of deterrence concerns and reputational concerns as additional predictors. Specifically, we will re-run the ANCOVA models described earlier, but this time including deterrence concerns, reputation concerns, and the deterrence concern \times observability and reputation concern \times observability interactions. If observability influences punishment via a deterrence mechanism, we expect to see a statistically significant deterrence concern \times observability interaction. If we find evidence of such an interaction, we will use simple slopes to test for effects of observability among participants with high versus low deterrence concerns. If observability has a stronger effect on punishment among individuals with high (rather than low) deterrence concerns, we will interpret this as supporting a general deterrence account. In contrast, if observability influences punishment via a reputation mechanism, we expect to see a statistically significant reputation concern \times observability interaction. If we find evidence of such an interaction, we will again use simple slopes to test for the effects of observability among participants with high versus low reputation concerns. If observability has a stronger effect on punishment among individuals with high (rather than low) reputation concerns, we will interpret this as supporting a reputation account.

Materials

Vignette Instructions

On the next page, you will read a scenario. We would like you to focus on how you would feel and react in response to the situation described in the scenario. Please read the scenario carefully and try to experience what is described in it as vividly as possible. After you read the scenario, we will ask you about how it made you feel and how you would react to it.

Moral Violation Vignettes

Vignette #1. Picture attending a party that is being hosted by a casual friend of yours. Some of your close friends are at the party, but most of the people there are just acquaintances. After you've been at the party for a while, you realize that you need to make a phone call. You go to the room where you and the other guests have left your coats to make the call. When you enter the room, you see that another guest – a man that you recognize, but whom you're not friends with – is smoking a cigarette and that he has been casually flicking ashes onto the top jacket on a pile of jackets.

[High severity] He looks at you and gives you a tight smile before he purposefully stubs out his cigarette on the jacket. You look closer and see that the jacket on the top of the pile has been badly damaged by the cigarette.

[Low severity] He looks at you and gives you a tight smile before flicking another bit of ash on the jacket. You look closer and see that the jacket on the top of the pile has been slightly stained by the ashes.

[High observability] Later that night, you bump into this same guest in the living room area. You look around and see that many other guests are still present.

[Low observability] Later that night, you bump into this same guest in the living room area.

You look around and see that a few other guests are still present.

Vignette #2. Picture attending a party that is being hosted by a casual friend of yours. Some of your close friends are at the party, but most of the people there are just acquaintances. After you've been at the party for a while, you decide to step outside to get some air. When you walk outside, you see that another guest – a man that you recognize, but whom you're not friends with – is making a phone call. While talking on the phone, he is casually pulling the flowers and the leaves off the plants.

[High severity] He looks at you and gives you a tight smile before he purposefully tears off the blossoms of several flowers. You look closer and find that there are many scattered leaves and that the plants are likely destroyed.

[Low severity] He looks at you and gives you a tight smile before he continues to play with the plants. You look closer and find that there are a few scattered leaves and that the plants are only slightly damaged.

[High observability] Later that night, you bump into this same guest in the living room area.

You look around and see that many other guests are still present.

[Low observability] Later that night, you bump into this same guest in the living room area.

You look around and see that a few other guests are still present.

Vignette #3. Picture attending a party that is being hosted by a casual friend of yours. Some of your close friends are at the party, but most of the people there are just acquaintances. After you've been at the party for a while, you decide to step outside to get some air. When walking outside, you see another guest – a man that you recognize, but whom you're not friends with.

[High severity] He is putting various expensive silverware that you previously saw on the dining table into his backpack. He looks at you and gives you a tight smile before he walks away.

[Low severity] He is putting a cheap ashtray that you previously saw on the balcony into his backpack. He looks at you and gives you a tight smile before he walks away.

[High observability] Later that night, you bump into this same guest in the living room area. You look around and see that many other guests are still present.

[Low observability] Later that night, you bump into this same guest in the living room area. You look around and see that a few other guests are still present.

Vignette #4. Picture attending a party that is being hosted by a casual friend of yours. Some of your close friends are at the party, but most of the people there are just acquaintances. After you've been at the party for a while, you begin to feel a little hungry. You go to the kitchen to get some food. When you enter the kitchen, you see another guest – a man that you recognize, but whom you're not friends with.

[High severity] He is hastily putting a bottle of expensive whiskey into his backpack. He looks at you and gives you a tight smile before walking away.

[Low severity] He is hastily putting a couple of beer cans into his backpack. He looks at you and gives you a tight smile before walking away.

[High observability] Later that night, you bump into this same guest in the living room area. You look around and see that many other guests are still present.

[Low observability] Later that night, you bump into this same guest in the living room area. You look around and see that a few other guests are still present.

Manipulation Checks

Next, we would like you to answer a few questions about the scenario you just read. '**The offender**' refers to the person who did something wrong in the scenario you read.

- How morally wrong do you think the offender's behavior was?

(1 = *not at all morally wrong*; 2 = *not morally wrong*; 3 = *neutral*; 4 = *morally wrong*; 5 = *extremely morally wrong*)

- How harmful do you think the offender's behavior was?

(1 = *not at all harmful*; 2 = *not harmful*; 3 = *neutral*; 4 = *harmful*; 5 = *extremely harmful*)

- How likely do you think it is that other guests will know your reaction to the offender's behavior?

- How likely do you think it is that only you will know your reaction to the offender's behavior?

(1 = *not at all likely*; 2 = *unlikely*; 3 = *neutral*; 4 = *likely*; 5 = *extremely likely*)

Punishment Responses

Punishment severity:

- To what extent do you think the offender should be punished?

(1 = *not at all*; 7 = *very much*)

On the next page, we will ask you to read some statements and rate how well each of them describes how you would act towards **the offender** (the person who did something wrong).

Direct punishment items:

- I would hit the offender.

- I would insult the offender to his face.

- I would shove the offender.

- I would get in the face of the offender.

- I would yell at or argue with the offender.

Indirect punishment items:

- I would spread negative information about what the offender did to other guests.
- I would mention something bad about what the offender did to others who know him.
- I would try to get other guests to dislike the offender based on what he did.
- I would express my disapproval about what the offender did to other guests.
- I would tell other guests what the offender did.

(1 = *not at all*; 7 = *very much*)

Emotional Responses

Now we would like to know how you felt while reading the scenarios. We would like you to look at some pictures of faces and rate how well each set matches how you felt.



fear

These faces match how I felt when I read the scenario.

(1 = *strongly disagree*; 7 = *strongly agree*)

*sadness*

These faces match how I felt when I read the scenario.

(1 = *strongly disagree*; 7 = *strongly agree*)

*anger*

These faces match how I felt when I read the scenario.

(1 = *strongly disagree*; 7 = *strongly agree*)



disgust

These faces match how I felt when I read the scenario.

(1 = *strongly disagree*; 7 = *strongly agree*)

Lexical items:

Next, we would like you to rate the extent to which you felt different emotions while reading the scenarios. (the items below will be presented in randomized order)

While I read the scenario:

- I felt *sad*.
- I felt *fearful*.
- I felt *angry*.
- I felt *grossed out/disgusted*.

(1 = *strongly disagree*; 7 = *strongly agree*)

Sentencing Goals Inventory (adapted from McKee & Feather, 2008)

Listed below are a number of statements that describe attitudes that different people have about justice in the community. There are no right or wrong answers, only opinions. Read each item and decide whether you agree or disagree and to what extent.

Retribution items:

- Severe sentences are appropriate for offenders who commit serious offenses.
- Offenders should be punished in proportion to the seriousness of their crimes.
- Harm to the victim should be considered when setting the punishment for a given offense.
- Offenders should be made to bear full responsibility for their actions.
- Harsher crimes deserve harsher punishment.

Deterrence items:

- Every punishment should be well publicized.
- Emphasis should be placed on keeping potential other offenders from doing any harm.
- Offenders should be harshly punished as examples to others.
- If there would be tougher punishments against offenses, there wouldn't be so many offenders.
- Light punishments do not provide enough threat to deter people from offenses.

Importance items:

Next, we would like you to rate how important you think each goal is on the 1 (*not at all important*) to 7 (*very important*) scales provided.

- To make sure that the offender “pays” in some way for what they have done.
- To deter other potential offenders.
- To deter the offender from committing similar offenses in the future.

These items will be rated on 7-point Likert scales (1 = *not at all important*; 7 = *very important*)

Brief Fear of Negative Evaluation Scale (adapted by Jordan & Rand, 2019)

These items will be rated on 7-point Likert scales (1 = *not at all characteristic of me*; 7 = *very characteristic of me*)

1. I am afraid that others will not approve of me.
2. I am afraid that people will find fault with me.

3. I often hope that I will say or do the right things.
4. Sometimes I think I am too concerned with other people liking me.
5. I hope that people will view me favorably.
6. I frequently hope that other people will notice my positive attributes.
7. I worry about what other people will think of me even when I know it doesn't make a difference.
8. When I am talking to someone, I worry about what they may be thinking about me.
9. I am frequently afraid of other people noticing my shortcomings.
10. I often worry that I will say or do the wrong things.
11. I hope that other people will like me even when I know it doesn't make a difference.
12. I hope that others will approve of me.
13. When I am talking to someone, I hope that they will be thinking positive things about me.
14. I am usually worried about what kind of impression I make.
15. Sometimes I think I am too concerned with what other people think of me.
16. I am usually excited about the idea of making a good impression.

References

- Archer, J. (2004). Sex differences in aggression in real-world settings: A meta-analytic review. *Review of General Psychology*, 8(4), 291-322. <https://doi.org/10.1037/1089-2680.8.4.291>
- Archer, J., & Coyne, S. M. (2005). An Integrated Review of Indirect, Relational, and Social Aggression. *Personality and Social Psychology Review*, 9(3), 212–230. https://doi.org/10.1207/s15327957pspr0903_2
- Ashton, M., & Lee, K. (2009). The HEXACO-60: A Short Measure of the Major Dimensions of Personality. *Journal of Personality Assessment*, 91(4), 340–345. <https://doi.org/10.1080/00223890902935878>
- Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*, 111(45), 15924-15927. <https://doi.org/10.1073/pnas.1413170111>
- Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2016). Altruistic punishment does not increase with the severity of norm violations in the field. *Nature Communications*, 7(1), 1-6. <https://doi.org/10.1038/ncomms13327>
- Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, 137(4), 594–615. <https://doi.org/10.1037/a0023489>
- Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, 27(5), 325–344. <https://doi.org/10.1016/j.evolhumbehav.2006.01.003>
- Baumard, N. (2010). Has punishment played a role in the evolution of cooperation? A critical review. *Mind & Society*, 9(2), 171-192. <https://doi.org/10.1007/s11299-010-0079-9>

- Baumert, A., Beierlein, C., Schmitt, M., Kemper, C. J., Kovaleva, A., Liebig, S., & Rammstedt, B. (2014). Measuring four perspectives of justice sensitivity with two items each. *Journal of Personality Assessment*, 96(3), 380–390.
<https://doi.org/10.1080/00223891.2013.836526>
- Baumert, A., Halmburger, A., & Schmitt, M. (2013). Interventions against norm violations: Dispositional determinants of self-reported and real moral courage. *Personality and Social Psychology Bulletin*, 39(8), 1053–1068.
<https://doi.org/10.1177/0146167213490032>
- Beersma, B., & Van Kleef, G. A. (2011). How the Grapevine Keeps You in Line: Gossip Increases Contributions to the Group. *Social Psychological and Personality Science*, 2(6), 642–649. <https://doi.org/10.1177/1948550611405073>
- Boehm, C. (1993). Egalitarian Behavior and Reverse Dominance Hierarchy. *Current Anthropology*, 34(3), 227–254. <https://doi.org/10.1086/204166>
- Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, 13(3), 171–195.
[https://doi.org/10.1016/0162-3095\(92\)90032-Y](https://doi.org/10.1016/0162-3095(92)90032-Y)
- Campbell, A. (1999). Staying alive: Evolution, culture, and women's intrasexual aggression. *Behavioral and Brain Sciences*, 22(2), 203–214.
<https://doi.org/10.1017/S0140525X99001818>
- 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., 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>
- 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. <https://doi.org/10.1037/xge0000018>
- Cushman, F., Sarin, A., & Ho, M. (2019). *Punishment as communication*. <https://psyarxiv.com/wf3tz>
- DeScioli, P., & Kurzban, R. (2009). Mysteries of morality. *Cognition*, 112(2), 281–299. <https://doi.org/10.1016/j.cognition.2009.05.008>
- DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin*, 139(2), 477–496. <https://doi.org/10.1037/a0029065>
- Eid, M., Gollwitzer, M., & Schmitt, M. (2011). *Statistik und Forschungsmethoden Lehrbuch*. Weinheim: Beltz.
- Fan, L., Molho, C., Kupfer, T., & Tybur, J.M. (2020). *Moral Emotions and Aggressive Tactics in Third-Party Punishment: The Effect of Welfare Tradeoff Ratio*. Manuscript in preparation
- Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960), 785–791. <https://doi.org/10.1038/nature02043>
- Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868), 137–140. <https://doi.org/10.1038/415137a>
- Feinberg, M., Willer, R., & Schultz, M. (2014). Gossip and Ostracism Promote Cooperation in Groups. *Psychological Science*, 25(3), 656–664. <https://doi.org/10.1177/0956797613510184>

- Foster, E. K. (2004). Research on Gossip: Taxonomy, Methods, and Future Directions. *Review of General Psychology*, 8(2), 78–99. <https://doi.org/10.1037/1089-2680.8.2.78>
- Gintis, H., Henrich, J., Bowles, S., Boyd, R., & Fehr, E. (2008). Strong Reciprocity and the Roots of Human Morality. *Social Justice Research*, 21(2), 241–253. <https://doi.org/10.1007/s11211-008-0067-y>
- Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3), 271–299. <http://dx.doi.org/10.1037/1089-2680.2.3.271>
- Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and wellbeing. *Journal of Personality and Social Psychology*, 85(2), 348–362. <http://dx.doi.org/10.1037/0022-3514.85.2.348>
- Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, 35(1), 1–15. <https://doi.org/10.1017/S0140525X11000069>
- Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2006). Costly Punishment Across Human Societies. *Science*, 312(5781), 1767–1770. <https://doi.org/10.1126/science.1127333>
- Hofmann, W., Brandt, M. J., Wisneski, D. C., Rockenbach, B., & Skitka, L. J. (2018). Moral Punishment in Everyday Life. *Personality and Social Psychology Bulletin*, 44(12), 1697–1711.
- Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, 345(6202), 1340–1343. <https://doi.org/10.1126/science.1251560>

- Jordan, J. J., & Rand, D. G. (2019). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology*, 118(1), 57–88. <https://doi.org/10.1037/pspi0000186>
- Krasnow, M. M., Cosmides, L., Pedersen, E. J., & Tooby, J. (2012). What Are Punishment and Reputation for? *PLoS ONE*, 7(9). <https://doi.org/10.1371/journal.pone.0045662>
- Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking Under the Hood of Third-Party Punishment Reveals Design for Personal Benefit. *Psychological Science*, 27(3), 405–418. <https://doi.org/10.1177/0956797615624469>
- Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, 28(2), 75–84. <https://doi.org/10.1016/j.evolhumbehav.2006.06.001>
- Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & Van Knippenberg, A. D. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and Emotion*, 24(8), 1377-1388.
- Leiner, D. J. (2019). Too Fast, too Straight, too Weird: Non-Reactive Indicators for Meaningless Data in Internet Surveys. In *Survey Research Methods* (Vol. 13, No. 3, pp. 229-248).
- Lenhard, W. & Lenhard, A. (2014). *Hypothesis Tests for Comparing Correlations*. Available at: <https://www.psychometrica.de/correlation.html>. Bibergau (Germany): Psychometrica. doi:10.13140/RG.2.1.2954.1367
- Lopez, L. D., Moorman, K., Schneider, S., Baker, M. N., & Holbrook, C. (2019). Morality is relative: Anger, disgust, and aggression as contingent responses to sibling versus acquaintance harm. *Emotion*, No Pagination Specified-No Pagination Specified. <https://doi.org/10.1037/emo0000707>

- Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Ensminger, J., Gurven, M., Gwako, E., Henrich, J., Henrich, N., Lesorogol, C., McElreath, R., & Tracer, D. (2008). More 'altruistic' punishment in larger societies. *Proceedings of the Royal Society B: Biological Sciences*, 275(1634), 587–592. <https://doi.org/10.1098/rspb.2007.1517>
- Marlowe, F. W., Berbesque, J. C., Barrett, C., Bolyanatz, A., Gurven, M., & Tracer, D. (2011). The 'spiteful' origins of human cooperation. *Proceedings of the Royal Society B: Biological Sciences*, 278(1715), 2159–2164. <https://doi.org/10.1098/rspb.2010.2342>
- McKee, I. R., & Feather, N. T. (2008). Revenge, Retribution, and Values: Social Attitudes and Punitive Sentencing. *Social Justice Research*, 21(2), 138. <https://doi.org/10.1007/s11211-008-0066-z>
- Molho, C., Tybur, J. M., Güler, E., Balliet, D., & Hofmann, W. (2017). Disgust and Anger Relate to Different Aggressive Responses to Moral Violations. *Psychological Science*, 28(5), 609–619. <https://doi.org/10.1177/0956797617692000>
- Molho, C., Tybur, J.M., Van Lange, P.A.M., & Balliet, D. (2020). Direct and indirect punishment of norm violations in daily life. *Nature Communications*, 11, 3432.
- Murphy, R. O., Ackermann, K. A., & Handgraaf, M. (2011). *Measuring Social Value Orientation* (SSRN Scholarly Paper ID 1804189). Social Science Research Network. <https://doi.org/10.2139/ssrn.1804189>
- Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics*, 92(1), 91–112. <https://doi.org/10.1016/j.jpubeco.2007.04.008>

- Pedersen, E. J., McAuliffe, W. H., Shah, Y., Tanaka, H., Ohtsubo, Y., & McCullough, M. E. (2019). When and why do third parties punish outside of the lab? A cross-cultural recall study. *Social Psychological and Personality Science*, 11(6), 846-853. <https://doi.org/10.1177/1948550619884565>
- Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. *Trends in Ecology & Evolution*, 30(2), 98–103. <https://doi.org/10.1016/j.tree.2014.12.003>
- Raihani, N. J., & Bshary, R. (2019). Punishment: One tool, many uses. *Evolutionary Human Sciences*, 1. <https://doi.org/10.1017/ehs.2019.12>
- Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences*, 106(35), 15073–15078. <https://doi.org/10.1073/pnas.0904312106>
- Twardawski, M., Tang, K. T. Y., & Hilbig, B. E. (2020). Is It All About Retribution? The Flexibility of Punishment Goals. *Social Justice Research*. <https://doi.org/10.1007/s11211-020-00352-x>
- Tybur, J. M., Lieberman, D., & Griskevicius, V. (2009). Microbes, mating, and morality: Individual differences in three functional domains of disgust. *Journal of Personality and Social Psychology*, 97(1), 103–122. <https://doi.org/10.1037/a0015474>
- Webster, G. D., DeWall, C. N., Pond, R. S., Deckman, T., Jonason, P. K., Le, B. M., Nichols, A. L., Schember, T. O., Crysel, L. C., Crosier, B. S., Smith, C. V., Paddock, E. L., Nezlek, J. B., Kirkpatrick, L. A., Bryan, A. D., & Bator, R. J. (2014). The brief aggression questionnaire: Psychometric and behavioral evidence for an efficient measure of trait aggression. *Aggressive Behavior*, 40(2), 120–139. <https://doi.org/10.1002/ab.21507>

Wu, J., Balliet, D., & Lange, P. A. M. V. (2016). Gossip Versus Punishment: The Efficiency of Reputation to Promote and Maintain Cooperation. *Scientific Reports*, 6(1), 1–8.

<https://doi.org/10.1038/srep23919>