What Motivates Direct and Indirect Punishment?

Extending the 'Intuitive Retributivism' Hypothesis

Catherine Molho[a*], Mathias Twardawski[b], & Lei Fan[c]

Forthcoming in *Zeitschrift für Psychologie* (2022), Special Issue "The Motivational Basis of Punishment: Conceptual Replications of the 'Intuitive Retributivism' Hypothesis". This copy may not exactly replicate the final, authoritative version of the article.

[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

Abstract

Punishment represents a key mechanism to deter norm violations and is motivated by retribution and/or general deterrence. Retribution-motivated punishment is tailored to offense severity, whereas deterrence-motivated punishment is tailored to different factors, including punishment observability. This study aimed to replicate and extend prior work by testing how offense severity and punishment observability motivate direct, confrontational punishment versus indirect, covert punishment. Participants (*N* = 308) read vignettes describing offenses with varying severity (high versus low) and punishment observability (high versus low). We then assessed their punishment tendencies—overall, direct, and indirect—and their endorsement of retribution and deterrence motives. Findings supported a 'strong version' of intuitive retributivism. Manipulating retribution-relevant information consistently influenced punishment: participants reported stronger overall, direct, and indirect punishment tendencies when severity was high (versus low). Self-reported deterrence (but not retribution) motives positively related to overall, direct, and indirect punishment tendencies. However, manipulating deterrence-relevant information did not influence punishment.

Keywords: punishment, gossip, motives, retribution, deterrence

Introduction

Punishment represents a key mechanism to 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). 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 means (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, reputation manipulation, and benefit withdrawal (Boehm, 1993; Raihani & Bshary, 2019). The present work aims to improve our understanding of informal punishment by examining the relative contribution of retribution- versus deterrence-relevant factors in determining tendencies to punish offenders through various means, including direct confrontation and more indirect reputation manipulation.

Motives Underlying Punishment

Moral philosophical theories and empirical research have distinguished between two broad classes of motives underlying punishment: 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 offense severity, with more severe offenses deserving harsher penalties. In empirical support of this view, norm violations are punished more 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 less sensitive to punishment observability. 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 from third parties (i.e., *general* deterrence). Deterrence-motivated punishment may be sensitive to distinct factors from those influencing retributive punishment. Specifically, punishment aiming at *general* deterrence (Twardawski, Tang et al., 2020) should depend on punishment observability, because widely observed penalties can be more effective at deterring onlookers from engaging in similar offenses (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 offense severity. 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 in motivating informal punishment. 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 information. A key goal of this research is to attempt to replicate these findings by experimentally manipulating offense severity—which should be more relevant when punishment is guided by retribution, but not deterrence, motives—and punishment observability—which should be more relevant when punishment is guided by deterrence, but not retribution, motives. In doing so, it will test two *alternative* hypotheses:

H1: Punishment tendencies will be stronger when offense severity is high (versus low), irrespective of punishment observability. [retribution perspective]

H2: Punishment tendencies will be stronger when punishment observability is high (versus low), irrespective of offense severity. [general deterrence perspective]

Moreover, this research will examine two versions of the ‘intuitive retributivism’ perspective, which suggests that retribution-relevant concerns have primacy over deterrence-relevant concerns in influencing punishment. 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’ punishment tendencies, 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 punishment, but that the former will have stronger effects 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., prison sentencing; 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—i.e., *direct* punishment—or covert and less costly—i.e., *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, via physical aggression or verbal reprimanding, can be more straightforwardly adjusted and scaled in proportion to offense severity.

In contrast, indirect punishment, which includes covert means of reputation manipulation (e.g., gossip and social exclusion; Feinberg et al., 2014; Wu et al., 2016), is less costly than direct confrontation, because it doesn't reveal the punisher's identity to the offender (Archer & Coyne, 2005); Dores Cruz et al., 2020). At the same time, indirect punishment 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 gossip especially—may be better suited to serve general deterrence goals. 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 punishment observability may influence punishment tendencies. As we have posited above, a general deterrence account suggests that punishment tendencies 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 punishment tendencies will be stronger when punishment can be observed, albeit for different reasons. According to this account, people upregulate 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 punishment observability influences punishment tendencies 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 direct *and* indirect punishment tendencies (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 indirect punishment tendencies (i.e., we will find support for H4).

Another way to test these two alternative explanations is by assessing how individual differences in general deterrence versus reputation concerns moderate the effects of observability on punishment tendencies. We elaborate on and test for these potential moderations in auxiliary analyses (see ESM).

Study Overview

In sum, this 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 punishment tendencies, rather than actual punishment decisions, and there are multiple reasons why the two may diverge (Baumert et al., 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; Tybur et al., 2020) or emotion regulation processes (Gross, 1998; Gross & John, 2003). Nevertheless, studying the factors driving punishment tendencies 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

Ethics. Before data collection, we obtained ethics approval from the Institute for Advanced Study in Toulouse (IAST) / Toulouse School of Economics (TSE) institutional review board. All participants provided informed consent.

Pre-registration. The study hypotheses, methods, and analysis plan were pre-registered and are available here: http://dx.doi.org/10.23668/psycharchives.4234

Materials, data, and code availability. Materials for this study are included in the ESM, which is available here: http://dx.doi.org/10.23668/psycharchives.4952. The data and code are available here: http://dx.doi.org/10.23668/psycharchives.4374

Power analysis. To determine our targeted sample size, we conducted an a priori power analysis for the 2 × 2 between-subjects design described below (see 'Design and Measures'). This power analysis suggested that we needed $N = 327$ participants, to obtain 80% statistical power to detect moderate effects (i.e., $f = 0.20$) of offense severity and punishment observability on overall punishment tendencies, with an $\alpha = 0.05$. We focused on these effects when calculating a priori power, because they are most relevant to testing H1/H2 and conceptually replicating the 'intuitive retributivism' account. Because we expected to exclude ~5% of participants based on inattentiveness, we aimed to recruit a sample of $N = 345$ participants.

Inclusion and exclusion criteria. Data was collected online via ZPID's PsychLab and participants were recruited by the panel company 'respondi'. We recruited individuals who were UK citizens, aged between 18-65 years, and fluent in English. We aimed to obtain an equal representation of male and female participants.

Survey completion time is one of the best identifiers of inattentive responding (Leiner, 2019) and was used as an exclusion criterion. Specifically, we calculated the median completion time of our survey (12.5 minutes) and then excluded participants who spent half of the time or less in completing it (≤ 6.25 minutes). This resulted in the exclusion of 42 out of 350 participants (12% of the recruited sample). In what follows, we report results after excluding inattentive participants, as pre-registered. Results using the full sample are reported in the ESM (see 'Robustness of Main

Analyses' and 'Robustness of Auxiliary Analyses') and show that findings are robust to including inattentive respondents.

Sample. Our final sample consisted of 308 participants (61.8% male; M_{age} = 47.5 years, SD_{age} = 12.23). In terms of educational attainment, seven participants had some high school education (2.3%); 77 had completed high school (25.0%); 99 had some college education (32.1%); 96 had obtained a bachelor's degree (31.2%); 24 a master's degree (7.8%); and 5 a doctoral degree (1.6%). In sum, we obtained a sample that was diverse in terms of age and educational background, though skewed toward including more male participants.

Design and Measures

First, participants read one out of four vignettes describing offenses occurring in a daily life setting. In a 2 × 2 between-subjects design, we manipulated offense severity (retribution-relevant factor: high versus low) and punishment observability (deterrence-relevant factor: high versus low). Vignettes were adapted from previous studies (Molho et al., 2017; Fan et al., 2020; Tybur 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 answered manipulation check questions. To assess perceptions of offense severity, we asked participants two questions assessing how morally wrong and how harmful they thought the offender's behavior was (1 = *not at all*; 7 = *extremely*). We calculated the bivariate correlation (two-tailed) between these items, $r = .59$, $p < .001$, which we considered strong enough (based on pre-registered criteria) to form an aggregate, with higher scores indicating perceptions of offenses as more severe. To assess participants' recollection of punishment observability, we asked two questions assessing how likely they thought it was for other guests to know their reaction to the offense (1 = *not at all*; 7 = *extremely*). We calculated the bivariate correlation (two-tailed) between these items, r = .54, p < .001, which we considered strong enough to form an aggregate, with higher scores indicating perceptions of punishment as more observable.

Then, we measured participants' punishment tendencies. To assess overall punishment tendencies, we asked participants the extent to which they thought the offender should be punished (1 = *not at all*; 7 = *very much*; Hofmann et al., 2018). Further, we measured 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 used five items for direct punishment and five items for indirect punishment (1 = *not at all*; 7 = *very much*; adapted from Molho et al., 2017; Fan et al., 2020). Punishment items were presented in randomized order. Following our pre-registration, we used Cronbach's alpha as an indicator of reliability for direct (α = .90) and indirect punishment items (α = .88) and calculated aggregates of each scale, with higher scores indicating stronger endorsement of the respective punishment type.

To perform auxiliary analyses on the associations between individuals' self-reported motives and their punishment tendencies, we assessed endorsement of retribution and deterrence motives, using items adapted from previous research (McKee & Feather, 2008). Specifically, we assessed participants' agreement with five items measuring retribution motives (α = .81) and five items measuring deterrence motives (α = .82; 1 = *strongly disagree*; 7 = *strongly agree*). We calculated aggregates for each scale, with higher scores indicating stronger endorsement of the respective motives. We also asked participants to rate the importance of three goals—retribution, special deterrence, and general deterrence—but do not analyze this data here. To disentangle general deterrence from reputation mechanisms, we used 16 items to measure reputation concern

(α = .96; Jordan & Rand, 2019). Participants used 7-point Likert scales to indicate how various statements characterized them (1 = *not at all characteristic of me*; 7 = *very characteristic of me*). We report results from auxiliary analyses aiming to disentangle general deterrence from reputation accounts in the ESM.

For exploratory purposes, we assessed participants' emotional responses to the offense (see pre-registration), but we do not analyze this data here. Finally, we measured other individual differences (SVO, trait aggression, and justice sensitivity, see pre-registration for details) and basic demographic information (gender, age, and level of education). No analyses were conducted before completing data collection.

Results

Manipulation Checks

We first conducted 2 × 2 ANOVAs testing the effects of the severity (*high* versus *low*) and the observability (*high* versus *low*) manipulations on the perceived severity and perceived observability aggregates (i.e., our manipulation checks). Our manipulation of offense severity worked as intended. Results showed a main effect of the severity manipulation on the perceived severity aggregate, $F(1, 304) = 20.66$, $p < .001$, $\eta^2 = 0.06$, with participants in the high severity condition ($N = 147$) perceiving offenses as more wrong and harmful ($M = 5.56$, $SD = 1.16$), compared to participants in the low severity condition ($N = 161$, $M = 4.89$, $SD = 1.39$). There was no main effect of the observability condition ($p = .490$) and no severity condition × observability condition interaction ($p = .974$) predicting perceived severity.

However, our manipulation of punishment observability did not work as intended. Results showed no main effect of the observability manipulation on the perceived observability aggregate, $F(1, 304) = 1.00$, $p = .318$, $\eta^2 < .01$, with participants in the high observability condition (N = 158) and those in the low observability condition (N = 150) perceiving punishment as similarly observable (*high*: M = 3.29, SD = 1.43; *low:* M = 3.49, SD = 1.61). In contrast, we observed a main effect of the severity manipulation on perceived observability, F(1, 304) = 7.22, p = .008, η^2 = 0.02, with participants in the high severity condition perceiving punishment as somewhat more observable (M = 3.63, SD = 1.54) than those in the low severity condition (M = 3.16, SD = 1.47). There was no severity condition × observability condition interaction (p = .663) predicting perceived observability.

Main Analyses

Overall punishment tendencies. To test **H1** and **H2**, we conducted 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 punishment tendencies (i.e., ratings of how much the offender should be punished). In our analyses, we included participant gender as a covariate, to account for well-documented sex differences in aggressive tendencies (Archer, 2004). In line with previous work, we expected men to report higher overall punishment tendencies compared to women.

According to a retribution perspective, we would expect to observe no severity × observability interaction, but a main effect of the severity manipulation on punishment tendencies, such that participants think the offender should be punished *more* when offense severity is high as compared to low (**H1**). Moreover, based on a 'strong' version of intuitive retributivism, we would expect to observe no significant main effect of the observability manipulation on punishment tendencies (**H1$_a$**). Based on a 'weak' version of intuitive retributivism, we might observe a main effect of the observability manipulation on punishment tendencies, but we would expect the severity manipulation to have a stronger effect than the observability manipulation (**H1$_b$**)[1].

Results showed that there was no severity condition × observability condition interaction (*p* = .857) affecting overall punishment tendencies. We observed a main effect of the severity manipulation, *F*(1, 303) = 26.90, *p* < .001, η^2 = 0.08, such that participants thought the offender should be punished *more* when the offense severity was high (*M* = 5.28, *SD* = 1.46) as compared to low (*M* = 4.35, *SD* = 1.62; see Figure 1, first panel). There was no main effect of the (unsuccessful) observability manipulation on overall punishment tendencies (*p* = .866). We also did not observe a main effect of gender on overall punishment (*p* = .140). In sum, results provide support for a ‘strong version’ of intuitive retributivism (**H1$_b$**).

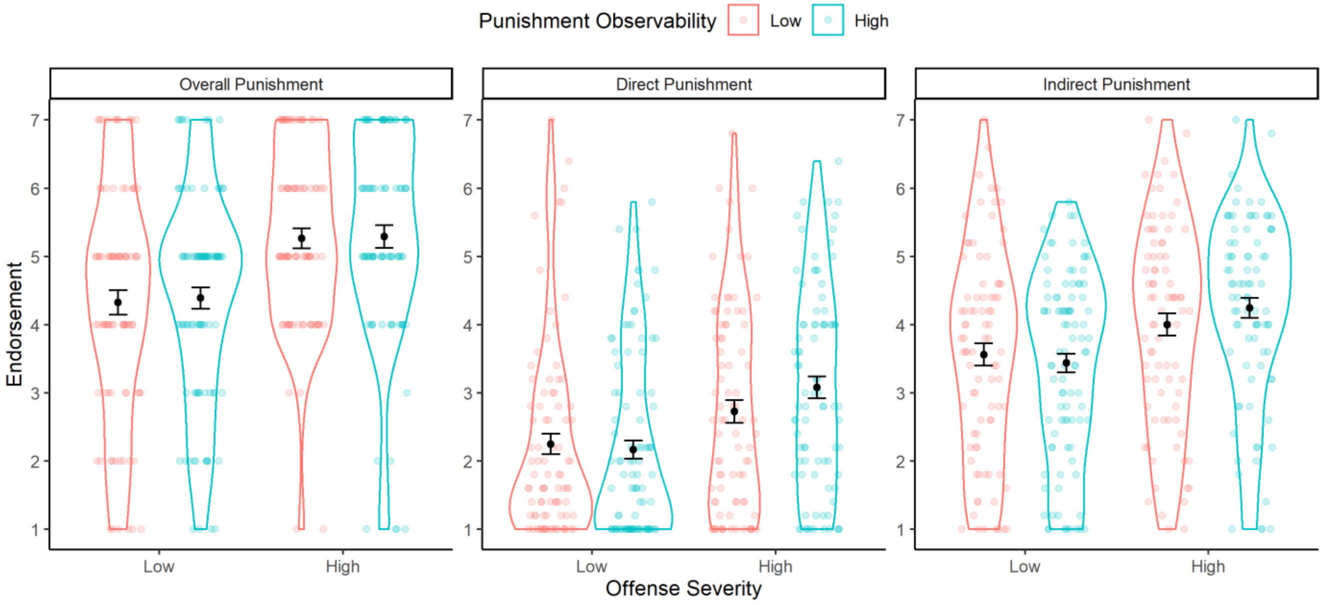

Figure 1. Endorsement of overall, direct, and indirect punishment tendencies depending on the severity condition (high versus low) and the observability condition (high versus low). Error bars indicate one standard error of the mean.

[1] Our pre-registration included explicit criteria for comparing the strength of effects under H1$_b$. However, we do not perform this comparison, as we did not observe a main effect of the observability manipulation.

Direct and indirect punishment. To test **H3** and **H4**, we conducted a mixed 2 (between-subjects severity: *high* versus *low*) × 2 (between-subjects observability: *high* versus *low*) × 2 (within-subjects punishment type: *direct* versus *indirect*) ANCOVA. The focus of these analyses was on the severity × punishment type and the observability × punishment type interactions. However, we also tested for main effects of the severity and observability manipulations and included the three-way interaction between severity × observability × punishment type in our model (for the sake of completeness). We used the direct and indirect punishment aggregates as two levels of the within-subjects punishment type factor. Again, we included participant gender as a covariate, and tested for previously documented sex differences in aggressive tendencies. Specifically, we tested for a main effect of gender, as well as the gender × punishment type interaction. Based on prior work, we expected men to report stronger direct punishment tendencies compared to women, and we also tested whether, reversely, women report stronger indirect punishment tendencies compared to men (though evidence for this latter difference is weaker; see Archer, 2004; Molho et al., 2017).

According to **H3**, we expected to observe a severity × punishment type interaction, such that offense severity would 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 expected to see an observability × punishment type interaction, such that punishment observability would 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.

Contrary to our expectations, we did not observe a severity manipulation × punishment type interaction (**H3**; p = .800), nor an observability manipulation × punishment type interaction

(**H4**; $p = .871$). Instead, consistently with our analyses on overall punishment tendencies, results showed a main effect of the severity manipulation on punishment, $F(1, 303) = 19.16$, $p < .001$, $\eta^2 = 0.06$, such that both direct and indirect punishment were higher when offense severity was high (*direct*: $M = 2.65$, $SD = 1.38$; *indirect*: $M = 4.04$, $SD = 1.45$) as compared to low (*direct*: $M = 2.04$, $SD = 1.22$; *indirect*: $M = 3.41$, $SD = 1.41$) (see Figure 1, second and third panels). There was no main effect of the observability manipulation on punishment ($p = .633$), nor a severity × observability interaction ($p = .087$). Importantly, there was a substantial difference in the overall endorsement of direct versus indirect punishment, $F(1, 303) = 360.26$, $p < .001$, $\eta^2 = 0.54$, with participants reporting stronger tendencies to intervene indirectly ($M = 3.71$, $SD = 1.47$) rather than directly ($M = 2.33$, $SD = 1.33$). Consistent with previous work, we observed that gender had a main effect on punishment tendencies, $F(1, 303) = 6.16$, $p = .014$, $\eta^2 = 0.02$, with women reporting weaker punishment tendencies than men; this effect was not qualified by punishment type ($p = .402$).

Auxiliary Analyses

As pre-registered, we conducted secondary auxiliary analyses to examine the relations of self-reported retribution and deterrence motives with overall, direct, and indirect punishment tendencies.

Self-reported motives and overall punishment tendencies. First, we run a general linear model testing the effects of retribution motives, deterrence motives, and their interaction on overall punishment tendencies. As in previous analyses, we controlled for participant gender. Contrasting findings from our main analyses, self-reported retribution motives had no main effect on overall punishment tendencies, $F(1, 303) = 0.16$, $p = .693$, $\eta^2 < .01$. Instead, when looking at participants' self-reported motives for punishment, we observed a positive main effect of deterrence motives on ratings of overall punishment, $b = 0.84$, $F(1, 303) = 5.79$, $p = .017$, $\eta^2 = 0.02$. There was no interaction between retribution and deterrence motives ($p = .320$), nor a main effect of gender ($p = .106$).

Self-reported motives and direct versus indirect punishment. We then run a general linear model testing the effects of retribution and deterrence motives on endorsements of direct versus indirect punishment tendencies (as two levels of a within-subjects punishment type factor). As with our main analyses, the focus here was on the retribution motives × punishment type and the deterrence motives × punishment type interactions. However, we also tested for main effects of retribution and deterrence motives across punishment types, and included the three-way interaction between retribution motives × deterrence motives × punishment type in our model (for the sake of completeness). Again, we tested for a main effect of gender, as well as the gender × punishment type interaction predicting endorsements of punishment.

We found no evidence in support of the idea that retribution and deterrence motives differentially relate to direct versus indirect punishment. Specifically, we did not observe a retribution motives × punishment type interaction ($p = .175$), nor a deterrence motives × punishment type interaction ($p = .505$). Instead, consistent with our results on overall punishment above, there was a positive, main effect of deterrence motives on punishment, $F(1, 303) = 12.32$, $p = .001$, $\eta^2 = 0.04$, which held both for direct punishment tendencies ($b = 1.04$, $p < .001$, $\eta^2 = 0.04$) *and* indirect punishment tendencies ($b = 0.85$, $p = .008$, $\eta^2 = 0.02$). In contrast, we did not observe a main effect of retribution motives on punishment, $F(1, 303) = 0.35$, $p = .555$, $\eta^2 < 0.01$. Finally, there was a main effect of gender on punishment endorsements, $F(1, 303) = 7.16$, $p = .008$, $\eta^2 = 0.02$), with men reporting stronger punishment tendencies than women; this effect was not qualified by punishment type ($p = .334$).

Discussion

The main goal of our research was to conceptually replicate a seminal study on the 'intuitive retributivism' hypothesis (Study 1; Carlsmith et al., 2002), showing that retribution-relevant concerns have primacy over deterrence-relevant concerns in determining punishment of norm violations. Following the original study, we presented participants with a hypothetical offense and manipulated retribution-relevant and deterrence-relevant factors in a 2 × 2 design. We based both experimental manipulations on the original study, varying (a) offense severity (retribution-relevant factor; high versus low) and (b) punishment observability (deterrence-relevant factor; high versus low). To improve ecological validity, we deviated from the original study by using vignettes of daily life offenses. Then, to replicate and extend previous findings, we measured overall punishment tendencies (i.e., the extent to which participants thought the offender should be punished), as well as tendencies to punish in distinct ways, using direct confrontation versus more indirect reputation manipulation.

Our experiment successfully replicated the original study findings, providing consistent evidence that retribution-relevant factors influence punishment tendencies. In support of **H1**, manipulating offense severity shifted overall punishment tendencies, such that participants reported stronger overall punishment when offense severity was high (compared to low). Notably, the effect size we observed ($\eta^2 = 0.08$) was substantially lower than the one in the original study ($\eta^2 = 0.26$). Moreover, offense severity had a similar effect on both direct, confrontational punishment *and* indirect punishment, via gossip and exclusion. Although this finding is not consistent with our prediction that retribution may specifically motivate direct punishment of offenders (**H3**), it further bolsters confidence that retribution concerns are key drivers of punishment.

Further, consistent with a ‘strong version’ of intuitive retributivism, and the findings of Carlsmith and colleagues’ study, we found no evidence that deterrence-relevant factors influence punishment tendencies. In support of **H1$_a$**, manipulating punishment observability did not shift participants’ overall punishment tendencies; instead, participants showed similar punishment tendencies irrespective of whether punishment observability was high or low. Moreover, we did not observe any effects of punishment observability on either direct or indirect punishment tendencies, further strengthening support for the intuitive retributivism hypothesis. This latter result is inconsistent with our proposition that deterrence specifically motivates indirect punishment (**H4**). In sum, results do not support a general deterrence account of why observability may influence punishment (because we observe no differential effect of observability on direct versus indirect punishment), nor do they support a reputational account (because we observe no main effect of observability on punishment tendencies; see ESM for auxiliary results).

Interestingly, while we found no evidence that manipulating deterrence-relevant information shifts punishment, results showed a different pattern at the self-report level. Specifically, we consistently observed that self-reported deterrence motives were positively related to overall, direct, and indirect punishment tendencies. In contrast, self-reported retribution motives were unrelated to punishment tendencies. Together, these results fit with prior research that found a substantial disconnect between what people *believe* is driving their punishment based on introspection (i.e., deterrence concerns) and what seems to *actually* drive their punishment (i.e., retribution-relevant information; cf. Carlsmith et al., 2002; Twardawski, Hilbig et al., 2020).

Limitations and Future Directions

Across our analyses, we did not find evidence that manipulating deterrence-relevant information, in terms of punishment observability, influences punishment tendencies. However, these findings should be interpreted with caution, because our manipulation of punishment observability did not work as intended. While our manipulation check analyses showed that participants correctly rated offenses in the high (compared to low) severity condition as more morally wrong, ratings of punishment observability were similar in both the high and low observability conditions. If participants indeed did not notice variation across observability conditions, this casts doubt on our conclusions regarding the relevance of deterrence factors in influencing punishment.

It is certainly possible that our manipulation of punishment observability was too subtle to be noticed by participants, or that it was overridden because of other situation-specific expectations (e.g., the expectation that behavior is generally observable in the highly social situation of a party). This issue could be remedied in future research by using starker differentiations between conditions. One alternative possibility, though, is that the manipulation check items we used were not well-designed to capture differences between observability conditions. In particular, one of our items asked participants: ‘How likely do you think it is that only you will know your reaction to the offender’s behavior?’ In retrospect, we realize that this item is not suitable to distinguish situations in which *a few* (low observability) versus *many* (high observability) other guests are present. That said, when we repeated our manipulation check analyses excluding this unsuitable item, we found qualitatively similar results: the severity manipulation, but *not* the observability manipulation, influenced perceived punishment observability.

Further, although our study showed that respondents differentiate between direct versus indirect punishment tactics—with a clear preference for the latter—we did not find support for the idea that retribution and deterrence concerns motivate distinct tactics. One explanation for this finding is that, as suggested earlier, people may not be aware of the *actual* factors driving their punishment and thus may have difficulty picking tactics that match them. Another possibility has to do with the fact that our study only considered offenses that target third parties (i.e., other-relevant offenses) and, as such, induce primarily indirect punishment rather than direct confrontation (Molho et al., 2017, 2020). To more clearly test how retribution and deterrence concerns motivate punishment, future work can also include offenses that target participants themselves (i.e., self-relevant offenses), which typically evoke both direct *and* indirect punishment (possibly conditional on the punishers' goals).

Conclusions

To conclude, our replication study provided support for the intuitive retributivism hypothesis. Based on people's introspection on the motives that drive punishment decisions, researchers may be tempted to conclude that deterrence concerns are key in determining penalties. Instead, consistent with findings from seminal work on intuitive retributivism, we find that retribution-relevant (but not deterrence-relevant) factors influence overall punishment tendencies, as well as distinct direct and indirect punishment tactics.

Acknowledgments

This study was preregistered on PsychArchives (http://dx.doi.org/10.23668/psycharchives.4234). We declare that we did not deviate from the preregistration in the data collection and that our analyses followed the preregistered plan as closely as possible. We note though that, although our preregistration included specific criteria for comparing the effects of the severity and observability manipulations, we do not perform these comparisons as we did not observe a main effect of the observability manipulation. There are no other deviations from the preregistered analyses. Electronic Supplementary Materials (ESM) are also available on PsychArchives (http://dx.doi.org/10.23668/psycharchives.4952). Data collection for this study was funded by PsychLab, a service of the Leibniz Institute for Psychology Information (ZPID). C.M. acknowledges IAST funding from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements d'Avenir program). M.T. received support by LMUexcellent, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder.

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

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., 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

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

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

Dores Cruz, T. D., Nieper, A., Testori, M., Martinescu, E., & Beersma, B. (2020). *An integrative definition and framework to study gossip.* Pre-print available at PsyArXiv: https://doi.org/10.31234/osf.io/b8x57

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. https://doi.org/10.1177/0146167218775075

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

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).

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. https://doi.org/10.1038/s41467-020-17286-2

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

Twardawski, M., Hilbig, B. E., & Thielmann, I. (2020). Punishment goals in classroom interventions: An attributional approach. *Journal of Experimental Psychology: Applied*, *26*(1), 61–72. https://doi.org/10.1037/xap0000223

Tybur, J.M., Molho, C., Çakmak, B., das Dores Cruz, T.D., Singh, G.D., & Zwicker, M. (2020). Disgust, anger, and aggression: Further tests of the equivalence of moral emotions. *Collabra: Psychology*, *6*, 34. https://doi.org/10.1525/collabra.349

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