Original Article

Examining Consequentialist Punishment Motives in One-Shot Social Dilemmas

Friederike Funk[1] and Dorothee Mischkowski[2]

[1]Faculty of Arts & Sciences, NYU Shanghai, China
[2]Social Cognition Center Cologne, Faculty of Human Sciences, University of Cologne, Germany

Abstract: We investigated whether consequentialist motives may underlie punishment decisions in single-round (i.e., one-shot) social dilemmas in which there is no prospect of reciprocity. In particular, we used an incentivized public goods game to examine how the prospect of receiving information on the effect of punishment (i.e., information that indicates potential regret and intention for future behavioral change on the part of the transgressor) affects people's punishment decisions. We also took person-situation interactions into account and studied whether prosocial individuals (i.e., persons high in Honesty-Humility and Social Value Orientation) punish more strongly when they receive consequentialist information. The data did neither reveal the hypothesized effects of information availability on punishment decisions nor were these effects conditional on dispositional prosociality. We discuss potential limitations of these findings as well as open questions for future research.

Keywords: punishment behavior, punishment motives, retribution, cooperation, social preferences

The question of whether punishment is conducted out of retributive or consequentialist motives represents an ongoing debate and has also been investigated in the context of social dilemmas and economic games. Social dilemmas are defined as situations in which the individual interest conflicts with the collective interest, for instance, when it comes to providing and maintaining public goods (Dawes, 1980; van Lange et al., 2013). As insufficient cooperation behavior in social dilemmas is often punished (e.g., Fehr & Gächter, 2002), such economic games (e.g., a public goods game, see Ledyard, 1995) can be used to investigate how retribution (i.e., an act of revenge to retaliate free-riding) and consequentialist motives (i.e., re-educating offenders or deterring offenders from further free-riding) contribute to punishment. Traditionally, the presence of punishment in single-round (i.e., one-shot) interactions has been interpreted as evidence for retributive motives that contradict consequentialist accounts of punishment, as there are no future interactions to deter or educate the offender for. In these settings, punishment is thought to be motivated by retribution and a desire for equality (Bone & Raihani, 2015; Mischkowski et al., 2018). Here, we critically examine the Intuitive Retributivism Hypothesis by investigating punishment behavior in a single-round interaction. In particular, we challenge this hypothesis by examining whether the outlook to receive information about the effects that punishment has on the transgressor may influence people's punitive behavior in a one-shot public goods game.

The current study conceptually replicates research on how consequentialist – as opposed to mere retributivist – motives affect punishment-related outcomes. Specifically, we relate our study to Funk and colleagues (2014, Study 1), where participants report being more satisfied if punishment is followed by feedback from the transgressor compared to when punishment lacks any kind of feedback. Their findings on people's hedonic reactions after punishment challenge the Intuitive Retributivism Hypothesis: If people punish for retributive reasons, punishment should be equally satisfying, regardless of its consequences on the transgressor. We test whether people's actual punishment behavior – instead of their hedonic reactions related to punishment – shows a similar pattern. Specifically, we investigate if the outlook to receive information about the effects of punishment on the transgressor affects people's actual punishment decisions, which would challenge the Intuitive Retributivism Hypothesis.

Previous research using one-shot economic games suggests that punishers adjust their punishment behavior depending on the consequences that punishment may have. In research on "hidden" versus "open" punishment, Crockett and colleagues (2014) found that participants are more

likely to punish and punish more harshly if their punished partners find out that their final payoff was the result of a punitive reduction (i.e., if punishment was "open"). Put differently, when the punished partners would only find out about their final payoff without any information about how it came about (i.e., if punishment was "hidden"), participants punished less. Crockett and colleagues interpret this difference as the effect of deterrent (i.e., consequentialist) motives in the "open" punishment condition or concern about communicating norms that accompanies people's retributive motives (which were identified in the "hidden" punishment condition).

Also, using one-shot interactions, Molnar and colleagues (2020) found that participants are willing to opt for a punishment for their partner that is less severe if, along with the punishment, their partner will read why their bonus has been reduced (e.g., "because you were unfair to your partner in the previous task"). Thus, punishers want transgressors to know that they are being punished and why (for similar findings from social psychology, see Gollwitzer & Denzler, 2009; Gollwitzer et al., 2011). People also punish less, for instance, if they can communicate their emotions (Xiao & Houser, 2005), and punishers expect harmless punishment to be as effective as harmful punishment if it is communicative (Sarin et al., 2020; see also Cushman et al., 2022). The present study looked at a related yet different facet of consequentialist punishment and manipulates a factor that has not been studied in the context of punishment decisions so far: we varied experimentally whether punishers knew they would receive information about the effect that their punishment had on the transgressor.

In studies that examine punishers' hedonic reactions, findings suggest that punishers care about knowing how punishment has affected a transgressor. People expect punishment to be more satisfying if they imagine receiving feedback from the transgressor compared to when they imagine receiving no feedback from the transgressor after punishment, and this difference in satisfaction can also be found if participants actually receive feedback or not (Funk et al., 2014, Study 1). Plus, the content of the feedback matters, such that information about a positive transgressor change after punishment makes punishment more satisfying than feedback about no change (see Funk et al., 2014, Study 2). While research on the hedonic effects of punishment suggests a role for consequentialism, it has not been studied so far whether the availability of such consequentialist information affects people's punitive decisions to begin with.

The Present Study

In the present study, we experimentally varied whether participants would be able to find out if their punishment has had an effect on the transgressor. Keeping people's retributive and expressive options constant between conditions, we examined how the availability of this kind of consequentialist information (i.e., the prospect that punishers will get to know which effect their punishment has on the transgressor) affects people's punitive reactions. If punishment in one-shot interactions is purely retributive, the availability of information on the effect of one's punishment should not influence the degree of punishment. However, if consequentialist motives contribute to punishment decisions in one-shot interactions, punishment behavior should increase when the corresponding consequentialist motives are addressed.

To account for individual differences, we assessed dispositional prosociality that might moderate the effect of information availability on people's punishment decisions. Specifically, we hypothesized increased consequentialist punishment motives for dispositional prosocials as these individuals should be more inclined to punish in order to establish prosocial norms that prevent them from (further) being exploited in the future. In this study, consequentialist motives would only be addressed when the corresponding information was provided, therefore we expected increased punishment behavior in the respective condition for dispositional prosocials as compared to when no consequentialist information was provided. As proself individuals per definition only maximize their own welfare, they were hypothesized to invest few monetary resources to punish, independent of whether consequentialist information would be provided or not.

To assess dispositional prosociality we relied on Social Value Orientation (SVO) as a measure of social preferences, next to Honesty-Humility as the related basic trait dimension of the HEXACO personality inventory (Ashton & Lee, 2007; Lee & Ashton, 2004). Specifically, SVO (Murphy et al., 2011; Van Lange, 1999) consists of an individual difference measure "defined in terms of the weights people assign to their own and others' outcomes in situations of interdependence" (Balliet et al., 2009, p. 533). SVO is operationalized as a series of Dictator Games in which individuals allocate monetary resources between themselves and an anonymous person. Even though SVO has been shown to be highly predictive of cooperative behavior in social dilemmas (for a recent meta-analysis, see Pletzer et al., 2018), there is heterogeneous evidence on whether prosocials (i.e., individuals who consider the other person's outcome when allocating resources) punish more harshly than proself (i.e., selfish) individuals. While some studies do not find a difference between prosocials and proselfs (e.g., Böckler et al., 2016; Mischkowski et al., 2018; Yamagishi et al., 2012), others do find such a difference – both, in the expected direction of increased punishment for prosocials (e.g., Bieleke et al., 2016; Haruno et al., 2014) as well as in the reverse direction of decreased punishment for

prosocials (Karagonlar & Kuhlman, 2013). We attempted to explain this heterogeneous evidence by identifying a potential boundary condition. We investigated whether prosocials punish more strongly, especially when consequentialist motives are addressed. Specifically, using the materials and measures of Mischkowski and colleagues (2018), our methods allowed for direct replication of the relation between SVO and punishment investments in the no information condition. We expected to replicate their identified null effect of SVO on punishment investments in the no information (i.e., baseline) condition. However, when consequentialist information was provided (i.e., in our experimental condition), we expected prosocials to punish more strongly than proselfs – thereby explaining heterogeneous evidence on the relation between SVO and punishment.

To validate the person-situation interaction with a related yet broader individual difference measure, we investigated whether a similar interaction pattern holds for individuals high in Honesty-Humility. Honesty-Humility "represents the tendency to be fair and genuine in dealing with others, in the sense of cooperating with others even when one might exploit them without suffering retaliation" (Ashton & Lee, 2007, p. 156). It thus represents a form of active (vs. reactive) prosociality and has been shown to be related to SVO (Hilbig et al., 2014). As individuals high in SVO and Honesty-Humility cooperate in the first place (e.g., Balliet et al., 2009; Hilbig et al., 2012), they inherently face a risk of being exploited. As this rationalizes consequentialist punishment to establish prosocial norms, we hypothesized that prosocials as measured by their SVO and Honesty-Humility punish more severely when consequentialist motives are addressed as compared to when they are not. In turn, we expected generally low punishment investments for individuals low in SVO and Honesty-Humility.[1]

Hypotheses

In the current study, we examined whether participants invest more resources to punish (i.e., punish more harshly) if they are able to find out its effects on the transgressor. Our preregistered hypotheses were as follows:

Hypothesis 1 (H1): If punishment is not purely retributive but also driven by consequentialist motives, participants who know that they will *receive information on the consequences of punishment on the transgressor* invest more resources to punish than participants who know they will not receive this information.

Additionally, we took potential person-situation interactions into account to offer an even further differentiated perspective on punishment behavior and its underlying motives. We hypothesized that the effect of information availability on punishment behavior would be more pronounced for prosocial individuals (see Figure 1).

Hypothesis 2a (H2a): There is an interaction between SVO and whether participants *receive information on the consequences of punishment* on people's punishment behavior, reflecting that prosocials increase their punishment investments when consequentialist motives are addressed. That is, we expect prosocials to show stronger punishment in the condition with available information in comparison to the condition without given information. We expect a lower increase – if any – in the punishment behavior of proselfs when consequentialist information is provided as compared to the control condition without this information.

Hypothesis 2b (H2b): We expected a similar interaction pattern as outlined in H2a for individuals high vs. low in Honesty-Humility.

Materials and Methods

All materials, codes, additional analyses, and anonymized data are available on the open science framework (https://osf.io/mpcyw/).

Data Collection

Data was collected via the Decision Lab Cologne, the database of the Social Psychology research group at the University of Cologne (UoC). The data base mainly consists of students from the UoC above 18 years who registered at the database to receive invitations to participate in psychological studies. The database is built in accordance with recent European data protection regulations ("DSGVO") and approved by the ethics committee of the UoC's Faculty of Human Sciences. Specifically, participants were informed and agreed in the personality base assessment that their personality data is going to be linked to their data of subsequent studies.

For the current study, participants were invited via email to take part in an online experiment on "decision

[1] It is important to note that Agreeableness, rather than Honesty-Humility, has been shown to be negatively related to punishment behavior (Hilbig et al., 2016; Thielmann et al., 2020). However, we focus less on the general relation between prosocial traits and punishment, but on potentially increased consequentialist punishment motives for (active) prosocials (i.e., individuals high in SVO and Honesty-Humility) who face a risk of being exploited in comparison to reactive prosocials (i.e., individuals high in Agreeableness). We report bivariate correlations between all HEXACO dimensions and punishment investments as well as the interaction effect of each HEXACO dimension with the information condition in an online appendix on the Open Science Framework (OSF; https://osf.io/gu4by/).

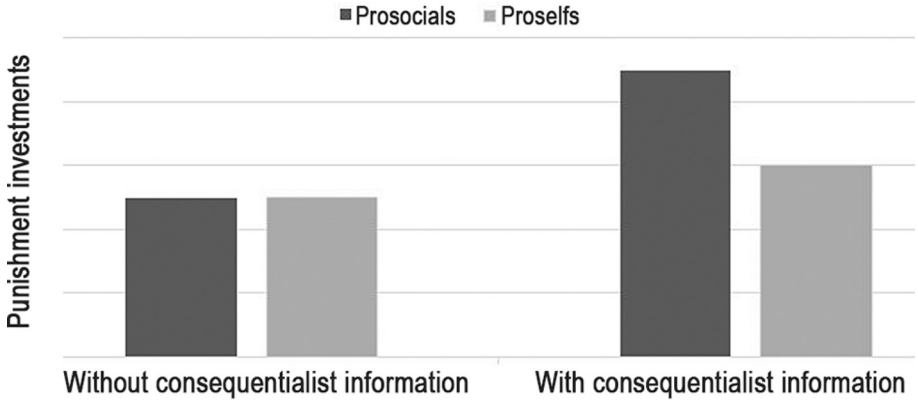

Figure 1. Illustration of the hypothesized interaction pattern (H2): Punishment investments as a function of the availability of consequentialist information and dispositional prosociality.

preferences." The invitation entailed information about the estimated duration, average payment, and broad information about the study's content and procedure. Participants who subscribed to a virtual session of the experiment were therein paired with three other participants and interacted with them in real-time. The study itself was run via oTree (Chen et al., 2016) in accordance with economic standards. That is, interactions were real, without any deception. Decisions (e.g., contributions to the public good, punishment investments) were incentivized, meaning participants were paid according to the decisions they and their group members made.

For each hypothesis, we conducted an a priori power analysis in G*Power (Faul et al., 2009) based on an alpha level of .05 and a power of .90 (see the Stage 1-Protocol as preregistration for details, https://osf.io/9672y/). To assure a high power for the interaction effect (H2), we preregistered to collect data of at least $N = 200$ participants with correct answers to the comprehension questions of the public goods game (i.e., indicating that they correctly understood the conflict of individual and collective interests, see below).

We further preregistered to include an important covariate in our model: the averaged contribution behavior of the other group members relative to one's own contribution, so that we could assess the effect of information availability independently of the degree of inequality in contributions. We complied with all of these aspects.

Sample

In total, 276 participants (76 male, 198 female, 1 diverse, 1 missing) between the ages of 18 and 81 ($M = 29.04$, $SD = 10.76$) successfully completed the experiment. At the beginning of the study, the software randomly assigned each group of four players either to the control ($n = 147$) or experimental condition ($n = 129$). Note that we report complete participation without considering participants who started but dropped out of the study, yielding sample sizes being not divisible by four. Out of this complete sample, 203 participants (56 male, 145 female, 1 diverse, 1 missing)[2] between the ages of 18 and 81 ($M = 28.36$, $SD = 9.14$) answered the comprehension check questions correctly and formed the final sample, out of which 116 were in the control condition (no availability of consequential information) and 87 were in the experimental condition (availability of consequential information).

Condition and Measures

As outlined, our manipulated variable is the *availability of information on the consequences of punishment*. To assess whether the impact of such consequentialist motives in one-shot interactions differs among prosocials and proselfs, we measured SVO (H2a) via the 15 items SVO Slider Measure (Murphy et al., 2011) and Honesty-Humility (H2b) as part of the HEXACO personality inventory (100 items version; Lee & Ashton, 2018). As a dependent variable, we assessed punishment magnitude (i.e., punishment investments, analyzed as percentual investments relative to the earnings of the public goods game; see section on Transformations).

As further measured variables, we assessed cooperation behavior as individual contributions to the public goods game. In addition, we asked participants at the end of the information condition whether they would behave differently in a similar future interaction and whether they regret their contribution behavior.

[2] One participant entered a non-retrievable Decision Lab ID and could therefore not be matched to the base assessment that includes the personality measures and demographics. That is why the number of observations is reduced to $N = 202$ when testing H2a and H2b and when reporting demographic variables.

Procedure

In general, when registering as a participant at the Decision Lab Cologne, participants first take part in a base assessment containing a broad range of personality questionnaires (e.g., basic traits, social preferences, cognitive reflection capacity, etc.). Data is pseudonymized via a Decision Lab ID, and participants have consented that data can be linked to subsequent studies. Thus, the relevant personality data (i.e., SVO and the HEXACO personality inventory containing Honesty-Humility) had already been collected prior to our study, allowing us to solely run the Public Goods Game as outlined below.

Focusing on the procedure of the data collection of the present study, participants first provided informed consent and received the instructions of a public goods game in groups of four players. Each of the players received 4€ that they could decide to keep or contribute to the common good. Participants were described that the money they contribute to the public good would be doubled before the overall sum of all players' contributions would subsequently be divided equally among all four group members, yielding a return of the public good of 0.5. Participants read examples about fictional payoffs depending on certain contribution patterns (for the exact wording of the instructions, see https://osf.io/uszje/). They were reminded that they were interacting with real partners and that their decisions were about real money. Participants were explicitly told that at a second stage of the game, they would be able to reduce the other players' outcomes at a cost, if they liked, with a cost-to-impact ratio of 1:4 (e.g., subtracting 4€ from a group member when deciding to invest 1€ for it). In a similar vein, participants were informed that their group members would also have the option to costly reduce their payoff. At the end of the same description page, participants would then be randomly assigned to either read that (1) at the end of the game they would get to know how the other players retrospectively evaluate their contributions and how they would behave in future interactions like this, versus (2) at the end of the game, they would not get to know how the other players retrospectively evaluate their contributions or how they would behave in future interactions like this. These wordings constituted our two experimental conditions (availability of consequentialist information/no availability of consequentialist information).

On the next page, participants were asked comprehension questions on the public goods game (for details, see Exclusion Criteria section and Instructions in Appendix I). Next, participants decided how much they wished to contribute to the public good (a round number between 0 and 400 Cents). Once all participants had made their decisions, participants were explained that they would now see everyone's contribution behavior and that they could decide to reduce a player's payoff by spending part of their money if they liked. Depending on the experimental condition, participants then read once again (1) that they would get to know afterwards how the other players retrospectively evaluate their contributions and how they would behave in future interactions like this, or (2) that they would not get to know afterwards about these things. On the next page, participants saw the respective information on every player's contribution behavior in their group, including their own (labeled as Player A, B, C, and D). This information was provided in absolute terms in Cents, as well as in relative terms to all contributions depicted in a pie chart.

Next, participants were asked to decide if they wished to deduct money from at least one other group member. They were reminded once again that they (1) would or (2) would not get to know afterwards how the other players retrospectively evaluate their behavior. If participants decided that they would like to deduct money from at least one other group member, they indicated separately for each player how much money exactly (in Cents) they wished to invest from their current payoff in order to subtract the fourfold amount from that player.

Participants would see a summary of how much money they had deducted from whom, followed by a page on which they would see a general summary about how much money they had at the beginning of the game after the contribution stage, how much costs had occurred because they had decided to reduce other players' payoffs, how much money was deducted from their own payoff because of the decisions made by all other players, as well as their resulting final payoff.

Lastly, since all interactions were real, participants in the information condition answered two questions about how much they regretted their contribution behavior (on a continuous slider ranging from not at all to very much) and how much they would contribute in future interactions like this (on a continuous slider ranging from nothing to everything, displaying the participant's contribution as a starting point at the corresponding part of the slider). These two answers served as consequentialist information that was then actually displayed to the other participants on the next page to comply with completely deception-free study standards. After having completed all stages of the public goods game, participants answered the manipulation check (if they had received information from the other players on how they retrospectively evaluated their behavior or not) along with a second item (indicating whether they had received information on their partners' contributions). Participants were thanked and debriefed and could leave feedback in an open-ended textbook. They received the payoff that resulted from their interaction behavior in the public goods game as payment plus an additional general participation payment of 2€ a few days later via bank transfer.

Table 1. Means, standard deviations (*SD*), and intercorrelations

Variable	Scale	Mean (*SD*)	Correlations					
			1.	2.	3.	4.	5.	6.
1. Punishment	0, 1	30.54%						
2. Punishment investments	≥ 0, in Cents	20.51 (49.93)	.62**					
3. Relative punishment investments	0–100	4.16 (12.06)	.52**	.96**				
4. Contributions	0–400 Cents	286.07 (137.27)	.16*	.21**	.20**			
5. Averaged difference of contributions	± 400 Cents	16.31 (161.36)	.29**	.34**	.36**	.86**		
6. Social value orientation angle	In degree (−16.26°–61.39°)	25.70 (13.70)	.07	.01	.00	.12	.11	
7. Honesty-Humility	0–5	3.54 (0.67)	−.07	−.05	−.02	.10	.08	.27**

Note. For the binary variable Punishment, we report point-biserial correlations. Number of observations for 1–4: $N = 203$, for 5–6: $N = 202$. $^{+}p < .1$; $^{*}p < .05$; $^{**}p < .01$ (two-sided).

Analyses and Results

Preprocessing

Exclusion Criteria

To assure sufficient game understanding, comprehension check questions to the public goods game were asked at the very beginning of the study and needed to be answered correctly in order to continue with the study. Participants read the description of the public goods game and were asked which contribution (a) maximizes their own outcome (correct response: none at all), (b) maximizes the group's outcome (correct response: contributing everything), and (c) how much they have to invest in order to substract 1€ from a group member (correct response: 25 Cents, see original materials, https://osf.io/mpcyw/). Up to three incorrect answers were allowed; after the third incorrect response, instructions were displayed again (the relevant parts being marked in bold). In case some participants were still not able to answer the control questions correctly, they were allowed to take part (in order to enable the remaining three participants to continue with the study), but their data were excluded from the main sample reported here.[3]

As preregistered, we did not exclude participants who answered the manipulation check at the end of the game incorrectly ($n = 28$ from the control condition, and $n = 4$ from the experimental condition). The findings do not differ when including the subgroup with incorrect answers to the manipulation check in the analyses, therefore we report the main sample here ($N = 203$), as preregistered in the Stage-1 protocol.

Transformations

Even when punishing to a similar extent (i.e., investing the same amount of resources on absolute terms), punishment investments differ in their severity for each participant depending on the outcome of the first stage of the public goods game (i.e., after each of the four group members made their contributions). To account for this relative difference, we transformed as preregistered punishment investments as a percentual investment to the earnings from the public good. Otherwise, no transformations were made. Note that punishment investments as a continuous variable also contained the binary decision, whether or not to punish at all, as the no-punishment decisions were coded as zero (Cents) investment.

Descriptives and Preliminary Analyses

For an overview of descriptive statistics and zero-order correlations, see Table 1.

On average, participants contributed almost three quarters of their endowment to the public goods game ($M = 286.07$ Cents, $SD = 137.27$, min = 0, max = 400). The punishment rate was relatively low, participants' binary decision to punish indicated that 141 participants (69.46%) did not decide to punish at all. The continuous punishment investments were correspondingly small (absolute $M = 20.51$ Cents, $SD = 49.93$, min = 0, max = 400; as percental investments relative to earnings from the public good $M = 4.16\%$, $SD = 12.06$, min = 0, max = 100). From the 62 participants who decided to punish, the absolute continuous investments ranged between 1 and 400 Cents ($M = 67.15$, $SD = 71.22$), the investments relative to earnings from the public good being on average 13.62% ($SD = 18.72$). Participants' final average payoff of the game (without the additional participation payment of 2€) was 561.10 Cents ($SD = 202.59$, min = 0, max = 925). As an indicator of altruistic (vs. antisocial) punishment, participants' relative punishment investments were positively related to how much participants' contributions had differed from the group's average, $r = .357$, $p < .001$, replicating earlier findings that participants who contributed above the average of their group members also punished more severely (see Mischkowski et al., 2018).

[3] As preregistered, we report findings from the whole sample online (see https://osf.io/mpcyw/). We checked for a systematic dropout between conditions and indeed found more participants in the experimental condition with a lack of game comprehension ($p = .032$). Note, however, that results are robust and do not change when including all 276 participants (i.e., those with a lack of game understanding).

Main Analyses and Results

To test the first hypothesis (H1), whether participants who *receive information on the consequences of punishment on the transgressor* punished more in that they invest more resources to punish than participants who know they will not receive this information, punishment investments were regressed on the information condition. Importantly, we controlled for the averaged group members' contribution behavior relative to the own contributions to the public good (i.e., the difference between own and averaged group members' contributions) as this reflects the degree of inequality in contributions that most likely influences the severity of punishment (see Mischkowski et al., 2018). By keeping its influence constant, we could independently assess the effect of information availability. There was no significant effect of experimental condition on relative punishment investments ($\beta = -.035$, $p = .595$), only the above described relation to participants' relative contribution behavior significantly predicted participants' relative punishment investments ($\beta = .357$, $p < .001$). Thus, the data did not allow to reject the Intuitive Retributivism Hypothesis: Punishment behavior did not increase when consequentialist information was provided.

To test the interaction effect (H2), that prosocials punish more severely when consequential motives are addressed as compared to when they are not, we conducted two analyses. First, we regressed the punishment magnitude on the experimental manipulation of information availability, SVO, and their interaction term (H2a), controlling for an individual's difference to the averaged group members' contribution behavior (see above). SVO was centered on the sample mean. A significant interaction would reflect that prosocials differ in comparison to proselfs with regard to their punishment magnitude when consequentialist information is given. The data did not reveal any significant effect of condition ($\beta = -.038$, $p = .563$), SVO ($\beta = .026$, $p = .780$), or the interaction of condition and SVO ($\beta = -.096$, $p = .306$) on participants' relative punishment investments. Again, only participants' relative contribution behavior positively predicted punishment ($\beta = .367$, $p < .001$).

Similarly, when conducting the same regression model with Honesty-Humility (HH) instead of SVO to assess dispositional prosociality with a broader basic trait (H2b), there was neither a significant effect of condition ($\beta = -.034$, $p = .612$), HH ($\beta = -.070$, $p = .416$), nor of the interaction of condition and HH ($\beta = .031$, $p = .713$) on participants' relative punishment investments. Once again, only participants' relative contribution behavior positively predicted punishment ($\beta = .365$, $p < .001$).

Since the distribution of punishment investments is highly skewed, we analyzed all hypotheses with punishment decision as binary dependent variable as a non-preregistered, exploratory robustness check. Results remain similar (all $p > .300$).

Discussion

The debate on the impact of retributive vs. consequentialist punishment motives is long-lasting and has been investigated with various approaches; for instance, by investigating the influence of motive-congruent information (e.g., Carlsmith, 2008), by tracking down information search behavior (e.g., Carlsmith, 2006), next to investigating punishers' hedonic reactions (e.g., Funk et al., 2014). While the first two areas of research have found support for the Intuitive Retributivism Hypothesis, findings from research on punishers' hedonic reactions have challenged it, as punishers' justice-related satisfaction depends on the effects of punishment. Punishers expect to be and indeed are more satisfied after punishment if transgressors understand that the bad treatment they receive is punishment for their wrongdoing (Gollwitzer et al., 2011; Funk et al., 2014, Study 1). In addition to a transgressor's mere understanding, punishers also react positively if they find out about the positive effect of punishment on the transgressor (Funk et al., 2014, Study 2).

In congruence with findings on the hedonic reactions, earlier studies have found that people's actual punishment decisions are affected by whether transgressors are or are not able to realize that the bad treatment they receive is punishment (for instance, when punishment is "open" or "hidden," see Crockett et al., 2014; or when punishers can make sure their message is received, see Molnar et al., 2020). In the present study, we aimed at further challenging the Intuitive Retributivism Hypothesis by assessing how the outlook to receive information about the effects of punishment (i.e., consequentialist information) affects participants' actual punishment decisions. From a retributive perspective, the availability of such information should not affect people's punishment behavior. And indeed, our data did not confirm this consequentialist alternative hypothesis. When it comes to actual punishment decisions, punishers were unaffected by whether they would or would not find out about the effects of their punishment decision on the transgressor. In addition, we could not identify this hypothesized effect to be conditional on dispositional prosociality (i.e., individuals high in SVO or Honesty-Humility) in that prosocials did not increase their punishment behavior more strongly than proselfs when consequentialist information was provided. Thus, our data did not provide any evidence that prosocials' punishment behavior may be particularly pronounced when consequentialist motives are addressed. As a consequence, future research is needed to reconcile heterogeneous results regarding the relation

between dispositional prosociality and punishment behavior (see, e.g., Bieleke et al., 2016; Böckler et al., 2016; Karagonlar & Kuhlman, 2013).

In sum, the present findings do not replicate and extend earlier findings regarding people's hedonic reactions to people's actual punishment behavior, and our findings do not allow us to reject the Intuitive Retributivism Hypothesis. A first possible explanation for why we could not conceptually replicate the effects on hedonic reactions with participants' actual punishment decisions is that, in the present study, manipulating the availability of consequentialist information might have been linked to a variable that we did not consider in our design: When participants learned that they would receive information on the effects of punishment, they knew that they would also potentially receive information that their punishment has had no effect. Thus, our experimental manipulation may also have increased the *salience of uncertainty* in regards to whether punishment would be able to bring about an effect. Such a sense of uncertainty was not present in studies on hedonic effects of punishment: Once participants' in those studies indicated their satisfaction, they already knew the exact effect of punishment on the transgressor (e.g., as indicated by the transgressor's feedback message, see Funk et al., 2014). By reversing any potential positive effect that consequentialist motives might have on people's punishment decisions, such uncertainty could have contributed to low punishment rates in the present study. In order to investigate this possibility, future studies should directly manipulate the degree to which punishment is likely to cause a change in the transgressor (thus manipulating a decision maker's sense of uncertainty) along with the availability of information, and examine how both factors affect people's actual punishment decisions.

A second possible explanation for the mismatch between findings of hedonic reactions versus actual punishment decisions is that those earlier findings on the hedonic effects may not be linked to "punishment" motives, but rather to more general "justice motives." It is possible that transgressor change generally has positive effects on people's hedonic reactions (e.g., Funk et al., 2014), independent of whether oneself has actually punished and as such elicited the transgressor change. Similarly, other lines of research have found that a renewed sense of value consensus with the transgressor restores a sense of justice in victims (see, e.g., Justice Restoration Theory; Wenzel et al., 2008; Okimoto & Wenzel, 2009). Thus, that justice-related satisfaction does not differ between people who punish and people who do not (e.g., Funk, 2015; Funk et al., 2014) can be interpreted as a challenge to the Intuitive Retributive Hypothesis, yet it could also suggest that interpreting hedonic reactions after punishment as an indicator for punishment motives is problematic.

With regard to the chosen paradigm and operationalization, future studies need to test if the null-effect that we found using a public goods game replicates in other settings, for instance, when using other kinds of economic games (e.g., punishment as the rejection of unfair offers in ultimatum games). In a similar vein, it remains to be examined whether the outlook to receive consequentialist information influences third-party punishment. Specifically, third-party punishment can be linked to moral indignation (Camerer, 2003), whereas second-party punishment (as investigated here) has been shown to be related to retributive anger (Mischkowski et al., 2018). Since consequentialist information might be particularly important to morally indignant external punishers, its influence might increase in third-party punishment settings. Thus, future research examining further (e.g., third-party) punishment settings is needed to challenge the Intuitive Retributivism Hypothesis. Lastly, punishment researchers should continue to combine a broad array of different approaches – including internally valid laboratory studies (e.g., economic games) as well as other externally generalizable methods (e.g., vignette or field studies) in which the different nuances of real-world experiences can be reflected to a broader extent. Importantly, as the present findings illustrate, because hedonic reactions after punishment and actual punishment decisions may diverge, it is crucial to study various punishment measures and not only focus on one of them when investigating people's punishment motives.

Limitations

There are also limitations of the present study that should not go unnoticed. First, it is possible that we could not identify any effect of consequentialist information on punishment decisions because the sample consisted of many people who decided not to punish at all (around 70%). The present sample behaved very prosocially, to begin with, in that the original contribution behavior was quite high. This is an interesting finding in itself and not necessarily a limitation, yet it might have led to the low punishment rate displayed by the participants: Participants simply did not have much reason to punish.[4]

[4] However, these sample characteristics have also been found in other punitive economic games on prosocial punishment (Mischkowski et al., 2018). Next to these descriptive results, we replicate two findings from Mischkowski and colleagues (2018), showing the meaningfulness of our data. First, participants who contributed above the average of their group members also punished more severely. Second, we again found no relation between SVO and punishment; yet since this is true for both of the conditions, this comes at the cost that we could not identify a boundary condition for the relation between dispositional prosociality and punishment behavior.

Second, given that our manipulation check was captured *after* the dependent variable (i.e., punishment investments) and – even more importantly – after participants had seen the consequentialist information in the experimental condition, we cannot rule out that participants only realized which information they were given based on the actual information and not based on the preannouncement in the instructions. The reduced number of participants who failed to pass the manipulation check in the experimental condition (n = 4) in comparison to the control condition (n = 28) might be an indicator that our MC was incapable of capturing whether participants knew *at the point of making their punishment decision* whether or not they would receive consequentialist information.

If the manipulation was unsuccessful and too many participants in the control group thought that they would receive information on the effects of punishment when they made their punishment decision, not finding any differences in punishment between the two experimental conditions would come as no surprise. However, it is more likely that the high number of failed manipulation checks in the control condition reflects a problem with the manipulation check item (and not with the manipulation as such). Participants in the control group may have confused, for instance, the information regarding their group members' contribution behavior with information on others' intentions how to behave in future interactions when they responded to the manipulation check item *after* they made their punishment decision. In this case, the uneven distribution of failed manipulation checks would not bias the null effects we found.

Conclusion

We have presented findings from a preregistered experimental study in which participants interacted with each other in an incentivized public goods game. In the present study, we did not find an effect of one particular type of consequentialist information on people's actual punishment decisions. Neither could we identify this effect to be conditional on participants' dispositional prosociality. Instead, only participants' relative contribution behavior was consistently related to their subsequent punishment decisions.

While the current findings do not allow to reject the Intuitive Retributivism Hypothesis, they should also not be interpreted as direct "confirmation" of the hypothesis, however, because in the present study, the Retributivism Hypothesis served as the null hypothesis. As such, our findings do not preclude other interpretations that need to be contrasted in future research. Just as our findings are in line with the Intuitive Retributivism Hypothesis, it is possible that punishers eliminate the relative disadvantage and restore equality by punishing, as competitive motives would suggest (Raihani & Bshary, 2019). These motives need to be contrasted in future research, emphasizing in more general terms that the research area of people's punishment motives remains an exciting field.

References

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11*(2), 150–166. https://doi.org/10.1177/1088868306294907

Balliet, D., Parks, C., & Joireman, J. (2009). Social value orientation and cooperation in social dilemmas: A meta-analysis. *Group Processes and Intergroup Relations, 12*(4), 533–547. https://doi.org/10.1177/1368430209105040

Bieleke, M., Gollwitzer, P. M., Oettingen, G., & Fischbacher, U. (2016). Social value orientation moderates the effects of intuition versus reflection on responses to unfair ultimatum offers. *Journal of Behavioral Decision Making, 30*(2), 569–581. https://doi.org/10.1002/bdm.1975

Böckler, A., Tusche, A., & Singer, T. (2016). The structure of human prosociality: Differentiating altruistically motivated, norm motivated, strategically motivated, and self-reported prosocial behavior. *Social Psychological and Personality Science, 7*(6), 530–541. https://doi.org/10.1177/1948550616639650

Bone, J. E., & Raihani, N. J. (2015). Human punishment is motivated by both a desire for revenge and a desire for equality. *Evolution and Human Behavior, 36*(4), 323–330. https://doi.org/10.1016/j.evolhumbehav.2015.02.002

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

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

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

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

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

Cushman, F. A., Sarin, A., & Ho, M. (2022). Punishment as communication. In J. Doris & M. Vargas (Eds.), *Oxford handbook of moral psychology*. Oxford University Press.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology, 31*(1), 169–193. https://doi.org/10.1146/annurev.ps.31.020180.001125

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

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

Funk, F. (2015). *Beyond retribution: The role of transformative justice motives for people's reactions to wrongdoers* (Dissertation). Princeton University.

Funk, F., McGeer, V., & Gollwitzer, M. (2014). Get the message: Punishment is satisfying if the transgressor responds to its

communicative intent. *Personality and Social Psychology Bulletin, 40*(8), 986–997. https://doi.org/10.1177/0146167214533130

Gollwitzer, M., & Denzler, M. (2009). What makes revenge sweet: Seeing the offender suffer or delivering a message? *Journal of Experimental Social Psychology, 45*(4), 840–844. https://doi.org/10.1016/j.jesp.2009.03.001

Gollwitzer, M., Meder, M., & Schmitt, M. (2011). What gives victims satisfaction when they seek revenge? *European Journal of Social Psychology, 41*(3), 64–374. https://doi.org/10.1002/ejsp.782

Haruno, M., Kimura, M., & Frith, C. D. (2014). Activity in the nucleus accumbens and amygdala underlies individual differences in prosocial and individualistic economic choices. *Journal of Cognitive Neuroscience, 26*(8), 1861–1870. https://doi.org/10.1162/jocn_a_00589

Hilbig, B. E., Glöckner, A., & Zettler, I. (2014). Personality and prosocial behavior: Linking basic traits and social value orientations. *Journal of Personality and Social Psychology, 107*(3), 529–539. https://doi.org/10.1037/a0036074

Hilbig, B. E., Thielmann, I., Klein, S. A., & Henninger, F. (2016). The two faces of cooperation: On the unique role of HEXACO Agreeableness for forgiveness versus retaliation. *Journal of Research in Personality, 64*, 69–78. https://doi.org/10.1016/j.jrp.2016.08.004

Hilbig, B. E., Zettler, I., & Heydasch, T. (2012). Personality, punishment and public goods: Strategic shifts towards cooperation as a matter of dispositional honesty-humility. *European Journal of Personality, 26*(3), 245–254. https://doi.org/10.1002/per.830

Karagonlar, G., & Kuhlman, D. M. (2013). The role of social value orientation in response to an unfair offer in the ultimatum game. *Organizational Behavior and Human Decision Processes, 120*(2), 228–239. https://doi.org/10.1016/j.obhdp.2012.07.006

Ledyard, J. O. (1995). Public goods: A survey of experimental research. In J. H. Kagel & A. E. Roth (Eds.), *The handbook of experimental economics* (pp. 111–194). Princeton University Press.

Lee, K., & Ashton, M. C. (2018). Psychometric properties of the HEXACO-100. *Assessment, 25*(5), 543–556. https://doi.org/10.1177/1073191116659134

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO Personality Inventory. *Multivariate Behavioral Research, 39*(2), 329–358. https://doi.org/10.1207/s15327906mbr3902_8

Mischkowski, D., Glöckner, A., & Lewisch, P. (2018). From spontaneous cooperation to spontaneous punishment – Distinguishing the underlying motives driving spontaneous behavior in first and second order public good games. *Organizational Behavior and Human Decision Processes, 149*, 59–72. https://doi.org/10.1016/j.obhdp.2018.07.001

Molnar, A., Chaudhry, S., & Loewenstein, G. F. (2020, January 24). *"It's not about the money. It's about sending a message!": Unpacking the components of revenge*. https://doi.org/10.2139/ssrn.3524910

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. J. (2011). Measuring social value orientation. *Judgment and Decision Making, 6*(8), 771–781. http://journal.sjdm.org/11/m25/m25.pdf

Okimoto, T. G., & Wenzel, M. (2009). Punishment as restoration of group and offender values following a transgression: Value consensus through symbolic labelling and offender reform. *European Journal of Social Psychology, 39*, 346–367. https://doi.org/10.1002/ejsp.537

Pletzer, J. L., Balliet, D., Joireman, J., Kuhlman, D. M., Voelpel, S. C., & Van Lange, P. A. (2018). Social value orientation, expectations, and cooperation in social dilemmas: A meta-analysis. *European Journal of Personality, 32*(1), 62–83. https://doi.org/doi.org/10.1002/per.2139

Raihani, N. J., & Bshary, R. (2019). Punishment: One tool, many uses. *Evolutionary Human Sciences, 1*, Article e12. https://doi.org/10.1017/ehs.2019.12

Sarin, A., Ho, M. K., Martin, J., & Cushman, F. A. (2020, March 26). *Punishment is organized around principles of communicative inference*. https://doi.org/10.31234/osf.io/2cyf7

Thielmann, I., Spadaro, G., & Balliet, D. (2020). Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin, 146*(1), 30–90. https://doi.org/10.1037/bul0000217

Van Lange, P. A. M. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology, 77*(2), 337–349. https://doi.org/10.1037/0022-3514.77.2.337

Van Lange, P. A. M., Joireman, J. A., Parks, C. D., & van Dijk, E. (2013). The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes, 120*, 125–141. https://doi.org/10.1016/j.obhdp.2012.11.003

Wenzel, M., Okimoto, T. G., Feather, N. T., & Platow, M. (2008). Retributive and restorative justice. *Law and Human Behavior, 32*, 375–389. https://doi.org/10.1007/s10979-007-9116-6

Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences, 102*, 7398–7401. https://doi.org/10.1073/pnas.0502399102

Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., Miura, A., Inukai, K., Takagishi, H., & Simunovic, D. (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proceedings of the National Academy of Sciences, 109*(50), 20364–20368. https://doi.org/10.1073/pnas.1212126109

History
Received March 31, 2020
Revision received February 5, 2021
Accepted April 23, 2021
Published online April 1, 2022

Acknowledgments
The research reported in this article was preregistered at the OSF (https://osf.io/mpcyw). We declare that we did not deviate from our preregistration in either data collection or analyses.
We would like to thank Léon Bartosch for his excellent research assistance.

Publication Ethics
Our study is in line with the ethical guidelines of the Deutsche Gesellschaft für Psychologie (German Association for Psychology) and was approved by the ethics committee of the Faculty of Human Sciences of the University of Cologne (Ethics proposal "DMHF0084").

Authorship
Friederike Funk and Dorothee Mischkowski share the first authorship and contributed equally, author names are listed in alphabetical order. You may also contact Friederike Funk, Faculty of Arts & Sciences, NYU Shanghai, 1555 Century Avenue, Pudong New Area, Shanghai, 200122, China; friederike.funk@nyu.edu

Open Data
Our preregistration (Stage 1 protocol), power analyses, study material, and raw data (including codebook) are available at the OSF at https://osf.io/mpcyw/.

Funding
Open access publication enabled by the University of Cologne.

ORCID
Friederike Funk
https://orcid.org/0000-0002-6514-3235

Dorothee Mischkowski
https://orcid.org/0000-0002-7563-402X

Dorothee Mischkowski
Social Cognition Center Cologne
Faculty of Human Sciences
University of Cologne
Richard-Strauss-Str. 2
50931 Cologne
Germany
dorothee.mischkowski@uni-koeln.de