Molnar, Andras; Chaudhry, Shereen J.; Loewenstein, George

Working Paper

"It's Not about the Money. It's about Sending a Message!" Unpacking the Components of Revenge

CESifo Working Paper, No. 8102

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Molnar, Andras; Chaudhry, Shereen J.; Loewenstein, George (2020) : "It's Not about the Money. It's about Sending a Message!" Unpacking the Components of Revenge, CESifo Working Paper, No. 8102, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/215104

CESifo WORKING PAPERS

8102
2020

February 2020

"It's Not about the Money. It's about Sending a Message!" Unpacking the Components of Revenge

Andras Molnar, Shereen J. Chaudhry, George Loewenstein

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 8102*

"It's Not about the Money. It's about Sending a Message!" Unpacking the Components of Revenge

Abstract

We examine whether belief-based preferences - caring about what transgressors believe - play a crucial role in punishment decisions: Do punishers want to make sure that transgressors understand why they are being punished, and is this desire to affect beliefs often prioritized over distributive and retributive preferences? We test whether punishers derive utility from three distinct sources: material outcomes (their own and the transgressor's payoff), affective states (the transgressor's suffering), and cognitive states (the transgressor's beliefs about the cause of that suffering). In a novel, preregistered experiment (N = 1,959) we demonstrate that consideration for transgressors' beliefs affects punishment decisions on its own, regardless of the considerations for material outcomes (distributional preferences) and affective states (retributive preferences). By contrast, we find very little evidence for pure retributive preferences (i.e., to merely inflict suffering on transgressors). We also show that people who would otherwise enact harsh punishments, are willing to punish less severely, if by doing so they can tell the transgressor why they are punishing them. Finally, we demonstrate that the preference for affecting transgressors' beliefs cannot be explained by deterrence motives (i.e., to make transgressors behave better in the future).

JEL-Codes: C910, D630, D820, D830.

Keywords: beliefs, belief-based utility, justice, fairness, morality, punishment, revenge.

*Andras Molnar**
Department of Social and Decision Sciences
Carnegie Mellon University
5000 Forbes Ave
USA – Pittsburgh, PA 15213
andrasm@andrew.cmu.edu

Shereen J. Chaudhry
Booth School of Business
University of Chicago
5807 S. Woodlawn Ave
USA – Chicago, IL 60637
shereen.chaudhry@chicagobooth.edu

George Loewenstein
Department of Social and Decision Sciences
Carnegie Mellon University
5000 Forbes Ave
USA – Pittsburgh, PA 15213
gl20@andrew.cmu.edu

*corresponding author

January 30, 2020

Author Contributions. All authors contributed to the development of the study concept and experimental design. Testing, data collection, and data analyses were performed by A. Molnar. All authors worked on the manuscript and provided critical revisions.

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations of interest. None.

Acknowledgments. The authors with to thank Russell Golman, Alex Imas, and Daniel Read for their valuable feedback.

> "Revenge is a very personal matter, and when it is inflicted, it is important that the target grasp the reason why. If the target does not know that he or she is paying the penalty because of his or her specific prior harming or injuring of someone or of the avenger himself or herself, the act of revenge has misfired."
>
> — *The virtues of vengeance* (French, 2001, p. 12)

INTRODUCTION

Imagine that you work at a company where one of your colleagues, Bob, has repeatedly claimed credit for work that you did, and as a result, is very likely to receive a large year-end bonus. Luckily, you are moving to a different state for a new job next week, so you won't ever have to work with him again. But before you leave, you are in a position to pay him back: You are about to submit the annual performance evaluations of your peers, and know that if you write a negative review of Bob, it will negatively impact his bonus. However, imagine that Bob would never find out that his bonus was reduced because of his bad behavior, or because of your review. Would your desire for payback be satisfied by writing that review? Would it be more satisfying if you knew Bob would learn the reason for his reduced bonus, even if you would never work with him again?

If Peter French is right in the quote above from his book on revenge, then most people would likely prefer that Bob know the reason he was punished, even if they received no material benefit from doing so. This intuition is ubiquitous in our culture: Countless classic "revenge stories"—e.g., *The Odyssey*, *Hamlet*, *The Count of Monte Cristo*, *Once Upon a Time in the West*, and *Gladiator*—depict protagonists going to great lengths, even risking their lives, just to make sure that the antagonists learn *why*, and *by whom*, they are punished. Doing so does not confer any instrumental value to the protagonist: Typically, the punishment itself is fatal; the antagonist meets their maker shortly after learning by whom and for what reason they are being punished, which leaves no room for deterrence (preventing them from transgressing again). Yet, as people in the audience witness these events unfold, most of them would feel utterly dissatisfied and frustrated if the protagonist just enacted their revenge, sending the transgressor to his grave with no awareness of the reason for his death. In this paper, we test whether, consistent with French's assertion and his book on revenge, as well as with the plot-line of so many stories, in taking revenge people are

motivated to affect the transgressor's *beliefs*. Consistent with beliefs mattering, we find that people who would be ready to punish severely are willing to reduce the magnitude of the punishment if in doing so they can increase the transgressor's awareness that they are being punished.

While punishment, or negative reciprocity, is one of the most well-studied topics in experimental and behavioral economics (for example, Fehr and Gächter's seminal paper on altruistic punishment has over 5,300 citations on Google Scholar, Fehr & Gächter, 2002), very little research has examined whether there are *belief-based motives* behind it. Most work has focused on drivers such as preferences for fairness in the distribution of material outcomes (e.g., Fehr & Gächter, 2000; Johnson, Dawes, Fowler, McElreath, & Smirnov, 2009; Raihani & McAuliffe, 2012), the desire to enforce, and conform to, social norms (e.g., Fehr & Fischbacher, 2004; Boyd, Gintis, & Bowles, 2010), the motivation to maintain a good reputation (e.g., Barclay, 2006; Kurzban, DeScioli, & O'Brien, 2007; Santos, Rankin, & Wedekind, 2011; Raihani & Bshary, 2015), and a desire to regulate one's emotions (e.g., Gollwitzer & Bushman, 2012; Dickinson & Masclet, 2015; Jordan, McAuliffe, & Rand, 2015).

Furthermore, most of the previous literature likely overestimated the desire for pure retribution or the restoration of distributive justice, because experiments allowed for *both* outcome-based and belief-based motives: Participants—both victims, transgressors, and punishers—had perfect information about the potential payoffs and their actual outcomes (e.g., Fehr & Gächter, 2002 and its replications; Fehr & Fischbacher, 2004; Henrich et al., 2006). This means that whenever a participant was punished for his or her misbehavior, it was immediately obvious to him or her not only that his or her payoff had been reduced, but also why it had been reduced. Even in anonymous settings, transgressors were fully aware that their partners (or fellow group members) were responsible for punishing them. However, in real-world situations, this information is often not obvious and transparent: People who have been punished often don't realize that they have been punished, or, if they do, why they have been punished or by whom.

In this paper, we examine the role of belief-based motives by introducing a paradigm that produces the type of asymmetric information that is common in real-world situations and that is necessary to detect such motives. In doing so, we disentangle the belief-based motives behind revenge from distributional preferences and even preferences over the emotional states of the transgressor.

Three Motives behind Punishment

Preference for material states (distributive justice). The first, and most straightforward motive for punishment is distributive justice: the desire to restore a fair allocation of resources (e.g., Bolton & Ockenfels, 2000; Charness & Rabin, 2002; Fehr & Gächter, 2000). People who care *only* about distributive justice are agnostic about how the transgressor feels and or what the transgressor believes, since they only care about the (reduction of the) objective welfare of the transgressor. A person who cares about material outcomes only, would be equally motivated to impose a "hidden" punishment (i.e., when the transgressor remains ignorant about the fact that he was punished) and an "open" punishment (i.e., when the transgressor is fully aware that he was punished). This also implies that the ability to inform the transgressor about *why* and *by whom* was the punishment enacted should not influence the punisher's choice.

Preference for affective states (retributive justice). In addition to the objective welfare of the transgressor, the punisher might also care about how the transgressor *feels* about the outcome. The same outcome might bring joy or suffering, depending on what the transgressor knows about the context (e.g., salient alternatives, awareness of a loss), and the punisher might want to make sure that the transgressor suffers sufficiently for their misbehavior, with the amount of suffering experienced by the transgressor being roughly equivalent to the amount of suffering they caused ("comparative suffering", see Gollwitzer, Meder, & Schmitt, 2011). This proportionality principle is also a key component of Just Deserts theory (Kant, 1952; Carlsmith, Darley, & Robinson, 2002): Transgressors deserve punishment proportional to the moral wrong they commit.

If an individual cares about both the objective material welfare and the subjective feelings of the transgressor, the punisher's decision depends on the expected affective reaction of the transgressor as well. If the punisher cares about the transgressor's affective reaction strongly enough, the punisher might even prioritize retributive justice over distributive justice, that is, let the transgressor get away with a less harsh punishment, as long as doing so will ensure that the transgressor suffers. For example, if the punisher is forced to choose between a harsh but hidden punishment (which would restore distributive justice but would not make the transgressor suffer) and a mild but open punishment (which would only partially restore distributive justice, but would make the

transgressor suffer), she might prefer the latter, even if ceteris paribus she would prefer to enact a harsher punishment. However, it is important to note that this individual would still not care about the transgressor's beliefs *directly*; only to the extent that such beliefs would make the transgressor suffer. That is, this individual would be agnostic about the source of the transgressor's suffering—would not care *why* the transgressor suffers or whether he or she understands the reason for their misfortune.

Preference for cognitive states (belief-based motives). Finally, the punisher might also care directly about the transgressor's beliefs, independent of his or her objective material welfare and suffering. The idea that people have preferences over their own and others' beliefs, regardless of the instrumental value of such beliefs, has many applications in economics (see, e.g., Bénabou & Tirole, 2016; Loewenstein & Molnar, 2018; Battigalli, Corrao, & Dufwenberg, 2019). Here, we focus on a specific category of beliefs: The transgressor's *understanding* of why they were punished. In particular, we hypothesize that punishers care directly about whether the transgressor shares a common understanding of the situation with the them (i.e., the punisher), consistent with a preference for belief consonance (Golman, Loewenstein, Moene, & Zarri, 2016).

In psychology, there is a line of research showing that victims are more satisfied if they can explain why they punished the transgressor, especially if the transgressor also acknowledges that he or she understands (the "expressive function" of punishment, Feinberg, 1965; the "understanding hypothesis", Gollwitzer et al., 2011; Funk, McGeer, & Gollwitzer, 2014). However, these papers have not investigated actual revenge behavior (only hypothetical), nor have they examined how people make trade-offs with respect to the various motives behind revenge—that is, it is unclear how much weight punishers place on impacting the transgressor's beliefs relative to concerns about distributive and retributive justice.

Similarly, the research on the role of symbolic punishment in economic games (e.g., Masclet, Noussair, Tucker, & Villeval, 2003; Xiao & Houser, 2005) has looked only at whether people use symbolic punishment when that is their *only* option to punish transgressors. This work does not investigate whether people would make trade-offs between different punishment goals. Would punishers be willing to compromise on the restoration of distributive and retributive justice in order to impact the transgressor's beliefs?

STUDY 1: UNCONSTRAINED HYPOTHETICAL CHOICE

Before we investigate whether people are willing to make trade-offs between different motives, we first look at punishers' behavior in an unconstrained setting. In Study 1, participants were presented with a hypothetical scenario in which they were treated unfairly by another person and could freely choose whether, and to what extent, to reduce this other person's payoff. They could also choose whether to send a message to this other person, and if they decided to do so, they could select from four preset messages, which differed in the amount of information conveyed about the punishment (i.e., whether they were punished, why they were punished, etc.). If punishers only care about the material welfare of the transgressor, then they should be indifferent about what the transgressor believes and therefore also be indifferent about the type of message to be sent to the transgressor. By contrast, if punishers do care about the affective and cognitive states of the transgressor, then we should see a tendency towards selecting more informative messages, i.e., messages that let the transgressor understand why their payoff has been reduced, and messages that let the transgressor know who reduced their payoff.

Methods

Participants. We recruited participants from two separate online participant pools: Amazon Mechanical Turk (MTurk) and Prolific. Since this study was exploratory, and did not involve any between-subjects manipulation or planned hypothesis testing, it was not preregistered. We aimed to collect 100 responses on each platform (i.e., 200 total), and ended up recruiting 201 participants: 100 participants on MTurk and 101 participants on Prolific.[1] Among the 201 recruited participants, we excluded 7 (3.5%) who failed the attention check question. The final sample contained 194 responses: 96 from MTurk (44% female; age: $M = 38.6$ years) and 98 from Prolific (45% female; age: $M = 32.5$ years).

Procedure. Eligible participants were directed to a Qualtrics survey (*Qualtrics*, 2019), in which they were instructed to read a hypothetical scenario. In this scenario the participant was interacting with another person who was responsible for the allocation of work (a numerical addition

[1] We recruited the latter group after we finished data collection on MTurk, and excluded those Prolific workers from participation who also had an MTurk worker account.

task) between him or her and the participant. In the scenario, both the participant and the other person received a fixed compensation for the work, regardless of how the task was split between them. Then, the participant was told that the other person had allocated 85% of the work to them (i.e., to the participant), which, given the fixed compensation, would be considered very unfair.

After learning about the unfair allocation of work, the participant had to indicate how he or she would have felt if this happened to them. First, the participant had to select from the following list of 12 feelings (presented in a randomized order): *amused, angry, betrayed, disappointed, envious, grateful, happy, pleased, proud, sad, satisfied, surprised.* Then, for each feeling that the participant selected, they had to indicate how strongly they would have experienced that feeling (continuous scale from *1: A little* to *7: Extremely*).

Next, the participant was informed that the other person was going to receive a surprise bonus of $1. Before receiving this bonus, however, the participant had the opportunity to reduce this bonus and to send a message to the other person. Using a slider, the participant could adjust the other person's bonus to any amount X between $0 and $1 (including $0 and $1). Simultaneously (on the same screen), the participant could choose one of the following five options, presented in a randomized order:

- Do NOT send any message. [ignorance]
- Send a message: "You received an extra bonus of $X." [bonus only]
- Send a message: "You received an extra bonus of $X out of $1." [suffering]
- Send a message: "You received an extra bonus of $X out of $1 because you were unfair to your partner." [justice]
- Send a message: "You received an extra bonus of $X out of $1. Your partner decided to reduce your extra bonus because you were unfair to them." [revenge]

As the participant adjusted the slider, the amounts X for all the options changed at the same time, and the individual could adjust the slider and choose between the options at will before clicking on "next" (see Figure 1). In subsequent sections we will refer to the above messages by the corresponding labels (in brackets; these labels were not displayed to participants).

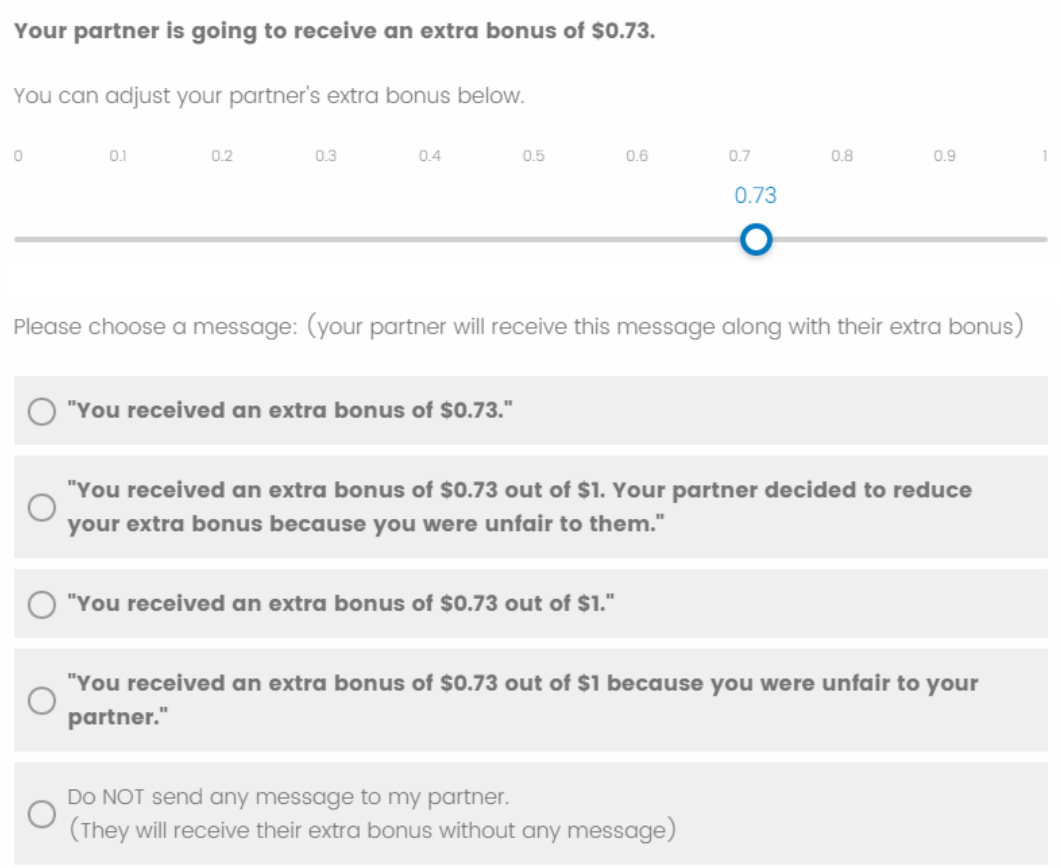

Figure 1: Main decision screen in Study 1: Participants could freely adjust the other person's payoff and could choose between not sending any message or sending one of the four preset messages.

After the participants selected the amount and the message, they wrote an open-ended explanation for their choices and answered the attention check question. Finally, we recorded participants' sex, age, and political affiliation.

Results

Punishment choices. Pooled across both samples, the overwhelming majority of participants, 169 people (87%), decided to reduce the other person's payoff (84% in the Prolific sample and 91% in the MTurk sample). The average reduction (including people who did not reduce) was \$0.76, leaving the unfair person only \$0.24 on average. The most frequently chosen amount (chosen by 40% of participants) was the full reduction (i.e., −\$1, leaving the partner \$0). These results indicate a strong overall preference for distributive justice, i.e., a desire to reduce the transgressor's material welfare. Figure 2 shows the full distribution of choices across the two subject pools.

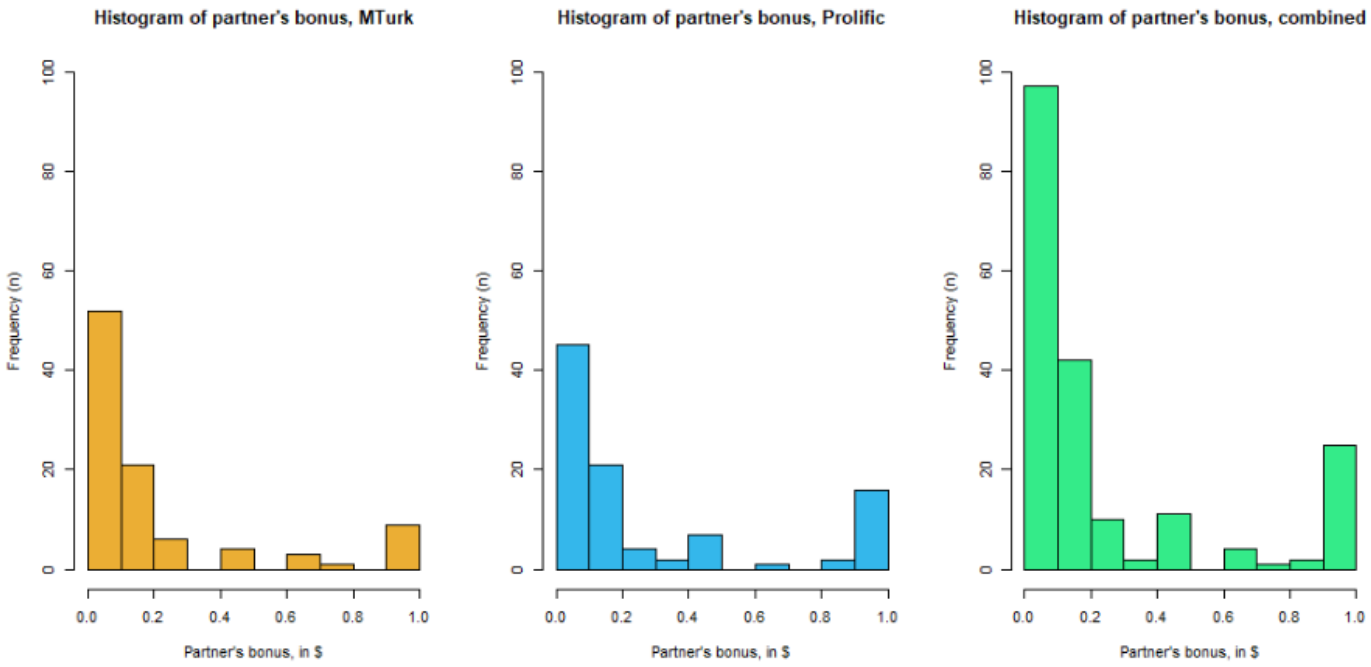

Figure 2: Histograms of punishment choices in Study 1 by subject pool: MTurk (left panel), Prolific (middle panel), and MTurk+Prolific combined (right panel). Lower values along the X-axes indicate harsher punishments, i.e. larger reductions of the unfair partner's bonus.

Message choices. The most frequently chosen message was the revenge message ("You received an extra bonus of $X out of $1. Your partner decided to reduce your extra bonus because you were unfair to them."): 81 participants (42%) chose this message. Thirty-seven participants (19%) chose not to send any message (ignorance), while 30 (15%) sent the bonus only message. Finally, 28 participants (14%) sent the justice message, and 18 (9%) sent the suffering message (see Figure 3, column 1). The above proportions simply capture the popularity of messages among *all* participants, regardless of their punishment decisions. However, it is more meaningful to separate the people who reduced their partner's bonus from those who did not, and investigate the distribution of messages within these two group separately.

Among the 25 participants who did not reduce their partner's payoff (Figure 3, column 2)., the vast majority (22, 88%) chose either the no message (5, 20%), the bonus only message (12, 48%), or the "suffering" message (5, 20%), the last of which simply informed the partner that they have received the full bonus (thus, receiving this message did not cause any suffering in this case). Only very few participants (3, 12%) chose either the justice or revenge messages.[2]

[2]Based on the open-ended explanations these participants provided, these choices were deliberate, and not mis-

Number (and proportion) of participants choosing each message, both participant pools (MTurk + Prolific)

Message	Participants							
	(1) Everyone (any X)		(2) Did not punish (X = $1)		(3) Punished (X < $1)		(4) Fully punished (X = $0)	
Ignorance (no message)	37	19%	5	20%	32	19%	18	23%
Bonus only ($X)	30	15%	12	48%	18	11%	2	3%
Suffering ($X/$1)	18	9%	5	20%	13	8%	2	3%
Justice ($X/$1 + unfair)	28	14%	1	4%	27	16%	9	12%
Revenge ($X/$1 + unfair + partner responsible)	81	42%	2	8%	79	47%	46	60%
Total	194		25		169		77	

Figure 3: Message choices in Study 1. Here we report the pooled data across the two samples, but the distribution of messages by punishment type are similar across the two subjects pools (see Appendix A1).

By contrast, among participants who decided to reduce the other person's payoff (Figure 3, column 3), the majority, 106 (63%) sent either the justice (16%) or revenge message (47%), thereby making sure that the other person understood the reason why he or she received a reduced bonus. Among participants who enacted the harshest punishment possible (i.e., reduced the other person's payoff to $0, Figure 3, column 4), an even larger proportion, 72% chose these messages. However, it is worth noting that there was a significant minority of participants who decided to keep this severe punishment hidden by not sending any message (23%). The open-ended explanations provided by these participants clarify that these were deliberate decisions, rather than errors: These participants had either a strong preference for fairness but did not want their partner's to know about the reduced bonus (e.g., "I don't think they deserve to receive a bonus. I'm not trying to be malicious so I'd prefer they didn't receive a message along with it") or they thought that their partner's action was so egregious that they *did not even deserve* an explanation (e.g., " I did not choose to send a message because he/she does not deserve an explanation").

takes: These participants wanted to let the other person know that they were unhappy, but they wouldn't want to reduce the other person's payoff (e.g., "I'd want to say something about making me do more work, but I wouldn't reduce the money"). This is consistent with the expressive function of punishment (e.g., Feinberg, 1965).

STUDY 2: INCENTIVIZED CHOICE AND TRADE-OFF BETWEEN MOTIVES

Overview and Hypotheses

In Study 1 we established that in this type of situation (unfair allocation of work), most people would severely punish their partner and also would prefer to send an explanation for why they did so. This finding already highlights that punishers are not indifferent about the transgressor's beliefs. However, Study 1 participants' choices were not constrained, and they were not forced to make a trade-off between different goals (e.g., distributive justice, retributive justice, belief-based motives), so we can't make any inference about the relative value punishers place on affecting the transgressor's beliefs compared to the transgressor's monetary outcomes.

To test whether people about the transgressor's affective reactions and beliefs, and whether they are willing to make trade-offs between these motives, we set up an incentivized experiment, in which participants who believed they had been harmed by a transgressor made choices between different levels of punishment and different communications that conveyed information to the perceived transgressor: Participants could either punish the transgressor severely, moderately, or not at all.

First, in a baseline condition we measured participants' preference for *pure* distributive justice, in the absence of any information about the punishment, i.e., in a case when punishments were completely hidden from the transgressors, and when transgressors were not even aware of the possibility of receiving a reduced payoff. Then, in three other conditions we gradually added more information (e.g., a notification to the transgressor informing them about the reduction; an explanation of why they were being punished) to the punishment option that was the *least preferred* in the baseline condition (moderate punishment), while keeping constant the other two punishment options: no punishment and severe punishment. By doing this, we were able to measure whether adding more information to the previously least preferred punishment option shifted participants' preferences towards what would otherwise be the non-preferred option. The study was designed to test between the following hypotheses:

H1. If participants care only about distributive justice, then we should not observe any difference in punishment behavior across conditions.

H2. If participants care about retributive justice (i.e., how much the transgressor suffers), in addition to distributive justice, then we should observe a shift towards the least preferred (moderate) punishment option in those conditions in which the added information makes the transgressor suffer.

H3. If participants care directly about the beliefs of the transgressor, in addition to distributive and retributive justice, then we should observe a shift towards the least preferred (moderate) punishment option in those conditions in which the added information lets the transgressor understand why he or she received the punishment.

H4. If participants care about whether the transgressor knows who punished them (i.e., the transgressor), then we should observe a shift towards the least preferred (moderate) punishment option in those conditions in which the added information reveals the identity of the punisher.

The experimental design allowed us to investigate both whether adding new information to the dominated (moderate) option "crowds in" non-punishers (i.e., participants selecting moderate punishment instead of no punishment) and whether more information "crowds out" severe punishers (i.e., participants selecting moderate punishment instead of severe punishment).

Methods

Participants. We recruited 1,959 participants on Prolific (https://prolific.ac).[3] Among recruited participants, 153 (7.8%) quit the experiment before being matched with another person (i.e., before being assigned to an experimental condition). Among the 1,806 participants who were matched with another person (i.e., 903 pairs), we excluded 101 (5.6%) participants from our analyses: Two (0.1%) duplicate responses (people who participated more than once), 17 (0.9%) incomplete responses (people who quit before completing the experiment), and 82 participants (4.5%) who failed at least one comprehension check question or who failed the attention check question (see Table 1). We stopped data collection as soon as we obtained 200 observations (i.e., 200 Recipients) in each of the four conditions, after applying the exclusion criteria. The exclusion criteria, as

[3]We used the following eligibility criteria: Participants must be U.S. citizens residing in the U.S., speak English as their first language, and have an approval rating of at least 99%. Participants could complete the experiment only once.

well as the data collection stopping rule were preregistered at AsPredicted.org (LINK). The final sample contained 1,705 responses (50.3% female; age: $M = 34.5$ years). The overall exclusion rates (duplicate + incomplete + failed checks) were not significantly different across conditions, $\chi^2(3, N = 903) = 1.126, p = .771$.

Table 1: Matched, excluded, and included participants across conditions in Study 2.

Role	Condition	Matched	Excluded	Included	Excluded %	χ^2 test
Allocator	ignorance	219	0	219	0.0%	
	suffering	229	0	229	0.0%	NA
	justice	228	2	226	0.9%	(expected $n = 0$)
	revenge	227	1	226	0.4%	
Recipient	ignorance	219	20	199	9.1%	
	suffering	229	28	201	12.2%	$\chi^2(3, N = 903) =$
	justice	228	25	203	11.0%	1.126, $p = .771$
	revenge	227	25	202	11.0%	

Note: Allocators had to pass only 3 comprehension checks, whereas Recipients had to pass 6 comprehension checks and an attention check, leading to higher exclusion rates for Recipients.

Procedure. Eligible participants were redirected to a survey designed in Qualtrics (*Qualtrics*, 2019). After reading the general instructions, participants had to answer three comprehension check questions about the experimental task and their payment. If participants answered all three questions correctly, they proceeded to the matching stage in which they were matched in pairs in real-time using SMARTRIQS (Molnar, 2019).

Within each matched pair of participants, one person was randomly assigned to role A (i.e., the Allocator), while the other participant was assigned to role B (i.e., the Recipient).[4] In the first stage of the experiment, the Allocator made a decision about how to divide work between him- or herself and the Recipient. Participants were told that they would receive a fixed compensation for their work ($1.50 each), regardless of how the work would be split between them. The work consisted of a real-effort "slider" task: Participants had to adjust several sliders to preset (random) values (see Figure 4).

[4]Throughout the study, participants were simply referred to as "A" and "B"—participants were not referred to as "Allocator" or "Recipient" at any point during the study to avoid potential biases associated with these labels.

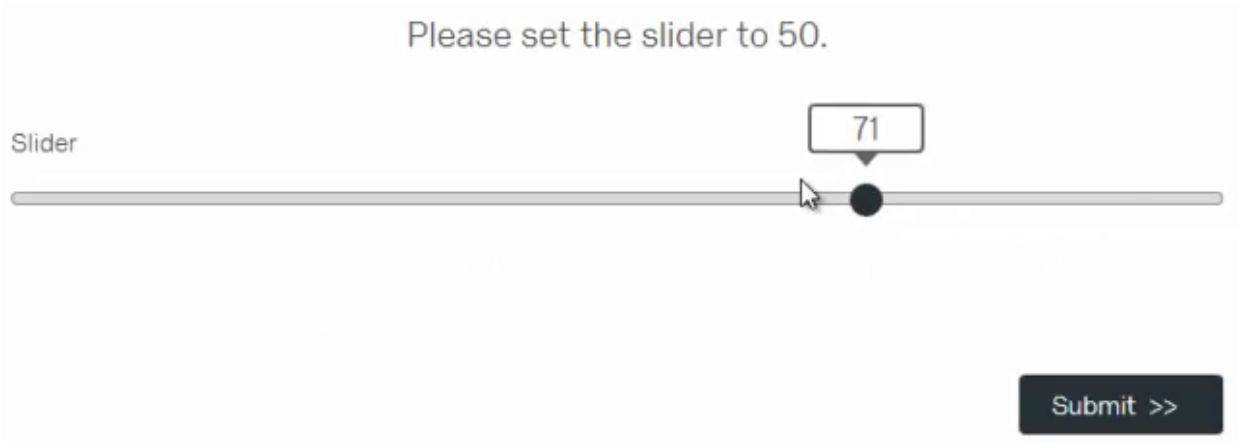

Figure 4: Sample screenshot of the slider task used in the first stage of Study 2. The task consisted of adjusting sliders to preset (random) values. Each pair of participants had to complete 50 of these sliders (combined).

This task is adopted from Gill and Prowse (2012), and is frequently used in economic experiments. The task is designed to capture exerted effort only (it is skill-independent) and to minimize intrinsic motivation. Each pair of participants had to complete 50 of these sliders combined, and the Allocator decided how to split this work within the pair. Unbeknownst to the Recipient, however, the Allocator could not allocate any number of sliders: The Allocator had to choose between two options, both of which allocated a disproportionate amount of work to the Recipient. One option left the Allocator with only 10 sliders, assigning the 40 remaining sliders to the Recipient.

The other option was even more unequal: It assigned 5 sliders to the Allocator and 45 sliders to the Recipient. Since the Recipient did not know about this constrained choice set, the Allocator's choice was perceived as unfair, regardless of which option was chosen (recall that participants received a fixed compensation for their work). This information asymmetry was a necessary element of the experimental design, because it guaranteed that all Recipients would face a very unfair allocation of work. Otherwise, most Allocators would have chosen fair allocations (e.g., 25/25), which would have made the experiment unsuitable for studying punishment behavior.

After the Allocator made a choice, the decision was transmitted to the Recipient. Then, the Allocator completed his or her share of work (either 5 or 10 sliders), after which the survey concluded for the Allocator.

The Recipient, however, was presented with a second (surprise) stage immediately after learning the Allocator's decision, but before starting to work on the slider task. This second stage served as the main part of the experiment. In this stage, the Recipient was told that both he or she and the Allocator would receive a surprise bonus of \$1.00 (each), in addition to the fixed compensation for the slider task. In addition, the Recipient had the opportunity to decrease the Allocator's surprise bonus, without any cost to the Recipient.[5] The Recipient's choice set was constrained, since we wanted to create trade-offs between different punishment motives. The Recipient could choose between three options:

- Do not decrease the other person's bonus, leaving them the full \$1.00. [**no punishment**]
- Decrease the other person's bonus by \$0.50, leaving them \$0.50.[6] [**moderate punishment**]
- Decrease the other person's bonus by \$0.90, leaving them only \$0.10. [**severe punishment**]

If the Recipient chose the no punishment or severe punishment options, the Allocator simply received the \$1.00 (or \$0.10) bonus after the experiment, without any further message or explanation. If the Recipient chose the moderate punishment, the Allocator received a message along with the \$0.50 bonus. These messages were manipulated between subjects across four conditions:

- No message. [*ignorance* condition]
- "Your bonus has been reduced by \$0.50." [*suffering* condition]
- "Your bonus has been reduced by \$0.50, because you were unfair to your partner in the previous task." [*justice* condition]
- "Your bonus has been reduced by \$0.50. Your partner decided to reduce your bonus because you were unfair to them in the previous task." [*revenge* condition]

[5] Unlike in most other experiments studying punishment behavior, punishment was not costly to the punisher in our experiment, since our hypotheses are agnostic about the trade-off between utility from the punisher's *own payoffs* and the other sources of her utility. Instead, we focus on the trade-offs between the punisher's utility from the *transgressor' payoff* and the punisher's utility from the *transgressor' suffering and beliefs*.

[6] We selected this particular level for the moderate punishment option (−\$0.50) based on the results of Study 1, which suggested that most participants (over 80%, see Figure 2) would prefer either the severe punishment (−\$0.90) or the no punishment (−\$0) over a reduction of −\$0.50. This allowed us to introduce a moderate punishment option that was less preferred than the other two options in the baseline (ignorance) condition.

Figure 5 depicts the Recipient's actual decision screen in the revenge condition:

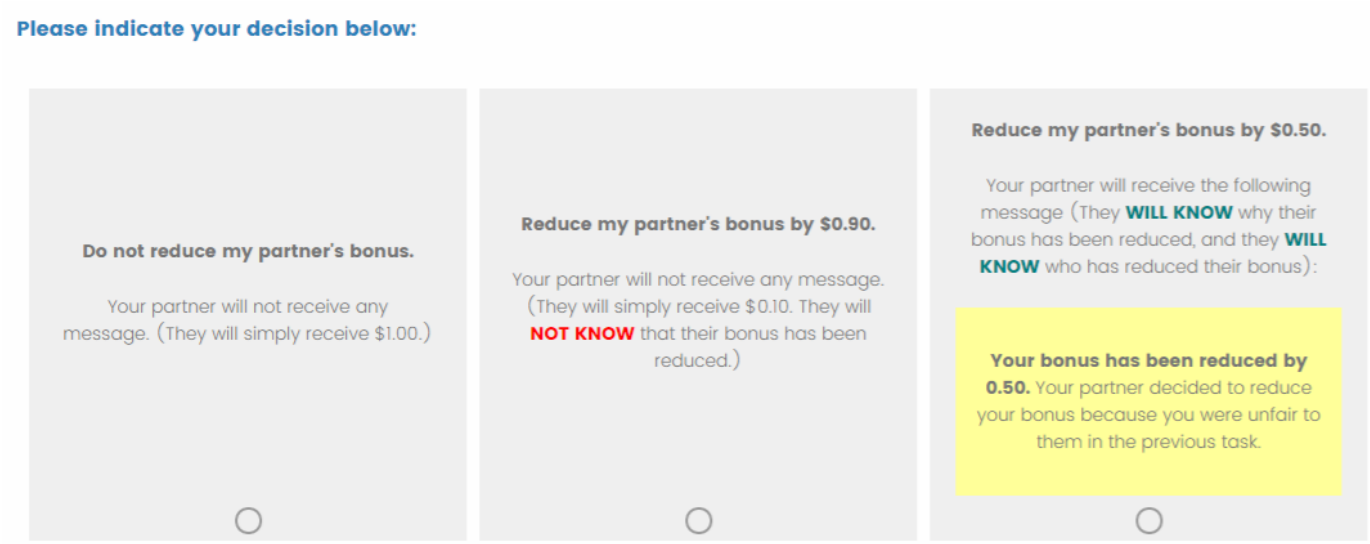

Figure 5: Sample screenshot of the punishment decision in the revenge condition in Study 2. The no punishment (left), moderate punishment (right), and severe punishment (middle) options were presented in a random order. The no and severe punishment options were identical in all four conditions, while the moderate punishment option was manipulated across conditions. Sample screenshots from all four conditions are included in Appendix A2.

Since the experimental manipulation was implemented between subjects, each Recipient made only a single decision.[7] After making a choice, the Recipient was asked to explain their choice (open-ended response), and then answered three comprehension check questions about the consequences of their choice. Then, the Recipient was asked to recall how he or she felt when they first saw the Allocator's decision. The Recipient was presented with the following list of feelings (presented in a random order, except for the last two items): *angry, beloved, betrayed, calm, disappointed, happy, relaxed, sad, satisfied, surprised, other (please specify)*, and *none of the above*. He or she first indicated for each, whether he/she had experienced the emotion, then for each identified as experienced, indicated how strongly they had experienced it (on a continuous scale from *1: a little* to *7: extremely*).

Following this, we elicited the Recipient's beliefs about four things: the morality of each of the three options that each was faced with, the downstream effect of each option on the Allocator's behavior, the effect of the options on the Allocator's feelings, and the effect of the options on the Allocator's sense of guilt.

[7] We decided to use a between-subjects design, instead of asking Recipients to make multiple decisions with different options, to minimize the potential experimental demand effects (Zizzo, 2010; Charness, Gneezy, & Kuhn, 2012).

- "How MORAL or IMMORAL would it be to choose the following options?"
  (continuous scale from *-100: very immoral* to *+100: very moral*)
- "Would YOUR PARTNER feel bad (experience suffering) or feel good (experience joy)?"
  (continuous scale from *-100: very bad* to *+100: very good*)
- "Would YOUR PARTNER FEEL GUILTY or would YOUR PARTNER FEEL PROUD about his or her allocation of work?"
  (continuous scale from *-100: very guilty* to *+100: very proud*)
- "Would YOUR PARTNER TREAT OTHERS WORSE or would YOUR PARTNER TREAT OTHERS BETTER in the future?"
  (continuous scale from *-100: much worse* to *+100: much better*)

To make sure that Recipients were paying sufficient attention to the above questions, we included an attention check question after these items, and excluded everyone who failed to answer the attention check correctly. Then, we recorded the Recipient's age and sex. Finally, the Recipient had to complete the work assigned to him or her: adjusting 40 (or 45) sliders.

At the end of the survey, we asked both Allocators and Recipients to indicate the extent to which they believed that they were interacting with a bot or a human partner (continuous scale from *-100: definitely a bot* to *+100: definitely a human*). We used this question to conduct robustness checks on our main analyses, by excluding, in a set of ancillary analyses, those participants who had a strong (but incorrect) belief that their partner was a bot, as this might have affected their decision.

Results

Allocators' task allocation. Out of the 900 Allocators who were included in the final sample, 810 (90%) chose the 10/40 allocation and 90 (10%) chose the 5/45 allocation.

Main results: Recipient's punishment decision. Out of the 805 Recipients who were included in the final sample, 270 (34%) chose no punishment, 245 (30%) chose moderate punishment, and 290 (36%) chose severe punishment across the four conditions combined.

In the ignorance condition only a minority of Recipients (20%) chose the moderate option, indicating that when only distributive preferences matter, most people prefer the no punishment (37%) or severe punishment (43%) options. Aligned with the hypothesis that people derive utility directly from the transgressor's understanding (i.e., the transgressor's beliefs), a significantly higher proportion of Recipients chose the moderate punishment option in the justice (41%) and revenge (34%) conditions than in the ignorance condition (20%), $\chi^2(1, N = 402) = 20.345$, $p < .001$ and $\chi^2(1, N = 401) = 8.694$, $p = .003$, respectively (see Figure 6). Recipients were also significantly more likely to choose the moderate punishment in the justice condition (41%) than in the suffering condition (26%), $\chi^2(1, N = 404) = 9.496$, $p = .002$.

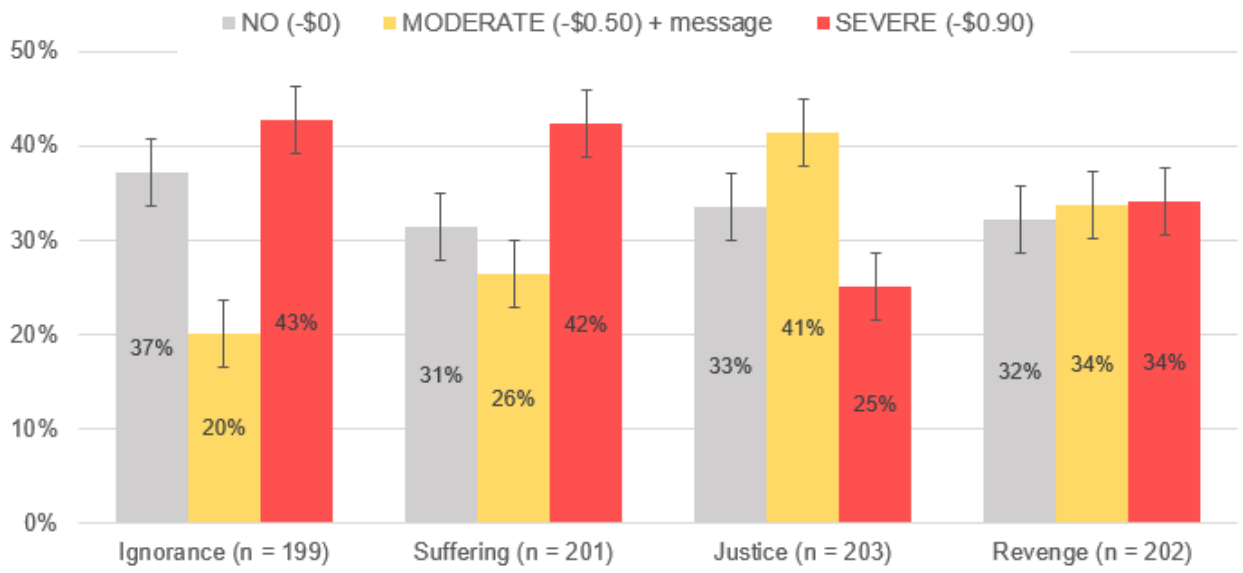

Figure 6: Proportion of Recipients choosing NO (grey), MODERATE (yellow), and SEVERE (red) punishment across conditions in Study 2. Error bars represent ±1 standard error.

Although the suffering condition (26%) was in-between the ignorance (20%) and the revenge conditions (34%), thus the results are directionally consistent with the hypothesis that Recipients derive some utility from making the Allocator suffer, these differences were not significant, $\chi^2(1, N = 400) = 1.864$, $p = .172$, and $\chi^2(1, N = 403) = 2.217$, $p = .137$, between suffering/ignorance and suffering/revenge, respectively. This result highlights the limited role of pure retributive motives (in absence of belief-based motives). We did not observe a significant difference between the justice (41%) and revenge (34%) conditions either, $\chi^2(1, N = 405) = 2.252$, $p = .133$.

If anything, disclosing that the Recipient was responsible for reducing the Allocator's payoff makes the corresponding message less appealing—compared to when the Allocator is informed about the reason of punishment only, i.e., the justice condition. The smaller proportion of Recipients choosing the moderate punishment option in the revenge condition than in the justice condition is consistent with the finding in Study 1 showing that among people who would reduce their partner's payoff *and* send an explanation (i.e., justice and revenge messages, $n = 106$), a significant minority of people ($n = 27$, 25%) would rather prefer *not* to disclose the identity. That is, these people selected the less informative justice message rather than the revenge message, even though their choice was unconstrained.

Crowding-in of non-punishers and crowding-out of severe punishers. Next, we conducted regression analyses to see whether the differences in the proportion of Recipients choosing the moderate punishment across conditions was driven by non-punishers "crowding in" (i.e., more Recipients selecting some punishment) or by the "crowding out" of severe punishers (i.e., fewer Recipients selecting severe punishment).

To investigate this, we created three dummy variables that correspond to the three pieces of information that were gradually added to the messages in the moderate punishment option. In this way, the coefficients on these terms capture the marginal effect of each of these pieces of information on punishment. The first dummy variable ("Suffer") indicates whether the maximum possible bonus was displayed in the message. This dummy is 0 in the ignorance condition and 1 otherwise. The second dummy ("Explain") indicates whether the message clarifies the reason why the Allocator's payoff has been reduced, and is 0 in the ignorance and suffering conditions and 1 otherwise. Finally, the third dummy ("Identity") indicates whether the Allocator is informed that the Recipient decided to reduce their payoff, and is 1 in the revenge condition and 0 otherwise.

We then conducted three separate OLS linear regressions, adding the above dummy variables as independent variables and the likelihood of choosing no, moderate, and severe punishment as dependent variables (see Table 2, Columns 1, 3, and 5).

Most importantly, these analyses replicate the results reported in the previous section: Participants were significantly more likely to choose the moderate punishment when the accompanying message clarified the reason for reducing the Allocator's payoff (i.e., Explain = 1, in the justice

and revenge conditions), $\beta = 0.150$, $t(801) = 3.320$, $p < .001$. Furthermore, participants chose the moderate punishment marginally significantly less likely when this option also revealed their identity to the Allocator (i.e., Identity = 1, in the revenge condition), $\beta = -0.077$, $t(801) = 1.709$, $p = .088$. Adding the maximum possible bonus to the messages, thus merely making the Allocator suffer (Suffer = 1), did not lead Recipients to select the moderate punishment more often, $\beta = 0.063$, $t(801) = 1.380$, $p = .168$, which highlights the limited role of pure retributive motives.

Table 2: Results of OLS regression analyses in Study 2: Likelihood of the Recipient choosing no, moderate, and severe punishment options.

	Dependent variable:					
	Likelihood of choosing NO punishment		Likelihood of choosing MODERATE punishment		Likelihood of choosing SEVERE punishment	
	(1)	(2)	(3)	(4)	(5)	(6)
Suffer[1]	−0.058	−0.057	0.063	0.058	−0.004	−0.0005
	(0.047)	(0.047)	(0.045)	(0.045)	(0.048)	(0.047)
Explain[2]	0.022	0.027	0.150***	0.151***	−0.172***	−0.178***
	(0.047)	(0.047)	(0.045)	(0.045)	(0.047)	(0.047)
Identity[3]	−0.013	−0.016	−0.077*	−0.077*	0.090*	0.092**
	(0.047)	(0.047)	(0.045)	(0.045)	(0.047)	(0.047)
Sex ($F = 1$)		0.072**		0.035		−0.107***
		(0.033)		(0.032)		(0.034)
Age (years)		0.003**		−0.002*		−0.001
		(0.001)		(0.001)		(0.001)
Constant	0.372***	0.222***	0.201***	0.271***	0.427***	0.508***
	(0.034)	(0.059)	(0.032)	(0.057)	(0.034)	(0.059)
Observations	805	805	805	805	805	805
R^2	0.002	0.017	0.030	0.035	0.022	0.036
Adjusted R^2	−0.002	0.011	0.026	0.029	0.019	0.030

Note: $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

[1] *Dummy: ignorance = 0; suffering = 1; justice = 1; revenge = 1*

[2] *Dummy: ignorance = 0; suffering = 0; justice = 1; revenge = 1*

[3] *Dummy: ignorance = 0; suffering = 0; justice = 0; revenge = 1*

None of the experimental manipulations had a significant effect on the likelihood of choosing no punishment, all $p > .217$, which suggests that having the ability to convey more information did not cause a "crowding in" of non-punishers: The proportion of participants who decided to let the Allocator's unfairness go unpunished remained relatively constant across conditions.

By contrast, participants were significantly less likely to choose the severe punishment when they could explain the punishment by choosing the moderate punishment (i.e., Explain = 1), $\beta = -0.172$, $t(801) = 3.625$, $p < .001$. In the revenge condition (Identity = 1), however, they were marginally significantly more likely to choose the severe punishment, $\beta = 0.090$, $t(801) = 1.911$, $p = .056$. These effects are directionally opposite, and of similar magnitude, to the effects observed for the likelihood of moderate punishment, suggesting that most of the main effects can be explained by the "crowding out" of severe punishers: Once Recipients were able to explain the punishment decision, they enacted a less severe punishment, even though in the absence of an explanation, they would have preferred a severe punishment. This indicates that at least some participants were willing to compromise on distributive justice, to make sure that their partner understood the reason for being punished.

Robustness checks: demographics, anger, and suspicion. In a second set of OLS regressions (see Table 2, Columns 2, 4, and 6) we also included sex and age as demographic controls. Although both sex and age have significant main effects on punishment choices—women are more likely to choose no punishment, while men are more likely to choose severe punishment; older participants are more likely to choose no punishment—the effects of the experimental manipulation are robust to the inclusion of these demographic controls.

Furthermore, all of the above results are robust if we limit our analyses to those Recipients only who reported that they were angry after seeing the Allocator's decision. Results are also robust to the inclusion or exclusion of suspicious Recipients (i.e., who did not believe that the Allocator was a human). We report the detailed results of these robustness checks in Appendix A3.

Alternative Explanation 1: The "Explanation Enhances Suffering" Hypothesis

Even though the main results strongly support the hypothesis that punishers care about transgressors' beliefs per se, and derive utility from being able to let the transgressors know why they are being punished, the observed pattern of results can be also consistent with an alternative hypothesis. If the transgressor *suffers more* when the punishment is accompanied by an explanation, then the higher willingness to enact a punishment with an explanation is also consistent with retributive motives, i.e., the desire to make the transgressor suffer for their misbehavior. This would

imply that punishers who prefer the moderate punishment in the justice and revenge conditions do not necessarily care about what the transgressor believes per se, but they do want to make these transgressors suffer more by clarifying why they received the punishment.

To test whether this alternative hypothesis can explain the main results, we compared Recipients' responses to the post-punishment question "[If you chose this option: ...] Would YOUR PARTNER feel bad (experience suffering) or feel good (experience joy)?" Recall that we asked this question about each of the available options (no, moderate, severe punishment), instead of asking the question only about the option that participants selected, therefore we obtained measures on all three options from everyone (avoiding a self-selection bias).

We found no significant differences in the ratings of the no punishment option between any two options (all $p > .215$, see Figure 7, gray bars). Recipients reported that the Allocator would feel very good in all four conditions if they decided to let them receive the full \$1 bonus (without letting them know anything else about the bonus): $M = 85.4$, 95% CI [81.7, 89.1], $M = 88.0$, 95% CI [84.5, 91.5], $M = 85.4$, 95% CI [81.3, 89.4], and $M = 84.8$, 95% CI [81.1, 88.5] in the ignorance, suffering, justice, and revenge conditions, respectively.

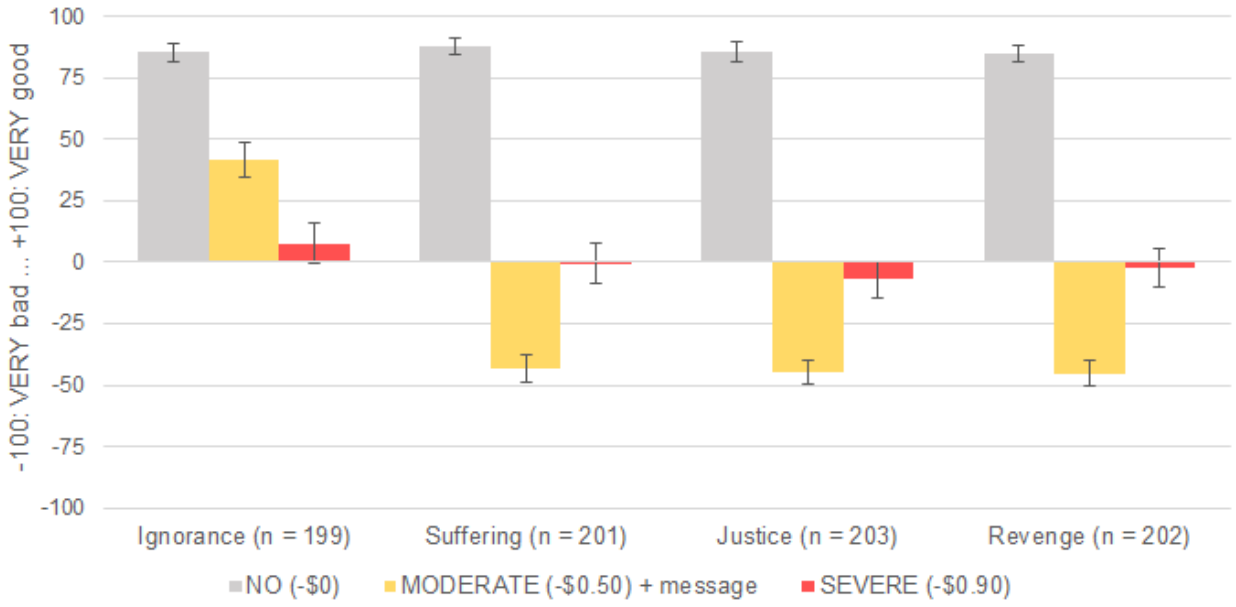

Figure 7: Means of Recipients' responses to the question: "Would YOUR PARTNER feel bad (experience suffering) or feel good (experience joy)?" in Study 2. Error bars represent 95% Confidence Intervals.

Similarly, we found no significant differences in the ratings of the severe punishment between any two conditions (all $p > .074$), except for between the ignorance and justice conditions, where we found a significant but weak effect, $t(400) = 2.633$, $p = .009$, $d = 0.26$ (Figure 7, red bars). More importantly, however, Recipients reported that the Allocator would feel neutral (i.e., not significantly different from 0) in all four conditions if they received a reduced bonus of \$0.10 (without knowing that the bonus was reduced): $M = 7.66$, 95% CI [-0.34, 15.7], $M = -0.58$, 95% CI [-8.61, 7.46], $M = -7.14$, 95% CI [-14.7, 0.45], $M = -2.59$, 95% CI [-10.4, 5.27] in the ignorance, suffering, justice, and revenge conditions, respectively. This indicates that even though this option reduced the Allocator's material welfare the most, Recipients (correctly) anticipated that their partner would not suffer because they would remain ignorant.

By contrast, Recipients reported that the Allocator would feel rather good when receiving the moderately reduced bonus in the ignorance condition ($M = 49.4$, 95% CI [34.8, 48.5]), and significantly better than when receiving the same bonus in any of the other three conditions (all $p < .001$), in all of which they would feel rather bad: $M = -43.2$, 95% CI [-48.6 -37.8], $M = -44.5$, 95% CI [-49.3, -39.7], $M = -45.2$, 95% CI [-50.2 -40.2] in the suffering, justice, and revenge conditions, respectively (see Figure 7, yellow bars).

The "explanation enhances suffering" hypothesis implies that Recipients would expect that the Allocator suffers more in the justice and revenge conditions than in the suffering condition. However, we found no such difference in the reported suffering between the suffering, justice, and revenge conditions (all $p > .595$), which indicates that letting the Allocator know the reason why their payoff was reduced was not expected to make them to suffer more.

To provide further support against the "explanation enhances suffering" hypothesis, we added Recipients' beliefs about the Allocator's suffering upon receiving the moderately reduced bonus to the regression analyses we conducted in the previous section. As Table 3 shows, not only does the effect of the Explain dummy remain significant after adding this new independent variable to the model, but the Recipients' belief about the Allocators' suffering is not a significant predictor of the Recipients' likelihood of choosing the moderate punishment: In other words, participants who believed that their partner would suffer more when enacting the moderate punishment, were not more likely to choose this option, compared to people who thought that their partner would suffer less.

Table 3: Regression results: Likelihood of the Recipient choosing MODERATE punishment in Study 2, controlling for the Recipient's belief about the Allocator's suffering upon receiving the moderately reduced bonus (moderate punishment).

	Dependent variable:		
	Likelihood of choosing MODERATE punishment		
	(1)	(2)	(3)
Suffer (dummy: ign. = 0; suff., just., rev. = 1)	0.063	0.056	0.063
	(0.045)	(0.057)	(0.057)
Explain (dummy: ign., suff. = 0; just., rev. = 1)	0.150***	0.150***	0.151***
	(0.045)	(0.045)	(0.045)
Identity (dummy: ign., suff., just. = 0; rev. = 1)	−0.077*	−0.077*	−0.077*
	(0.045)	(0.045)	(0.045)
Recipient's belief about Allocator's suffering		−0.0001	0.0001
		(0.0004)	(0.0004)
Sex (F = 1)			0.036
			(0.032)
Age (years)			−0.002*
			(0.001)
Constant	0.201***	0.204***	0.269***
	(0.032)	(0.036)	(0.058)
Observations	805	805	805
R^2	0.030	0.030	0.035
Adjusted R^2	0.026	0.025	0.028
Note:	$^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$		

Thus, the above results strongly support that the Recipients' higher willingness to choose the moderate punishment in the justice and revenge conditions cannot be explained their motive for retributive justice, i.e., by their desire to inflict more suffering. This very limited role of pure retributive motives behind punishment decisions is consistent with the finding in Study 2 that introducing the suffering component per se (difference between the ignorance and suffering conditions) did not significantly increase the proportion of participants who chose the moderate option. It is also consistent with the finding in Study 1 showing that among people who would punish their partner ($n = 169$), only a negligible fraction ($n = 13$, 8%) would make their partner suffer without sending them any explanation.

Alternative Explanation 2: Teaching a Lesson (Deterrence Motives)

The other potential alternative explanation is that Recipients had deterrence motives: They wanted to teach the Allocator a lesson. While this interaction was one-shot and anonymous—i.e., sending an explanation could not confer any instrumental value to the Recipient—Recipients might still want to send an explanation to the Allocator if they care about how the Allocator behaves towards others in the future and if they believe that sending a message would change the Allocator's future behavior (altruistic punishment).

To investigate whether such considerations could explain the observed results, we compared Recipients' responses to the post-punishment question "[If you chose this option: ...] Would YOUR PARTNER TREAT OTHERS WORSE or would YOUR PARTNER TREAT OTHERS BETTER in the future?" Recall that we asked this question about each of the available options (no, moderate, severe punishment), therefore we obtained measures on all three options from everyone (avoiding a self-selection bias). While Recipients thought that enacting no or severe punishment would not change their partner's future behavior significantly (or make it even worse), they thought that the moderate option would make their partner behave better in the future in all but the ignorance condition (see Figure 8).

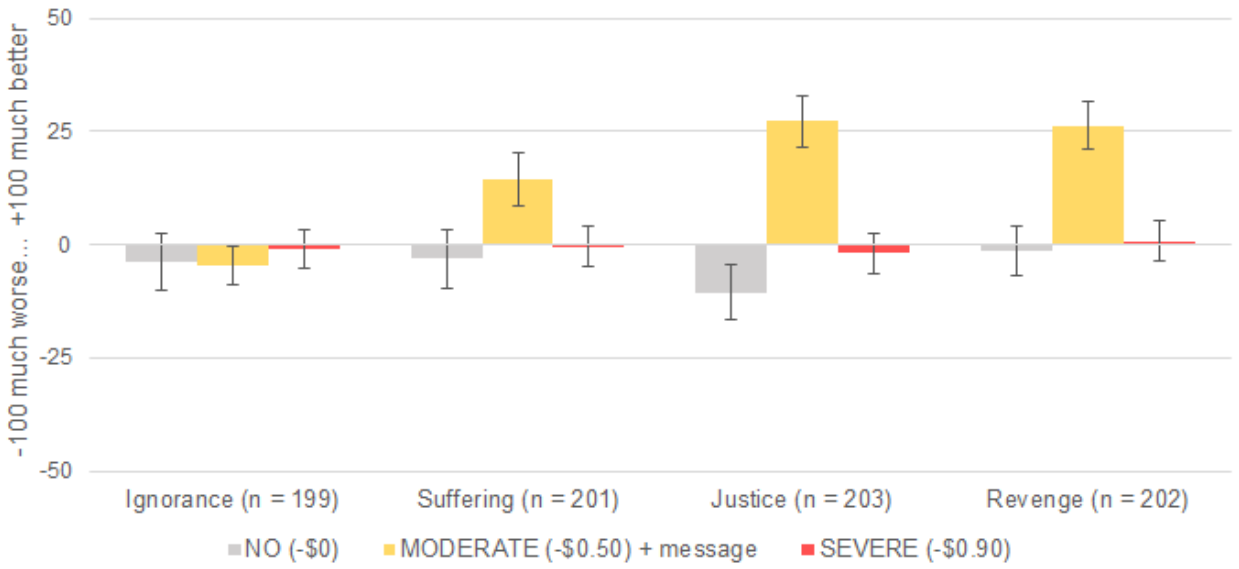

Figure 8: Means of Recipients' responses to the question: "Would YOUR PARTNER TREAT OTHERS WORSE or would YOUR PARTNER TREAT OTHERS BETTER in the future?" in Study 2. Error bars represent 95% Confidence Intervals.

This already suggests that *some* of the increased willingness to choose the moderate punishment in the justice and revenge conditions might be explained by deterrence motives, i.e. to make the other person behaves better in the future. To investigate whether this motive was solely responsible for the observed differences between conditions, we added this independent variable to the regression conducted in the previous section. As Table 4 shows, Recipients who thought that the moderate punishment would make their partner behave better in the future, were more likely to choose the moderate punishment, supporting the claim that participants were at least partially motivated by deterrence, $\beta = 0.002$, $t(800) = 4.876$, $p < .001$. Moreover, including this variable weakened the coefficient of the Explain dummy (cf. Columns 1 and 2 in Table 4). More importantly, however, the Explain dummy is still significant, and the reduction of its coefficient is only 2.6 percentage points (or about 17% of the original effect), $\beta = 0.124$, $t(800) = 2.764$, $p = .006$.

Table 4: Regression results: Likelihood of the Recipient choosing the MODERATE punishment option in Study 2, controlling for the Recipient's motive for improving the Allocator's future behavior.

	Dependent variable:		
	Likelihood of choosing MODERATE punishment		
	(1)	(2)	(3)
Suffer (dummy: ign. = 0; suff., just., rev. = 1)	0.063	0.024	0.020
	(0.045)	(0.045)	(0.045)
Explain (dummy: ign., suff. = 0; just., rev. = 1)	0.150***	0.124***	0.125***
	(0.045)	(0.045)	(0.045)
Identity (dummy: ign., suff., just. = 0; rev. = 1)	−0.077*	−0.075*	−0.075*
	(0.045)	(0.045)	(0.044)
Future behavior of the Allocator		0.002***	0.002***
		(0.0004)	(0.0004)
Sex (F = 1)			0.032
			(0.032)
Age (years)			−0.002*
			(0.001)
Constant	0.201***	0.210***	0.271***
	(0.032)	(0.032)	(0.056)
Observations	805	805	805
R^2	0.030	0.058	0.062
Adjusted R^2	0.026	0.053	0.055
Note:			*$p < .1$; **$p < .05$; ***$p < .01$

Thus, most of the effect of the explanation cannot be explained by participants' desire to teach a lesson: Participants have a preference for enacting a punishment with an explanation, beyond merely wanting to improve their partner's future behavior by doing so.

SUMMARY

We show that punishers take into account what a transgressor thinks after being punished, and these motives affect the punisher's decision after taking account of purely distributional preferences (i.e., the transgressor's material outcome), the punisher's suffering, and the potential deterrence effects of the punishment. Our results strongly support the idea that punishment decisions are at least partially motivated by *belief-based preferences*—the punisher's preferences over what the transgressor believes—and that these preferences can often dominate distributional preferences. Specifically, punishers have a strong desire for transgressors to know that they have been punished, and for what.

To our knowledge, this experiment is the first that clearly disentangles these three potential motives behind punishment—material, affective, and cognitive. The most closely related study that tried to disentangle distributive preferences from other considerations is by Crockett, Özdemir, and Fehr (2014). They conducted a 3-player one-shot anonymous Trust Game in which a third-party observer could punish an untrustworthy participant (who was given funds but did not return them). The authors manipulated whether the punishment was *open* or *hidden*—i.e., whether the transgressor would realize that they had been punished by their payoff being reduced by the third party—and found that people were significantly more willing to punish in the open condition. The authors conclude that the preference to communicate norms through punishment plays an important role for punishment decisions. However, their experimental design did not force participants to make any trade-off with respect to the various motives behind punishment, and it did not allow for disentangling the comparative suffering hypothesis from the understanding hypothesis: The higher likelihood of punishing transgressors in the open condition is consistent with both a preference for retributive justice and belief-based preferences.

By contrast, in our experiment we isolated retributive motives from belief-based preferences, and demonstrated that most of the difference between open and hidden punishment is driven by

belief-based preferences, and not by retributive motives. Furthermore, we also show that once we allow participants to choose punishment options that provide an explanation to the transgressor (and thus, provide additional belief-based utility for the punisher), we observe a crowding-out effect of severe punishers, but no crowding-in effect of people who would otherwise not punish.

Thus, the availability of explanations can increase overall social welfare by reducing the severity of material punishments. For instance, this insight has implications for the escalation of punitive damages in the legal system, which which legal scholars have argued are getting out of hand (e.g., Huber, 1989, Sunstein, Kahneman, & Schkade, 1998). If victims are provided alternative ways of communicating messages to transgressors, through, for instance, mediation or victim impact statements, they may be willing to settle legal suits for lower amounts or to ask for lower punitive damages. This is consistent with evidence showing that providing legal cover for transgressors and victims to communicate to each other (i.e., through "apology laws") is associated with faster settlements and lower settlement amounts (Ho & Liu, 2011).

This work expands our understanding of belief-based preferences to include preferences over second-order beliefs, i.e., the beliefs of others, that go beyond image motivation and strategic considerations (e.g., Battigalli et al., 2019). Most previous work on belief-based preferences has examined preferences for cognitive states about material outcomes affecting the self, such as whether one has an untreatable illness (Oster, Shoulson, & Dorsey, 2013; Ganguly & Tasoff, 2017) or whether one's stock portfolio has decreased in value (Karlsson, Loewenstein, & Seppi, 2009). Research on image motivation (e.g., Ariely, Bracha, & Meier, 2009; Soetevent, 2011) has found that people also have a preference for certain cognitive states about others' cognitive states, but specifically with respect to image. That is, people care what others think about them.

People's concern about what goes on in other people's minds—their thoughts, beliefs, preferences and feelings, however, goes well beyond the desire for others to hold a positive image of them, and has diverse consequences for interpersonal relations, politics and economics. For example, research in psychology suggests that people not only have a desire to be perceived positively, but also realistically (as perceived by themselves), even if this implies a negative image (see self-verification theory: Swann, Pelham, & Krull, 1989; Swann, 2011).

These motivations undoubtedly play an important role in the growth of social networking sites such as Facebook, Instagram and Snapchat, and also contribute to the proliferation of websites

devoted to broadcasting people's anonymous confessions, typically of their less-than-admirable activities. Confessions are especially difficult to reconcile with standard economic accounts of behavior since they typically consist of information, revelation of which would be antithetical to the revealer's interests. Likewise, people care deeply about what other people believe, and about how other people's beliefs align with their own (c.f., Golman et al., 2016), which has important consequences for geographic mobility (Bishop, 2009; Motyl, Iyer, Oishi, Trawalter, & Nosek, 2014) and politics (Iyengar & Westwood, 2015; McCoy, Rahman, & Somer, 2018).

Our paper finds evidence for an additional interest in others' beliefs that seems to be unrelated to what others perceive about them: People care what transgressors think about punishment they receive—that they understand they have been punished, and why, though, according to our admittedly preliminary research, not necessarily by who. We have demonstrated that punishers want transgressors to have a particular understanding of their outcomes, and are willing to compromise on distributive and retributive justice to fulfill this goal.

REFERENCES

Ariely, D., Bracha, A., & Meier, S. (2009). Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. *American Economic Review*, *99*(1), 544–555. doi: 10.1257/aer.99.1.544

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*(5), 325–344. doi: 10.1016/j.evolhumbehav.2006.01.003

Battigalli, P., Corrao, R., & Dufwenberg, M. (2019). Incorporating belief-dependent motivation in games. *Journal of Economic Behavior & Organization*, *167*(February), 185–218. doi: 10.1016/j.jebo.2019.04.009

Bénabou, R., & Tirole, J. (2016). Mindful Economics: The Production, Consumption, and Value of Beliefs. *Journal of Economic Perspectives*, *30*(3), 141–164. doi: 10.1257/jep.30.3.141

Bishop, B. (2009). *The big sort: Why the clustering of like-minded America is tearing us apart.* Houghton Mifflin Harcourt.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review*, *90*(1), 166–193. doi: 10.1257/aer.90.1.166

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated Punishment of Defectors Sustains Cooperation and Can Proliferate When Rare. *Science*, *328*(5978), 617–620. doi: 10.1126/science.1183665

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. doi: 10.1037/0022-3514.83.2.284

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1–8. doi:

10.1016/j.jebo.2011.08.009

Charness, G., & Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, *117*(3), 817–869. doi: 10.1162/003355302760193904

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

Dickinson, D. L., & Masclet, D. (2015). Emotion venting and punishment in public good experiments. *Journal of Public Economics*, *122*, 55–67. doi: 10.1016/j.jpubeco.2014.10.008

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

Fehr, E., & Gächter, S. (2000). Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives*, *14*(3), 159–182. doi: 10.1257/jep.14.3.159

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

Feinberg, J. (1965). The Expressive Function of Punishment. *Monist*, *49*(3), 397–423. doi: 10.5840/monist196549326

French, P. (2001). *The virtues of vengeance*. Kansas: The University Press of Kansas.

Funk, F., McGeer, V., & Gollwitzer, M. (2014). Get the Message. *Personality and Social Psychology Bulletin*, *40*(8), 986–997. doi: 10.1177/0146167214533130

Ganguly, A., & Tasoff, J. (2017). Fantasy and Dread: The Demand for Information and the Consumption Utility of the Future. *Management Science*, *63*(12), 4037–4060. doi: 10.1287/mnsc.2016.2550

Gill, D., & Prowse, V. (2012). A Structural Analysis of Disappointment Aversion in a Real Effort Competition. *American Economic Review*, *102*(1), 469–503. doi: 10.1257/aer.102.1.469

Gollwitzer, M., & Bushman, B. J. (2012). Do Victims of Injustice Punish to Improve Their Mood? *Social Psychological and Personality Science*, *3*(5), 572–580. doi: 10.1177/1948550611430552

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

Golman, R., Loewenstein, G., Moene, K. O., & Zarri, L. (2016). The Preference for Belief Consonance. *Journal of Economic Perspectives*, *30*(3), 165–188. doi: 10.1257/jep.30.3.165

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., ... Ziker, J. (2006). Costly Punishment Across Human Societies. *Science*, *312*(5781), 1767–1770. doi: 10.1126/science.1127333

Ho, B., & Liu, E. (2011). Does sorry work? The impact of apology laws on medical malpractice. *Journal of Risk and Uncertainty*, *43*(2), 141–167. doi: 10.1007/s11166-011-9126-0

Huber, P. (1989). No-Fault Punishment. *Alabama Law Review*, *40*(3), 1037–1052.

Iyengar, S., & Westwood, S. J. (2015). Fear and Loathing across Party Lines: New Evidence on Group Polarization. *American Journal of Political Science*, *59*(3), 690–707. doi: 10.1111/ajps.12152

Johnson, T., Dawes, C. T., Fowler, J. H., McElreath, R., & Smirnov, O. (2009). The role of egalitarian motives in altruistic punishment. *Economics Letters*, *102*(3), 192–194. doi: 10.1016/j.econlet.2009.01.003

Jordan, J., McAuliffe, K., & Rand, D. (2015). The effects of endowment size and strategy method on third party punishment. *Experimental Economics*, *19*(4), 741–763. doi: 10.1007/s10683-015-9466-8

Kant, I. (1952). The science of right (W. Hastie, Trans.). In R. Hutchins (Ed.), *Great books ofthe western world: Vol. 42. kant* (p. 397–446). Chicago: Encyclopedia Brittanica.

Karlsson, N., Loewenstein, G., & Seppi, D. (2009). The ostrich effect: Selective attention to information. *Journal of Risk and Uncertainty*, *38*(2), 95–115. doi: 10.1007/s11166-009-9060-6

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*(2), 75–84. doi: 10.1016/j.evolhumbehav.2006.06.001

Loewenstein, G., & Molnar, A. (2018). The renaissance of belief-based utility in economics. *Nature Human Behaviour*, *2*(3), 166–167. doi: 10.1038/s41562-018-0301-z

Masclet, D., Noussair, C., Tucker, S., & Villeval, M.-C. (2003). Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism. *American Economic Review*, *93*(1), 366–380. doi: 10.1257/000282803321455359

McCoy, J., Rahman, T., & Somer, M. (2018). Polarization and the Global Crisis of Democracy: Common Patterns, Dynamics, and Pernicious Consequences for Democratic Polities. *American Behavioral Scientist*, *62*(1), 16–42. doi: 10.1177/0002764218759576

Molnar, A. (2019). SMARTRIQS: A Simple Method Allowing Real-Time Respondent Interaction in Qualtrics Surveys. *Journal of Behavioral and Experimental Finance*, *22*, 161–169. doi: 10.1016/j.jbef.2019.03.005

Motyl, M., Iyer, R., Oishi, S., Trawalter, S., & Nosek, B. A. (2014). How ideological migration geographically segregates groups. *Journal of Experimental Social Psychology*, *51*, 1–14. doi: 10.1016/j.jesp.2013.10.010

Oster, E., Shoulson, I., & Dorsey, E. R. (2013). Optimal Expectations and Limited Medical Testing: Evidence from Huntington Disease. *American Economic Review*, *103*(2), 804–830. doi: 10.1257/aer.103.2.804

Qualtrics. (2019). Provo, Utah, USA: Qualtrics. Retrieved from https://www.qualtrics.com

Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. *Trends in Ecology & Evolution*, *30*(2), 98–103. doi: 10.1016/j.tree.2014.12.003

Raihani, N. J., & McAuliffe, K. (2012). Human punishment is motivated by inequity aversion, not a desire for reciprocity. *Biology Letters*, *8*(5), 802–804. doi: 10.1098/rsbl.2012.0470

Santos, M. d., Rankin, D. J., & Wedekind, C. (2011). The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1704), 371–377. doi: 10.1098/rspb.2010.1275

Soetevent, A. R. (2011). Payment Choice, Image Motivation and Contributions to Charity: Evidence from a Field Experiment. *American Economic Journal: Economic Policy*, *3*(1), 180–205. doi: 10.1257/pol.3.1.180

Sunstein, C. R., Kahneman, D., & Schkade, D. (1998). Assessing Punitive Damages (With Notes on Cognition and Valuation in Law). *The Yale Law Journal*, *107*(7), 2071. doi: 10.2307/797417

Swann, W. B. (2011). Self-verification theory. In *Handbook of theories of social psychology (vol. 2)* (pp. 23–42).

Swann, W. B., Pelham, B. W., & Krull, D. S. (1989). Agreeable fancy or disagreeable truth? Reconciling self-enhancement and self-verification. *Journal of Personality and Social Psychology*, *57*(5), 782–791. doi: 10.1037/0022-3514.57.5.782

Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences*, *102*(20), 7398–7401. doi: 10.1073/pnas.0502399102

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, *13*(1), 75–98. doi: 10.1007/s10683-009-9230-z

APPENDIX

Appendix A1. Robustness Check: Study 1

Number (and proportion) of participants choosing each message, MTurk only

Message	Participants							
	Everyone (any X)		Did not punish (X = $1)		Punished (X < $1)		Fully punished (X = $0)	
Ignorance (no message)	19	20%	2	22%	17	20%	8	20%
Bonus only ($X)	12	13%	3	33%	9	10%	1	2%
Suffering ($X/$1)	8	8%	2	22%	6	7%	2	5%
Justice ($X/$1 + unfair)	15	16%	1	11%	14	16%	2	5%
Revenge ($X/$1 + unfair + partner responsible)	42	44%	1	11%	41	47%	28	68%
Total	96		9		87		41	

Figure 9: Message choices in Study 1, MTurk sample only.

Number (and proportion) of participants choosing each message, Prolific only

Message	Participants							
	Everyone (any X)		Did not punish (X = $1)		Punished (X < $1)		Fully punished (X = $0)	
Ignorance (no message)	18	18%	3	19%	15	18%	10	28%
Bonus only ($X)	18	18%	9	56%	9	11%	1	3%
Suffering ($X/$1)	10	10%	3	19%	7	9%	0	0%
Justice ($X/$1 + unfair)	13	13%	0	0%	13	16%	7	19%
Revenge ($X/$1 + unfair + partner responsible)	39	40%	1	6%	38	46%	18	50%
Total	98		16		82		36	

Figure 10: Message choices in Study 1, Prolific sample only.

Appendix A2. Punishment Decision Screens in Study 2 (All Conditions)

Ignorance condition

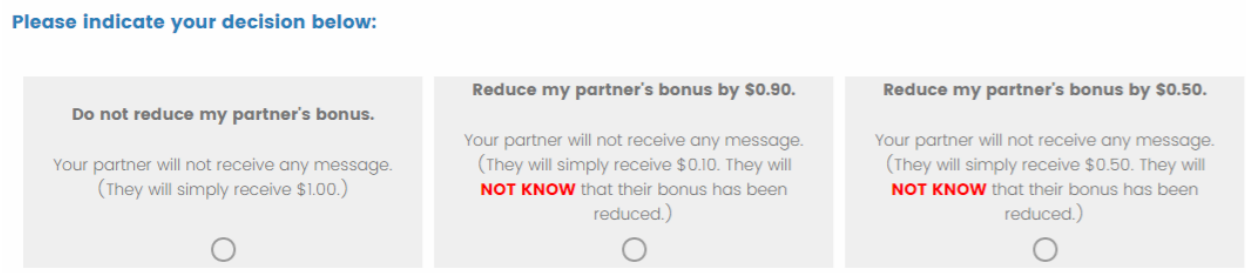

Figure 11: Sample screenshot of the punishment decision in the ignorance condition. The no punishment (left), moderate punishment (right), and severe punishment (middle) options were presented in a random order.

Suffering condition

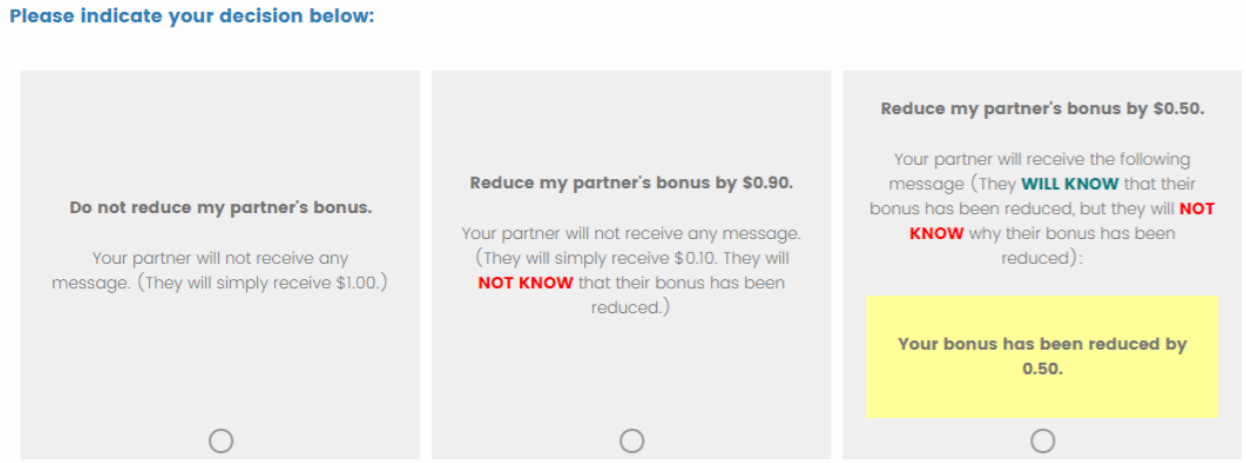

Figure 12: Sample screenshot of the punishment decision in the suffering condition. The no punishment (left), moderate punishment (right), and severe punishment (middle) options were presented in a random order.

Justice condition

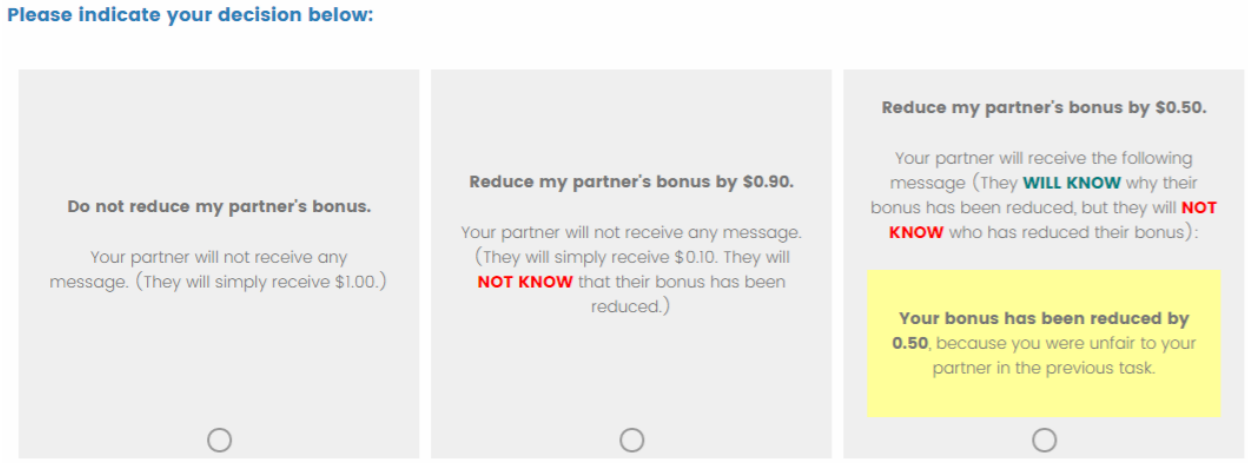

Figure 13: Sample screenshot of the punishment decision in the justice condition. The no punishment (left), moderate punishment (right), and severe punishment (middle) options were presented in a random order.

Revenge condition

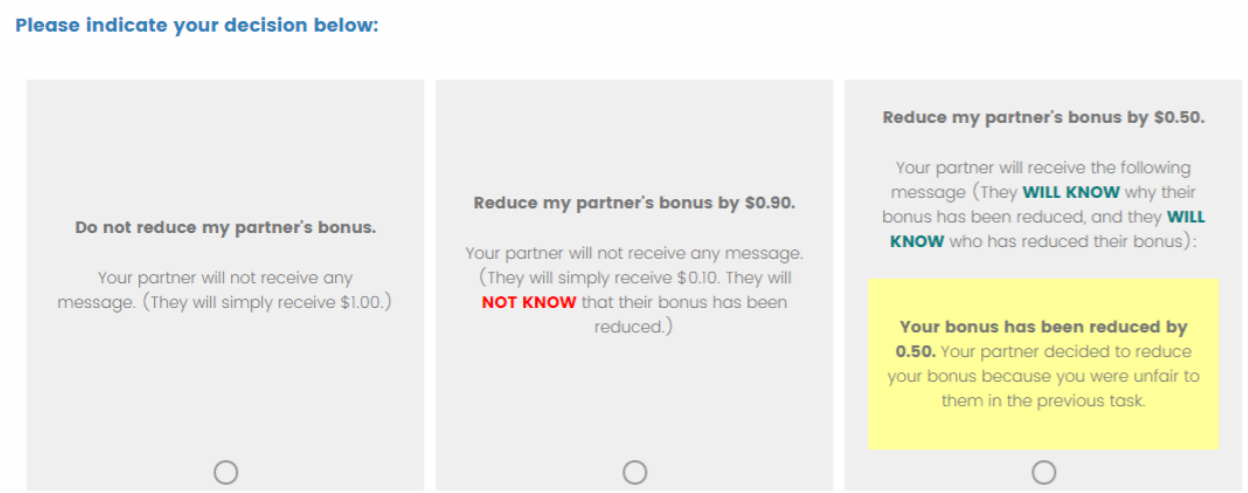

Figure 14: Sample screenshot of the punishment decision in the revenge condition. The no punishment (left), moderate punishment (right), and severe punishment (middle) options were presented in a random order.

Appendix A3. Robustness Checks — Study 2

Anger. In this robustness check we looked at the role of anger. It can be argued that when people are in a "hot" state, they have stronger retributive motives, and think less about the consequences of their actions, thus might have weaker belief-based preferences, and would prefer to inflict greater pain to the transgressor. If this is the case, we should expect a shift towards the moderate punishment option when we introduce the suffering component (ignorance → suffering condition), but not when we introduce the explanation component (suffering → justice condition).

To test this,we analyzed the choices of Recipients who reported that they were angry when they saw the Allocator's decision, i.e., who selected *angry* from the list of 12 feelings ($n = 242$). While these participants were more likely to choose the severe punishment overall, we still observe a significant shift towards the moderate punishment option (crowding out), once this option allowed for sending an explanation (see Figure 15). If anything, the main effect between conditions is even more pronounced for angry Recipients than for the full sample.

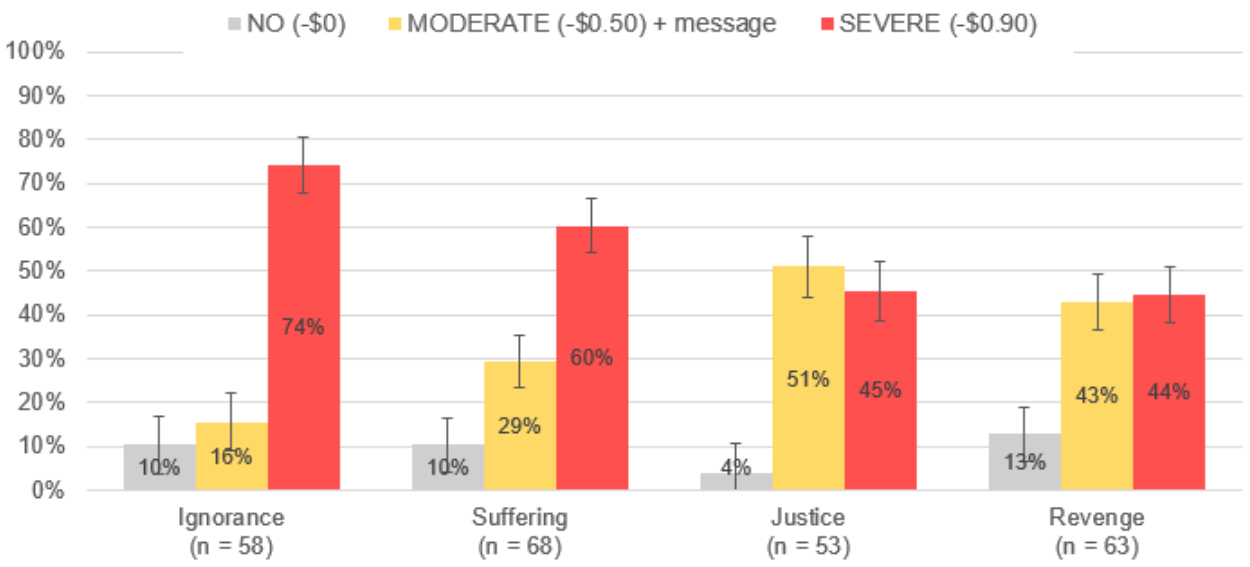

Figure 15: Proportion of Recipients choosing NO (grey), MODERATE (yellow), and SEVERE (red) punishment across conditions, among those Recipients who were angry. Error bars represent ±1 standard error.

All of the main results are robust: A significantly higher proportion of Recipients chose the moderate punishment option in the justice condition (51%) and in the revenge condition (43%) than in the ignorance condition (16%), $\chi^2(1, N = 111) = 14.285$, $p < .001$, and $\chi^2(1, N = 121) = 9.532$, $p = .002$, respectively. The proportion of moderate-choosers was also significantly higher in the

justice condition (51%) than in the suffering condition (29%), $\chi^2(1, N = 121) = 4.942$, $p = .026$.

There is no significant difference between the revenge and suffering, ignorance and suffering, and justice and revenge conditions, $\chi^2(1, N = 131) = 2.019$, $p = .155$, $\chi^2(1, N = 126) = 2.672$, $p = .102$, and $\chi^2(1, N = 116) = 0.466$, $p = .495$, respectively.

The main results of the OLS regression hold as well: When limiting the analysis to angry Recipients only, the justice and revenge conditions are associated with a significantly higher likelihood of choosing the moderate option (Table 5):

Table 5: Regression results: Likelihood of choosing punishment options, among those Recipients who reported being angry ($n = 242$)

	Dependent variable:					
	Likelihood of choosing NO punishment		Likelihood of choosing MODERATE punishment		Likelihood of choosing SEVERE punishment	
	(1)	(2)	(3)	(4)	(5)	(6)
Suffer	−0.001	−0.009	0.139*	0.142*	−0.138	−0.133
	(0.053)	(0.052)	(0.082)	(0.083)	(0.087)	(0.087)
Explain	−0.065	−0.066	0.215**	0.207**	−0.150*	−0.141
	(0.054)	(0.054)	(0.084)	(0.085)	(0.089)	(0.089)
Identity	0.089	0.089	−0.081	−0.088	−0.008	−0.002
	(0.055)	(0.055)	(0.086)	(0.086)	(0.090)	(0.090)
Sex (F = 1)		0.090**		0.043		−0.133**
		(0.038)		(0.060)		(0.062)
Age (years)		−0.001		−0.003		0.004
		(0.002)		(0.003)		(0.003)
Constant	0.103***	0.084	0.155**	0.234**	0.741***	0.682***
	(0.039)	(0.068)	(0.060)	(0.107)	(0.064)	(0.112)
Observations	242	242	242	242	242	242
R^2	0.012	0.035	0.076	0.082	0.058	0.081
Adjusted R^2	−0.001	0.015	0.064	0.063	0.047	0.061
Note:						*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Interestingly, we also observe a marginally significant effect of the suffering component, which suggests that angry Recipients had at least some retributive motives, in addition to distributive and belief-based preferences.

Suspicion (not believing that the interaction was real). In this robustness check we excluded all participants who indicated a strong (but incorrect) belief that they were interacting with a robot, as opposed to interacting with another human. In the following analyses we excluded everyone who responded below -50 to the following question (i.e., indicated a serious doubt that their partner was real): "Using the slider below, please indicate the extent to which you believed you were interacting with a bot or human partner" (continuous scale from *-100: definitely a bot* to *+100: definitely a human*)

By applying the above criteria, we excluded 251 Recipients (31%), while 554 (69%) were retained in the sample. Figure 16 summarizes the main results:

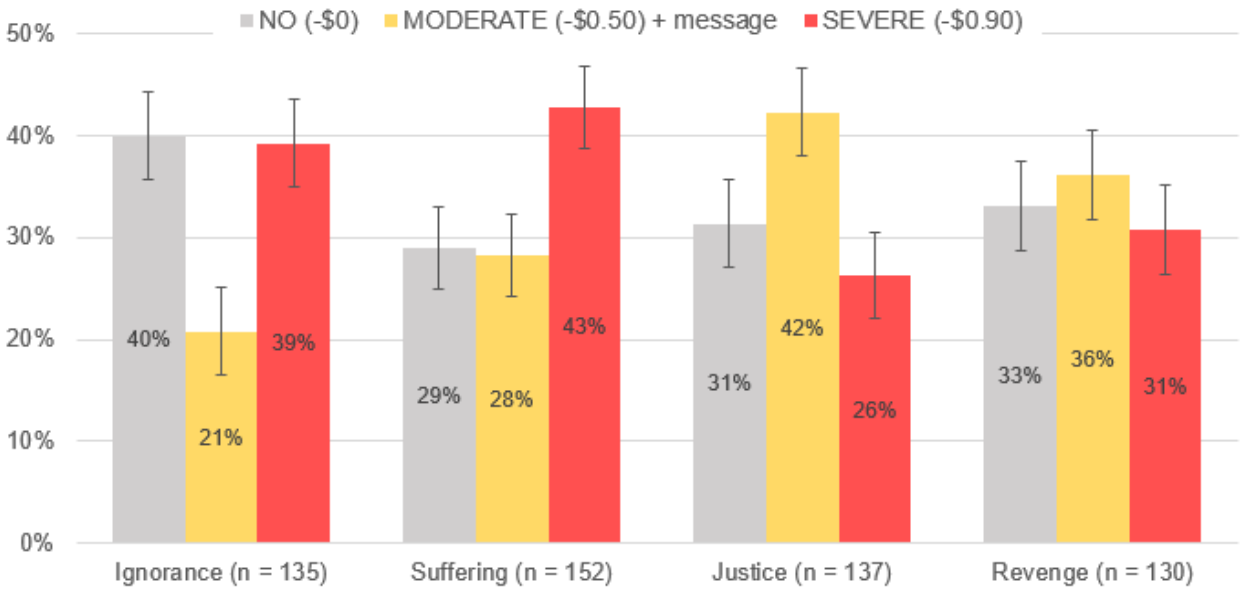

Figure 16: Proportion of Recipients choosing NO (grey), MODERATE (yellow), and SEVERE (red) punishment across conditions, among only those Recipients who had at least some reasonable confidence that they were interacting with another human. Error bars represent ±1 standard error.

All of the main results are robust to excluding suspicious Recipients: A significantly higher proportion of Recipients chose the moderate punishment option in the justice condition (42%) and in the revenge condition (36%) than in the ignorance condition (21%), $\chi^2(1, N = 272) = 13.684$, $p < .001$, and $\chi^2(1, N = 265) = 7.012$, $p = .008$, respectively. The proportion of moderate-choosers was also significantly higher in the justice condition (42%) than in the suffering condition (28%), $\chi^2(1, N = 289) = 5.651$, $p = .002$.

There is no significant difference between the revenge and suffering, ignorance and suffering,

and justice and revenge conditions, $\chi^2(1, N = 282) = 1.649$, $p = .199$, $\chi^2(1, N = 287) = 1.802$, $p = .180$, and $\chi^2(1, N = 267) = 0.825$, $p = .364$, respectively.

The main results of the OLS regression hold as well: Adding the explanatory component to the moderate option (i.e., justice and revenge conditions) is associated with a significantly higher likelihood of choosing the moderate option and a significantly lower likelihood of choosing the severe option (Table 6). In addition, adding the suffering component to the moderate option (i.e., all but the ignorance condition) is associated with a significantly lower likelihood of choosing no punishment, but this effect holds only if we exclude suspicious participants (cf. Table 2).

Table 6: Regression results: Likelihood of choosing punishment options, excluding Recipients who believed that they were interacting with a robot.

	Dependent variable:					
	Likelihood of choosing NO punishment		Likelihood of choosing MODERATE punishment		Likelihood of choosing SEVERE punishment	
	(1)	(2)	(3)	(4)	(5)	(6)
Suffer	-0.111^{**}	-0.108^{*}	0.075	0.071	0.035	0.037
	(0.056)	(0.056)	(0.054)	(0.054)	(0.056)	(0.056)
Explain	0.024	0.026	0.140^{***}	0.140^{**}	-0.165^{***}	-0.165^{***}
	(0.055)	(0.055)	(0.054)	(0.054)	(0.056)	(0.056)
Identity	0.017	0.018	-0.062	-0.062	0.045	0.044
	(0.058)	(0.057)	(0.056)	(0.056)	(0.058)	(0.058)
Sex (F = 1)		0.052		0.038		-0.089^{**}
		(0.040)		(0.039)		(0.040)
Age (years)		0.003^{*}		-0.003^{*}		0.0002
		(0.002)		(0.002)		(0.002)
Constant	0.400^{***}	0.271^{***}	0.207^{***}	0.298^{***}	0.393^{***}	0.431^{***}
	(0.041)	(0.071)	(0.040)	(0.070)	(0.041)	(0.072)
Observations	554	554	554	554	554	554
R^2	0.008	0.017	0.030	0.038	0.019	0.028
Adjusted R^2	0.002	0.008	0.025	0.029	0.014	0.019
Note:						$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$