Abstract

There are two broad functional explanations for second-party punishment: fitness-leveling and deterrence. The former suggests that people punish to reduce fitness differences, while the latter suggests that people punish in order to reciprocate losses and deter others from inflicting losses on them in the future. We explore the relative roles of these motivations using a pre-registered, two-player experiment with 2,426 US participants from Amazon Mechanical Turk. Participants played as the "responder" and were assigned to either a Take or Augment condition. In the Take condition, the "partner" could steal money from the responder's bonus or do nothing. In the Augment condition, the partner could augment the responder's bonus by giving them money at no cost to themselves or do nothing. We also manipulated the responders' starting endowments, such that after the partner's decision, responders experienced different payoff outcomes: advantageous inequity, equality, or varying degrees of disadvantageous inequity. Responders then decided whether to pay a cost to punish the partner. Punishment was clearly influenced by theft and was most frequent when theft resulted in disadvantageous inequity. However, people also punished in the absence of theft, particularly when confronted with disadvantageous inequity. While the effect of inequity on punishment was small, our results suggest that punishment is motivated by more than just the desire to reciprocate losses. These findings highlight the multiple motivations undergirding punishment and bear directly on functional explanations for the existence of punishment in human societies.

1.1. Introduction

Punishment has traditionally been defined as a costly behavior aimed at those who cause harm or violate social norms (Clutton-Brock & Parker, 1995) and is thought to be critical for maintaining cooperation in humans (Boyd et al., 2010; Boyd & Richerson, 1992; Fehr & Fischbacher, 2004; but see Raihani & Bshary, 2019). Human adults show striking willingness to pay a cost to punish defecting partners in both second-party contexts in which the punisher is directly harmed by the transgressor (Fehr & Gachter, 2000; Güth et al., 1982; Bone & Raihani, 2015), and in third-party contexts in which the punisher is an uninvolved observer (Fehr & Fischbacher, 2003). Costly punishment in these contexts is seen across human societies (Balliet & Van Lange, 2013; Heinrich et al., 2006), highlighting its ubiquity in human social interactions. Nevertheless, other studies indicate that punishment might undermine, rather than support, cooperation (e.g. Dreber et al. 2008; Nikiforakis 2008; Wu et al. 2009). Recent studies of the proximate motives underpinning punishment suggest that it is often used in the absence of any cheating by the target (e.g. Herrmann et al., 2008; Abbink & Sadrieh, 2009; Raihani & McAuliffe, 2012), and when there is no possibility for the target to change their behavior in future interactions (reviewed in Raihani & Bshary, 2019).

There are two broad functional explanations for why we punish others. The first explanation is that punishment serves a *deterrence* function by deterring the target from harming the punisher in the future (Delton & Krasnow, 2017; McCullough, Kurzban, & Tabak, 2013). The deterrence hypothesis predicts that individuals should be motivated to punish those who intentionally harm them, regardless of whether punishment equalizes the payoff differentials between defectors and punishers. A second functional explanation is that punishment evolved to

serve a competitive function, by increasing relative fitness differences between punishers and cheaters (Bone & Raihani, 2015; Price et al., 2002). The competition hypothesis predicts that individuals should punish in order to increase their relative payoffs and that any deterrence function would occur as a by-product (Raihani & Bshary, 2019). Subsumed under the competition hypothesis is fitness-leveling theory, which predicts that individuals will punish in order to equalize payoffs and reduce their fitness disadvantage. Although these explanations do not uniquely apply to second-party punishment, here we focus solely on this form of punishment.

Recent empirical work has investigated these functional explanations by exploring the proximate mechanisms that motivate punishment decisions. Specifically, if punishment serves a deterrence function, then punishment should be proximately motivated by the desire to reciprocate losses (henceforth 'revenge'). If punishment serves a competitive or fitness-leveling function, then punishment should be proximately motivated by the desire to increase relative payoffs and inequity aversion, respectively. Studies that hold losses constant (e.g. Raihani & McAuliffe 2012) have suggested that punishment is frequently motivated by disadvantageous inequity aversion. In Raihani & McAuliffe (2012), participants played as a responder in a two-person theft game in which an actor (thief) could cheat by stealing $0.20 from the responder. Responders could then choose to pay $0.10 to reduce the actor's bonus by $0.30. Importantly, responders who interacted with a thief experienced the same magnitude of losses but different kinds of outcomes. Depending on which of three starting-endowment conditions responders were assigned to, stealing resulted in advantageous inequity, equality, or disadvantageous inequity for the responder. Responders who experienced losses that resulted in disadvantageous inequity were the most likely to punish the thief, whereas responders who experienced losses that did not

results in inequity were less likely to punish, suggesting that punishment was motivated primarily by disadvantageous inequity rather than a desire for revenge. However, the design of the experiment did not completely disambiguate the role of disadvantageous inequity from experiencing losses because participants only experienced disadvantageous inequity if they experienced losses, rendering it impossible to fully isolate their respective roles in motivating punishment.

In a follow-up study, Bone & Raihani (2015) used a similar two-person theft game to disentangle whether responders used punishment to equalize outcomes between themselves and the thief. This study was similar to Raihani & McAuliffe (2012) but, rather than punishment being a binary choice, responders could purchase from 0-4 'punishment points', each of which cost $0.05 to purchase and which inflicted costs of $0.05 and $0.15 on the target in the *ineffective* and *effective* conditions, respectively. As before, participants played in the role of responder and were assigned to one of five different starting endowments where they experienced advantageous inequity, equality, or various levels of disadvantageous inequity if the partner stole. In this study, responders punished theft more than non-theft, even when they did not experience disadvantageous inequity or when punishment was ineffective, suggesting that punishment was partly motivated by the desire to reciprocate losses. However, in the effective punishment condition, punishment increased if theft resulted in disadvantageous inequity, and punishers also tailored their investment in punishment to create equal outcomes when this was possible. When achieving equal outcomes was not possible, punishers typically invested in the harshest punishment possible. In line with Bone & Raihani (2015), Bone, McAuliffe, & Raihani (2016) found that punishment was motivated by both revenge and inequity aversion and that

these motivations might differ across countries. Together, these studies suggest that punishment is motivated by the punisher's desire to reciprocate losses as well as by the desire to increase their relative payoffs. These motives are consistent with punishment serving both a deterrent and a fitness-leveling function, respectively.

However, recent work has cast some doubt on these conclusions. In a recent study, participants played a version of the two-person theft game with three conditions in which they always assumed the role of responder (Marczyk, 2017). The partner either took 20 points from the punisher (Take condition), destroyed 20 of the punisher's points (Destruction condition), or added 20 points to their own score (Augmentation condition). Responders could then punish the partner by deducting 30 points from their bonus for free. Punishment was most common when the partner inflicted costs on the responder (i.e. the Destruction and Take conditions) and inequity aversion only became relevant as a secondary input once costs had been inflicted. These results align with the previously discussed findings, showing that punishment decisions are sensitive to both losses and disadvantageous inequity (Raihani & McAuliffe, 2012; Bone & Raihani, 2015). However, in conflict with past work, the low levels of punishment observed in the Augmentation condition were interpreted as evidence against inequity aversion as a proximate motive for punishment, with the inference being that inequity aversion only motivates punishment when it is accompanied by loss.

Before ruling out the role of inequity aversion in motivating punishment in the absence of losses, it is important to note that the low levels of punishment in the Augmentation condition in Marczyk (2017) may have resulted from the design of the Augmentation condition. In this

study's Augmentation condition, the actor was afforded the opportunity to add 20 points to their *own* endowment, resulting in disadvantageous inequity for responders. It is possible that punishment was infrequent in this condition because, despite experiencing disadvantageous inequity, responders did not wish to punish the other player for making a decision to increase their own endowment when this decision was detached from the responder's endowment. In other words, responders might not have been motivated to punish the other player for accepting free money, despite experiencing disadvantageous inequity, because the other player's augmentation decision had no connection to their own payoff. Additionally, given that previous work has found that inequity aversion does plays a role in punishment (Raihani & McAuliffe, 2012; Bone & Raihani, 2015), we should not disregard the role of inequity aversion on the specifics of one study's design. Thus, we cannot yet rule out the potential role that inequity aversion—the disutility associated with having less than someone else—plays in motivating punishment in the absence of losses.

A stronger test of the independent role that inequity aversion plays in motivating punishment would investigate whether it matters in the absence of losses, yet in a situation where the partner's decision actually affects the punisher's payoff. In other words, to understand the role of inequity aversion in motivating punishment decisions, it is necessary to examine how responders behave when the partner's decision to augment is connected to the responder's own payoffs. Understanding the role of inequity aversion in the absence of loss will help inform our understanding of why we punish others, painting a fuller picture of the motives and possible functions of punishment.

Here, we use a two-person punishment game to test whether inequity aversion promotes punishment in the absence of experiencing loss. In a preregistered, 3 × 4 × 2 between-subject design we manipulated the structure of the previously used two-person theft game, the degree of inequity, and the partner's behavior. Participants were assigned to the responder role in either a Take or one of two Augment conditions. In the Take condition, the partner had the option to steal $0.10 from the responder or do nothing. In the Augment condition, the partner had the option to augment the responder's bonus by $0.10 at no cost to themselves or do nothing. Based on the starting endowments, theft or augmentation resulted in various degrees of disadvantageous inequity, equality, and advantageous inequity. We also compared punishment decisions between cases where the partner decided to steal or augment and when they decided to do nothing.

If punishment is motivated by inequity aversion, we expect to see responders using punishment most frequently in cases where there they experience disadvantageous inequity, regardless of whether they also experienced losses. If punishment is motivated by revenge, then punishment should occur most frequently when responders incur losses, such as in the Take condition when the partner steals, regardless of inequity. However, if both inequity aversion and revenge motivate punishment, then we expect to see punishment in cases where responders experience both disadvantageous inequity and losses, with the highest levels of punishment when theft results in disadvantageous inequity.

1.2. Method

1.2.1. Participants

We tested N = 2,426 participants (55.89% female), aged 19-81 (M = 38.42) using Amazon's Mechanical Turk, a crowdsourcing website, and TurkPrime, a user interface for Mechanical Turk (Buhrmester, Kwang, & Gosling, 2011; Litman, Robinson, & Abberbock, 2017). Data collected online using Mechanical Turk has been found to be comparable to data collected in the lab with US participants (Amir & Rand, 2012; Horton, Rand, & Zeckhauser, 2011; Rand, 2012). Participants received a base payment of $0.50 for participating and had the potential to earn a bonus payment ranging from $0.05 to $0.90 depending on the condition they were assigned to and their choices in the experiment. Participants (1) were located in the United States (2) had completed between 100 and 10,000 HITs prior to this study; (3) had a HIT approval rate between 97–100%, and (4) had a unique IP address. In addition, 178 participants were tested but excluded for failing to complete the survey in its entirety and 55 participants were excluded for answering any comprehension question incorrectly more than once (see Procedure section below). Comprehension questions were coded by the first author and a second coder. Inter-rater reliability was excellent (Kappa = 95.5%) and discrepancies were resolved by agreement between coders.

1.2.2. Design

Participants provided informed consent and were assigned to the role of responder and randomly assigned to one of three conditions (Take, Augment 1, and Augment 2), one of four initial starting endowment treatments (treatments A-D; Table 1), and one of two partner behaviors (steal/augment, do nothing). In total, participants were assigned to one of 24 different combinations of condition, treatment, and partner behavior. We used an ex-post matching procedure to pair responder and partner behavior: after participants made their decisions as

responders, they assumed the partner role and made a series of decisions for each condition and treatment (Rand 2012; Bone & Raihani, 2015) using the strategy method (Fischbacher, Gachter, & Quercia, 2012; Jordan, McAuliffe, & Rand, 2016). We collected these decisions in order to pair up participants without using deception and therefore were not interested in analyzing these data. We included the Augment 2 condition in order to compare punishment in the Take condition when the partner did something (i.e., stole) with punishment in the Augment condition when the partner did nothing (i.e., did not augment) when the degree of inequity between the responder and the partner was held constant between them.

Figure 1. Diagram illustrating the possible actions in the Take and Augment conditions of the game (Augment 2 was identical to Augment 1 in terms of available actions). In Stage 1, both participants were given an endowment and the partner decided whether to steal $0.10 from the responder (or augment the responder's bonus by $0.10 at no cost to themselves) or do nothing. In Stage 2, the responder saw the partner's decision and decided whether to pay $0.05 to reduce the partner's endowment by $0.15.

1.2.3. Procedure

Participants provided informed consent and were given instructions for the task that detailed the rules of the 2-player punishment game. After reading the instructions, participants answered two comprehension questions to ensure they understood the task (see supplementary materials). If they answered any comprehension question incorrectly, they were redirected back

to the task instructions and asked to answer the question again before moving onto the rest of the survey. Participants who answered any comprehension question incorrectly twice were excluded from analysis (see information on exclusions above). After answering the comprehension questions, participants began the task.

The game consisted of two stages. In Stage 1 of the Take condition, the responder was told that the partner either decided to steal $0.10 from them or do nothing. In Stage 1 of the Augment condition, the responder was told that the partner either decided to augment the responder's endowment by $0.10 at no cost to themselves or do nothing. In Stage 2, across conditions, the responder decided whether or not to pay $0.05 to reduce the bonus of the partner by $0.15. The two Augment conditions were identical with the exception that the starting endowments in the Augment 2 condition were altered to match the post-theft payoffs in the Take condition (see Table 1). Doing so allowed us to directly compare punishment when the partner decided to steal $0.10 and when the partner decided *not* to augment the responder's bonus by $0.10 while controlling for payoffs.

The responder always experienced the same magnitude of loss or gain when the partner stole (or augmented) $0.10. However, depending on the treatment, the partner's decision to steal resulted in equal payoffs (70:70, 70 for the responder and 70 for the partner, or differing levels of disadvantageous inequity for the responder (50:70, 30:70, 10:70). In Augment 1, the partner's decision to augment resulted in two levels of advantageous inequity (90:60,70:60), and two levels of disadvantageous inequity (30:60, 50:60) for the responder. In Augment 2, the partner's

decision to augment resulted in one level of advantageous inequity (80:70), and three levels of disadvantageous inequity (20:70, 40:70, 60:70) for the responder.

Table 1. The payoffs experienced by the responder and the partner for treatments A-D (in cents). Stage 1 endowments varied according to treatment and Stage 2 payoffs varied depending on whether the partner acted (stole/augmented or did nothing). Degree of inequity is the payoff difference between the responder's and the partner's bonus.

After participants made their punishment decisions as the responder, we elicited partner behavior using the strategy method (Fischbacher, Gachter, & Quercia, 2012; Jordan, McAuliffe, & Rand, 2016). All participants made a series of 12 decisions, one for each treatment (A-D) and condition (Take, Augment, Augment 2), to either steal $0.10 from the responder or do nothing for the Take condition and to either augment the responder's bonus by $0.10 or do nothing for the Augment conditions. This resulted in 24 different outcomes, one for each possible combination of condition, treatment, and partner behavior. Before making their decisions, participants were instructed that there was a chance that one of their decisions would be randomly selected and shown to the responder and that the responder might or might not react to that decision. One participant's decision as the partner was randomly selected for each of the 24 condition/treatment/partner decision combinations and matched with all responders who were randomly assigned to that combination (approximately 100 per cell). Bonuses for the partner

performance were paid out to workers whose partner decisions were chosen and matched with other workers.

After making their decisions as the responder and the partner, participants completed Likert scale measures of childhood and current socioeconomic status and demographic questions assessing age, race, and educational attainment. They also answered two self-report questions assessing whether they believed the other player in the task was real, to what extent they had previously participated in similar economic games. The childhood and current socioeconomic status questions were asked for exploratory purposes and are not presented here.

1.3. Analyses

All analyses were conducted in R version 3.5.1 (R Core Team, 2018). For all models, we used generalized linear models (GLMs) with punishment as a binary response term (1 = punished, 0 = did not punish) and conducted model comparisons using likelihood ratio tests.

To test whether punishment was predicted by degree of inequity, condition (Take/Augment), or partner behavior (did something/did nothing), we first constructed a full model with all three of our explanatory terms of interest and their interactions, as well gender and age (in years) as covariates. We then compared the full model with a model without the three-way interaction term but that included all three possible two-way interactions. The model with the three-way interaction term did not offer a significantly better fit to the data than the one without it (see Table 2 for model output), so we dropped the three-way interaction term. Because we were not interested in the interaction between condition and degree of inequity (see

preregistration), we then compared the model with both two-way interaction terms of interest (condition × partner behavior; degree of inequity × partner behavior) to models without those terms (see Table 2 for model output). The model including the condition × partner behavior interaction significantly better fit the data than one without it while the model including the degree of inequity × partner behavior interaction did not. This resulted in a final model with the minimal number of terms that provided the best fit to the data. This reduced model had five significant predictors of punishment: condition, degree of inequity between the responder and the partner, partner behavior, age, and the interaction between condition and partner behavior (see Table 2 for model output).

For all analyses reported here, including both preregistered and exploratory, we combine the Augment 1 and Augment 2 conditions into a single Augment condition unless otherwise specified. For transparency, we also report the results for the degree of inequity term from two analyses including each Augment condition separately (e.g., one model with only Augment 1 and one model with only Augment 2) and we report the results for the full models in the supplementary materials. All models included only participants who selected male or female as their gender because we had insufficient observations to make comparisons for participants who selected a different response to this question (i.e., transgender, prefer not to answer); thus our models include 12 fewer participants than in our total sample size. As per our preregistration, we also ran the same analyses (with punishment as a binary response term) with partner behavior coded as "nasty" (the partner decided to steal or not augment) or "nice" (the partner decided to not steal or augment). We report those results in the supplementary materials but note that the

results of these analyses do not differ substantially in interpretation from the ones we report below.

As a post-hoc exploratory analysis, we investigated whether the effect of inequity on punishment was robust to different analytical approaches by constructing a model with degree of inequity treated as a categorical variable with three levels (disadvantageous inequity = treatments A & B, equality = treatment C, advantageous inequity = treatment D). This model included the six terms present in our reduced model: condition (Take/Augment), partner behavior (did something/did nothing), the condition × partner behavior interaction, degree of inequity, and age and gender as covariates. We also replicated our preregistered analyses using contrast coding for the condition and partner behavior variables, which allowed us to test for the main effect of degree of inequity while still including terms of interest. We report the results of these models in the supplement but note that the results do not differ substantially in interpretation from those reported here. Finally, we also conducted an exploratory analysis testing for a non-linear relationship between degree of inequity and punishment. We report these results in the supplementary materials but note here that these analyses found that a quadratic term for degree of inequity better fit the data than a linear term, suggesting the relationship between degree of inequity and punishment is better explained as curvilinear than linear. However, because this analysis was introduced during the review process and because we had no a priori reason to predict a non-linear relationship we do not detail those findings here.

1.4. Results

Looking at the overall percentage of punishment reveals that participants seemed to punish more frequently in the Take condition (12.64%) than the Augment conditions (9.71%). Participants punished the partner most frequently in treatment B (12.79%), followed by treatment A (11.35%), treatment C (11.29%), and treatment D (7.33%; see Figure 2 for proportions of punishment across all conditions, treatments, and partner behaviors). The full model including the three-way interaction between condition, degree of inequity, and partner behavior did not provide a significantly better fit to the data than the model without it (LRT, $\chi^2(1) = 0.03$, $p = .87$). We next examined the partially reduced model that included both the degree of inequity × partner behavior and condition × partner behavior interactions and compared it to models without either interaction (see Table 2 for model output). The model that included the degree of inequity × partner behavior interaction did not provide a significantly better fit to the data than the model without it (LRT, $\chi^2(1) = 0.92$, $p = .34$). The degree of inequity × partner behavior interaction term was not significant (B = 0.006, SE = .006, $p = .34$, OR: 1.01, 95% CI: 0.99, 1.02), suggesting that punishment did not vary by the degree of inequity across the partners' behavior differently. The model that included the condition × partner behavior interaction provided a significantly better fit to the data than the model without it (LRT, $\chi^2(1) = 61.69$, $p < .001$). The condition × partner behavior interaction was significant (B = -2.44, SE = .33, $p < .001$, OR: 0.09, 95% CI: 0.05, 0.17); we interpret this interaction when reporting the results of our reduced model below.

In our final, reduced model (see Table 2 for the model output), we found that punishment was most frequent when the partner did something (i.e., stole) in the Take condition and when the partner did nothing (i.e., did not augment) in the Augment condition (B = -2.40, SE = .33, $p <$

.001, OR: 0.09, 95% CI: 0.05, 0.17; Figure 3). Critically, responders were more likely to punish when they experienced a greater degree of disadvantageous inequity (B = -.007, SE = .003, *p* = .01, OR: 0.99, 95% CI: 0.98, 0.99). Finally, we found that age predicted punishment, such that as age increased, participants showed a decreasing tendency to punish (B = -0.02, SE = .006, *p* < .001, OR: 0.98, 95% CI: 0.97, 0.99) and there was no sex difference in the tendency to punish (B =-0.09, SE =.14, *p* = .48, OR: 0.91, 95% CI: 0.69, 1.19).

Figure 2. A bar graph comparing the frequency of punishment decisions between conditions (Take, Augment 1 and Augment 2), starting endowment treatments (A: strong disadvantageous inequity, B: disadvantageous inequity, C: equality, D: advantageous inequity), and partner behavior (do something – take or augment, or do nothing). Numbers in brackets refer to the starting endowments (responder:partner) for the responder before the partner chose to steal from/augment the responder's payoff. Error bars show standard deviation.

We next re-analyzed the reduced model examining Take vs. Augment 1 (i.e., excluding Augment 2). We found that responders were more likely to punish as the degree of inequity increased (B = -.01, SE = .004, p = .003, OR: 0.99, 95% CI: 0.98, 0.99). We then re-analyzed the reduced model examining Take vs. Augment 2 (i.e., excluding Augment 1). Unlike our results from the analyses using solely Augment 1 or combing both Augment conditions, the degree of inequity between the responder and the partner did not significantly predict punishment (B = -.005, SE = .004, *p* = .16, OR: 0.99, CI: 0.99, 1.00). We interpret this finding in the discussion.

Next, we explored whether the partner stealing was perceived similarly to the partner not augmenting. If stealing $0.10 is perceived as similarly egregious as not giving $0.10, then the frequency of punishment between these two cases should be the same. However, if stealing is perceived as worse than not giving, we expect to see more punishment when the partner stole than when the partner did not give. For this analysis only, we used the results from the Augment 2 condition because the starting endowments were equivalent in this condition to the final payoffs when the partner stole in the Take condition. Stealing was punished more than not augmenting (stealing: 20.45%; not augmenting: 14.39%), suggesting that stealing was perceived as a worse transgression than not augmenting ($\chi^2(1) = 4.75$, $p = .029$). We replicate this analysis using the Augment 1 condition in the supplementary materials.

We then asked whether not stealing (in the Take condition) was punished as frequently as failing to augment in the Augment 1 condition. For this and the following analysis we used the results from the Augment 1 condition because the starting endowments in this condition were equivalent to the starting endowments in the Take condition (see the supplement for these analyses using Augment 2). Not augmenting was punished more than not stealing (not augmenting: 13.67%; not stealing: 4.96%), suggesting that failing to augment was perceived as worse than not stealing ($\chi^2(1) = 16.99$, $p < .001$). Additionally, we compared cases where the partner chose to not steal $0.10 from the responder with cases where the partner chose to augment the responder's endowment by $0.10. If not stealing was perceived as equivalently wrong as augmenting, then we should see equal levels of punishment. Augmenting and not stealing were punished at equivalently low levels (augmenting: 4.90%; not stealing: 4.96%),

suggesting that augmenting and not stealing were perceived as equally wrong ($\chi^2(1) = 0.00$, $p = 1.0$).

Lastly, the results of the exploratory analysis with degree of inequity as a categorical variable found that participants were significantly more likely to punish when they experienced disadvantageous inequity compared to equality (B = -0.53, SE = .19, p = .006, OR: 0.59, 95% CI: 0.40, 0.85), as well as when there was disadvantageous inequity compared to advantageous inequity (B = -0.50, SE = .22, p = .025, OR: 0.61, 95% CI: 0.38, 0.93). The full model output is reported in Table 2 in the SOM.

Table 2. Estimate and standard error of fixed effects in Generalized Linear Models predicting punishment. Baselines were set as follows: condition = Take, gender = male, partner behavior = nothing. The full model includes all three explanatory terms of interest and their interactions, as well as gender and age as covariates. The partially reduced model was the intermediate model between the full model and the reduced model which contained both two-way interactions of interest. The reduced model was the final model with the minimal number of terms that provided the best fit to the data. Table also shows goodness-of-fit statistics.

Figure 3. Plot of predicted effects from models of punishment for when the partner did something (took $0.10/augmented $0.10) or did nothing across the degree of inequity between responder and partner (negative values correspond to the responder having less than the partner). Each panel represents a different condition (Take, Augment 1, Augment 2). Note that plots include extrapolated values for the degree of inequity between responder and partner. Ribbons depict 95% confidence intervals.

1.5. Discussion

Our main aim was to investigate whether inequity aversion motivates punishment in the absence of experiencing losses. Responders who had money stolen from them by the partner were the most likely to punish, however, disadvantageous inequity also predicted punishment, even in the absence of theft. In other words, responders who experienced disadvantageous inequity were more likely to punish, although the effect size of the inequity term was small, indicating a relatively weak effect. Additionally, while we found a significant linear effect of inequity on punishment, this linear trend might be better explained by a categorical difference between disadvantageous inequity and non-disadvantageous inequity. In other words, people are more likely to punish when experiencing disadvantageous inequity relative to equality or advantageous inequity, but the likelihood of punishment does not increase as disadvantageous inequity increases. Thus, our results suggest that, while punishment is primarily motivated by the desire to reciprocate losses, this is not the sole motivation; inequity aversion also plays a minor role in motivating punishment. Our findings illuminate the motivations that underlie punishment, highlighting the potentially unique role of inequity aversion as a motivation in human

punishment. This supports past work suggesting that punishment is motivated by both the desire to reciprocate losses and, to a lesser extent, inequity aversion, and that punishment is most likely when losses also result in disadvantageous inequity (Bone & Raihani, 2015; Bone, McAuliffe, & Raihani, 2016).

Punishment is a behavior that is sensitive to many, often competing, concerns (Raihani & Bshary 2019). It is likely motivated largely by the desire to reciprocate losses but, as we show here, can be triggered by the experience of having less than others even when no losses occur (Dawes et al. 2007). Our results are consistent with both the deterrence and fitness-leveling explanations of punishment. Specifically, our finding that revenge motivates punishment is in line with deterrence theory (Delton & Krasnow, 2017) and our finding that inequity aversion motivates punishment is in line with fitness-leveling theory (Price et al., 2002). Since these explanations are not mutually exclusive, it is likely that both contribute and interact to shape second party punishment in humans.

While a previous study found evidence that inequity aversion does not motivate punishment in the absence of losses, this conclusion rested on the results of an augmentation condition, where people did not punish partners who augmented their own income and became richer by doing so (Marcyzk, 2017). That finding conflicts with the evidence from our 'Augment 1' condition, where we found that punishment increased as the degree of inequity between the responder and partner became larger. The critical difference between our augment conditions is that in Marcyzk (2017), the partner decided to augment their own endowment whereas in our study the partner could augment the responder's endowment, thereby tying the partner's

augmentation decision to the responder's payoff. This suggests that inequity aversion might be most likely to influence punishment when the partner's decision is directly tied to the punisher's own outcome. In other words, if an actor's decision results in disadvantageous inequity but the decision is removed from the punisher's own payoff or outcome, then inequity aversion might be less likely to trigger punishment. Relatedly, participants might have inferred different motives in the partners decision to augment between the Augmentation condition in Marcyzk (2017) and our Augment conditions. Namely, participants might have inferred relatively benign motives in the partner's decision to augment their own endowment in the Augmentation condition used in previous work whereas in our Augment conditions, participants might have inferred more malicious intentions in the partner's decisions not to augment. Indeed, recent work has found that attributions of harmful intent in ambiguous social interactions positively predict punishment (Raihani & Bell, 2019). Future work should continue to investigate how inferences about a partner's intentions when deciding to not augment or steal influence punitive behavior.

While the differences between our Augment conditions might explain why our results differ from Marcyzk (2017), we also found different results between the two Augment conditions used in our study, such that degree of inequity predicted punishment in Augment 1 but not Augment 2. This suggests that punishment in the Augment conditions might be sensitive to minor differences in the design, such as the starting endowments and relative cost of punishment. We discuss the differences between our Augment conditions in greater detail below. When taken together, the results from our Augment conditions and Marcyzk's (2017) Augmentation condition suggest that inequity aversion motivates punishment, but only under specific

conditions, such as when augmenting is tied to the responder's payoff and punishment is relatively effective at reducing resource differentials.

The difference in punishment observed between our study and past work suggests that when deciding to punish, people might be sensitive to how punishment is perceived. Namely, punishment of a partner's individual decision to augment their own endowment, even if it results in disadvantageous inequity, might be perceived as less permissible than punishment of a partner who fails to confer a benefit which results in disadvantageous inequity. Punishment is most effective in changing behavior when it is perceived as fair and legitimate (Fehr & Rockenbach, 2003; Houser & Xiao, 2010; Raihani & Bshary, 2019). When punishment is perceived as morally illegitimate, such as when it is motivated by selfishness, it is less likely to deter future cheating and can increase the risk of retaliation (Raihani & Bshary, 2019; Xiao, 2013). Our study does not directly address the role of moral permissibility in punishment decisions which might differ between cases where the partner's decision to augment is a disconnected decision or tied to the responder's own payoff. Understanding the role of moral permissibility in punitive behavior will be a fecund area for future work.

Our data also provide some support for the theory that punishment may serve a competitive function to increase relative payoffs and status (Raihani & Bshary, 2019). All punishment in this study used a 1:3 fee-to-fine ratio, the most commonly used ratio in economic games studying punishment (Raihani & Bshary, 2019). Because this ratio allows punishers to gain a relative advantage, it is difficult to disentangle any punishment found here from a competitive motivation to increase relative payoffs. Additionally, we found that around 5% of

responders punished the partner for either augmenting or not stealing, suggesting that a small minority of participants were punishing regardless of the partner's behavior, which might correspond to a competitive motivation to increase relative payoffs.

One question arising from our results is why degree of inequity did not predict punishment when including only the Augment 2 condition. This difference might stem from the different starting endowments between the two Augment conditions. Specifically, the Augment 2 condition differed from the Augment 1 condition in that it was skewed towards strong disadvantageous inequity (in terms of responders' outcomes) and there was no advantageous inequity outcome. Because participants punished less when there was strong disadvantageous inequity compared to weak disadvantageous inequity, this might have contributed to the non-significant effect of degree of inequity in the Augment 2 condition. However, this in turn raises the question of why participants did not punish more when there was strong disadvantageous inequity, as in treatment A, relative to when there was weak disadvantageous inequity, such as in treatment B. On its face, this result is counterintuitive as a strong account of fitness leveling theory would predict higher levels of punishment as disadvantageous inequity increases. However, this finding makes more sense when considering how ineffective punishment was in this treatment. Namely, punishment in treatment A was relatively more costly compared to treatment B—participants in this treatment had to pay 16-50% of their endowment, depending on the condition and the partner's behavior, to punish the other player. Additionally, punishment in this condition was highly inefficient at reducing the relative degree of inequity between players; even after punishment, the partner would still have a relatively larger endowment than the responder. In other words, punishment may have been less frequent in treatment A than

treatment B because it was ineffective in the sense that it substantially decreased the responders' own endowments while doing little to close the gap in resources between the responder and the partner.

That participants were sensitive to the cost and effectiveness of punishment in our study suggests that people have multiple, competing concerns that motivate punishment beyond a motivation to level payoff differentials and reciprocate losses. This finding is in line with recent work which argues that punishment often stems from multiple motives and serves different ultimate functions (Raihani & Bshary, 2019). Additionally, that participants were most likely to punish when they experienced intermediate levels of disadvantageous inequity suggests that inequity aversion is most likely to motivate punishment when punishment is effective at reducing relative resources differences. An important area of future work will be to disentangle how the relative cost, effectiveness, and degree of inequity influence punishment.

Our finding that participants were willing to pay a cost to reduce inequity aversion is in line with previous work using the Ultimatum Game. In the Ultimatum Game, one player is given an endowment of which they can share any portion of with another player. The second player then can either accept the offer, in which case both players receive their portion of the endowment, or reject, in which case both players get nothing. Rejecting any non-zero offer in the Ultimatum Game can be viewed as a form of punishment because it involves incurring a personal cost to reduce the other player's payoff. A significant body of work has demonstrated that humans are often willing to reject offers that are perceived as unfair in the Ultimatum Game (Güth et al., 1982; Heinrich et al., 2006), a finding that has been suggested to be motivated by

status concerns (Yamagishi et al., 2012). Furthermore, these findings from the Ultimatum Game provide indirect evidence that people are willing to punish inequity aversion even in the absence of experiencing loses. Thus, our results and the findings from the Ultimatum Game provide converging evidence that people are willing to punish inequity aversion in the absence of experiencing losses. Moreover, our findings help tackle this question more directly by comparing proximate motivations within one paradigm. Indeed, our study clearly demonstrates the relative strength of inequity aversion and revenge, as well as their additive influence, on motivating punishment decisions—something we would have been unable to do by just comparing data from previous Ultimatum and theft games. More generally, our data contribute to an emerging understanding of punishment as a behavior that is likely driven by a myriad of motivations.

One concern regarding our findings is that not giving in the Augment conditions might have been encoded by participants as a form of loss. That is, responders might have held the expectation that the partner would augment their endowment, thus responders might have perceived their endowment as a loss relative to their endowment if the partner had chosen to augment. In other words, if responders expected their partner to augment their endowment, then those responders whose partners failed to augment may have experienced a loss in the same way as participants who had money taken from them. While we recognize that this is possible, two lines of evidence suggest that this is unlikely to account for our pattern of findings. First, we saw differences in punishment across conditions suggesting that participants perceived these situations differently. Secondly, our Chi-Squared analyses show that stealing was punished more than not augmenting, and not augmenting was punished more than not stealing. This suggests that stealing was perceived as a worse transgression than failing to augment and that failing to

augment was perceived as worse than not stealing. However, it is still possible that not augmenting was perceived as a loss to a smaller extent than stealing. Because our data cannot disentangle these possibilities, future work should further explore the extent to which failing to augment another person's bonus at no cost is perceived as a form of loss.

A potential limitation of our study design is the extent to which the small endowments used here actually correspond to real-world decisions and, even more broadly, to fitness. Previous work has demonstrated that there are no significant differences in behavior in respect to stake size in studies run on Mechanical Turk (Amir & Rand, 2012; Horton, Rand, & Zeckhauser, 2011), which suggests that people treat the relatively small endowments ($.50) in the game similarly to more substantial endowments ($10) in lab. Regardless of stake size, while endowments used in economic games cannot directly measure fitness, they can be used to track preferences, the latter of which been shaped by selection due to their effects on material payoffs in social interactions (Alger, Weibull, & Lehmann, 2020).

While in our study we focused solely on an American sample, future work should explore how culture affects the motivations underlying punishment. Recent work suggests that there are meaningful cultural differences in punishment, such that participants from India punish at higher levels overall than American participants (Bone, McAuliffe, & Raihani, 2016) and that countries with weak rule of law experience higher levels of antisocial punishment—the punishment of cooperative individuals (Herrmann et al., 2008). It is currently unclear what explains these differences in punitive behavior. One possible explanation is that countries with weak rule of law that lack formal enforcement mechanisms have greater levels of local competition for resources,

resulting in higher levels of competitive punishment aimed at increasing relative payoffs and status (Herrmann et al., 2008; Sylwester et al., 2013). Future work should explore the role of revenge and inequity aversion in cultures that rely on more informal enforcement mechanisms. For example, inequity aversion might motivate punishment more in cultures that lack strong formal enforcement mechanisms and where there is more intense local competition for resources, environments where fitness-leveling is an especially beneficial strategy (Bone, McAuliffe, & Raihani, 2016).

In conclusion, we investigated whether punishment was motivated by inequity aversion in the absence of losses. Punishment was most common when participants experienced losses, especially when losses resulted in disadvantageous inequity. However, punishment also occurred as a result of disadvantageous inequity, even in the absence of losses. These findings provide support for both revenge and inequity aversion as motivations for punishment and shed light on why we punish others. At the functional level, our results provide evidence for both the deterrence and fitness-leveling theories of punishment. While deterrence is a widely accepted function of punishment, recent work suggests that the role of punishment as a tool to create a competitive advantage has been overlooked in the literature (Raihani & Bshary, 2019). Our study supports the idea that punishment also serves a competitive function, allowing individuals to reduce fitness differentials between themselves and others. Our findings shed light on the relative importance of the different motivations that underly punishment. Understanding the ultimate and proximate mechanisms that underlie punishment is important because punishment is critical for maintaining cooperation in large scale societies (Fehr & Fischbacher, 2004). A better

understanding of punishment can help us strengthen and maintain the high levels of cooperation essential to the success of complex societies.

1.5. Acknowledgements

1.6. Data Availability

The data associated with this research are available at [https://osf.io/xz3sc/].

1.7. Open Practices

This project was preregistered prior to data collection [http://aspredicted.org/blind.php?x=md765v]. All of our data and code are available in an online repository here [https://osf.io/xz3sc/].

1.8. References

Abbink, K., & Sadrieh, A. (2009). The pleasure of being nasty. *Economics Letters*, *105*(3), 306 308.

Alger, I., Weibull, J. W., & Lehmann, L. (2020). Evolution of preferences in structured populations: genes, guns, and culture. *Journal of Economic Theory*, *185*, 104951.

Amir, O., & Rand, D. G. (2012). Economic games on the internet: The effect of $1 stakes. *PloS One*, *7*(2), e31461.

Balliet, D., & Van Lange, P. A. (2013). Trust, punishment, and cooperation across 18 societies: A meta-analysis. *Perspectives on Psychological Science*, *8*(4), 363-379.

Bone, J. E., McAuliffe, K., & Raihani, N. J. (2016). Exploring the motivations for punishment: framing and country-level effects. *PloS One*, *11*(8), e0159769.

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.

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, *13*(3), 171-195.

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, *328*(5978), 617-620.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3-5.

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature*, *373*(6511), 209.

Delton, A. W., & Krasnow, M. M. (2017). The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior*, *38*(6), 734-743.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature*, *452*(7185), 348.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63-87.

Fehr, E., & Gachter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980-994.

Fischbacher, U., Gächter, S., & Quercia, S. (2012). The behavioral validity of the strategy method in public good experiments. *Journal of Economic Psychology, 33*(4), 897-913.

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature, 422*(6928), 137.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization, 3*(4), 367-388.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J.C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2006). Costly punishment across human societies. *Science, 312*(5781), 1767-1770.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362-1367.

Houser, D., & Xiao, E. (2010). Inequality-seeking punishment. *Economics Letters*, *109*(1), 20 23.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14(3), 399-425.

Jordan, J., McAuliffe, K., & Rand, D. (2016). The effects of endowment size and strategy method on third party punishment. *Experimental Economics, 19*(4), 741-763.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*(2), 433-442.

Marczyk, J. (2017). Human punishment is not primarily motivated by inequality. *PloS One, 12*(2), e0171298.

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences*, *36*(1), 1-15.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics*, *92*(1-2), 91-112.

Price, M. E., Cosmides, L., & Tooby, J. (2002). Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior*, *23*(3), 203-231.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raihani, N. J., & Bell, V. (2018). Conflict and cooperation in paranoia: a large-scale behavioural experiment. *Psychological Medicine*, *48*(9), 1523-1531.

Raihani, N.J., & Bshary, R. (2019). Punishment: One Tool, Many Uses. *Evolutionary Human Sciences*, 1..

Raihani, N. J., & McAuliffe, K. (2012). Human punishment is motivated by inequity aversion, not a desire for reciprocity. *Biology Letters*, *8*(5), 802-804.

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*, 172-179.

Sylwester, K., Herrmann, B., & Bryson, J. J. (2013). Homo homini lupus? Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics*, *6*(3), 167.

Wu, J. J., Zhang, B. Y., Zhou, Z. X., He, Q. Q., Zheng, X. D., Cressman, R., & Tao, Y. (2009). Costly punishment does not always increase cooperation. *Proceedings of the National Academy of Sciences*, *106*(41), 17448-17451.

Xiao, E. (2013). Profit-seeking punishment corrupts norm obedience. *Games and Economic Behavior*, *77*(1), 321-344.

Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., ... & 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.