Why do we punish? A multinomial analysis of the mechanisms underlying moral punishment

Inaugural-Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Ana Isabel Philippsen

aus Düsseldorf

Düsseldorf, September 2024

aus dem Institut für Experimentelle Psychologie

der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung

der Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

Berichterstatter

1. Prof. Dr. Raoul Bell

2. Prof. Dr. Axel Buchner

Tag der mündlichen Prüfung: 13.11.24

Table of content

Abstract

The human species is unique in their capacity for large-scale cooperation. What is particularly remarkable is the extent to which cooperators are willing to sacrifice own resources to punish others for non-cooperation, referred to as *moral punishment*. While this behavior can be easily explained when interacting with partners repeatedly, it is more challenging to explain why people consistently engage in costly moral punishment even in one-shot interactions where they cannot benefit from forcing others to cooperate. In light of the ubiquity of this behavior as well as the crucial role of such moral punishment in sustaining cooperation, understanding what motivates people to morally punish in one-shot interactions is an important but yet unanswered question. In the present dissertation, this gap in research was addressed in seven experiments using a simultaneous one-shot Prisoner's Dilemma game. The established multinomial cooperation-and-punishment model was used to obtain an adequate estimate of the probability of moral punishment and to thereby test different mechanisms proposed to underlie the inclination to morally punish, namely emotion communication, conformity and deliberation. The findings revealed that a) communicating one's emotions after an interaction could partly substitute costly moral punishment, indicating that moral punishment to some extent serves to communicate an emotional evaluation to the partner; b) people do not punish to enforce conformity but, instead, primarily punish defection; and c) moral punishment occurs deliberately rather than intuitively, in line with the hypothesis that deliberating fairness helps overcome selfish, profit-oriented impulses. In sum, while the experiments shed light on different mechanisms underlying moral punishment, they jointly reveal a remarkably robust preference to morally punish defection across a variety of contexts.

Introduction

Cooperation, that is, to be bearing own costs for the benefit of others and the collective good, has been observed in a variety of animal species. Bees, for example, cooperate in building honeycombs and meerkats guard each others' offspring when the others go hunting. Humans, however, are special in that they cooperate extensively even among non-kin and in one-shot interactions where there are no future benefits expected through direct or indirect reciprocity (Bowles & Gintis, 2004; Clutton-Brock, 2009). Researchers agree that this capacity for large-scale cooperation has crucially fostered human evolution and the establishment of modern societies (Boyd & Richerson, 2005; Henrich et al., 2003; Hill, 2002; Nowak, 2006; Tomasello et al., 2005). Still, as cooperation implies costs, many situations raise an incentive to shy away from these costs and to free ride on the cooperation maintained by others. If too many people free ride, cooperation continuously loses its appeal and impends to collapse (Andreoni, 1995; Boyd et al., 2003; Boyd & Richerson, 1992; Fehr & Fischbacher, 2004). This clash of individual and collective interests is called a social dilemma (cf. Kollock, 1998). Given that many current global challenges resemble a social dilemma in which an individual incentive to free ride (e.g., by not investing in environmentally friendly products) clashes with the collective demand for cooperation (to preserve our planet), it is crucial to understand which mechanisms foster and maintain cooperation.

One mechanism that helps sustain cooperation is punishment of defection which removes the incentive of defection and thereby effectively enforces cooperation (Axelrod, 1986; Boyd & Richerson, 1992; Hua & Liu, 2023; Ostrom et al., 1992; Yamagishi, 1986). Similar to cooperation, punishment entails costs to the punisher in the form of time, money or the risk of retaliation. The costly punishment of defectors by cooperators is therefore often considered a moral act, thus referred to as *moral punishment* (cf. Kurzban et al., 2007; Mieth et al., 2021a, 2021b). Despite its costs, it is

well established that people consistently engage in costly moral punishment of defection even in one-shot interactions where they cannot personally benefit from coercing others to cooperate because they cannot reasonably expect to interact with the person again. This phenomenon is not confined to laboratory experiments (e.g., Barclay, 2006; Carpenter & Matthews, 2012; Falk et al., 2005; Fehr & Gächter, 2000, 2002; Henrich et al., 2006; Walker & Halloran, 2004) but can also be observed in the field (Artavia-Mora et al., 2016; Carpenter, 2004; Price, 2005) and in everyday social interactions. For example, people often invest a lot of time to write negative online reviews after transactions they feel were unfair even though they will never interact with the reviewed person again. In light of this ubiquity and the important role of moral punishment in sustaining cooperation, it is crucial to investigate what drives this puzzling yet socially tangible behavior.

The aim of the present dissertation was therefore to test three different hypotheses regarding the mechanisms behind moral punishment. First, in Experiments 1.1 and 1.2 the hypothesis was tested that moral punishment serves to privately express or to communicate one's emotions after an unfair interaction. Second, in Experiments 2.1 and 2.2 it was tested whether punishment is directed primarily at punishing defection or merely directed at punishing deviations from the majority behavior. Lastly, in Experiments 3.1, 3.2 and 3.3 it was tested whether moral punishment occurs intuitively or relies on deliberation.

In all experiments, cooperation and punishment behavior was investigated in a simultaneous one-shot Prisoner's Dilemma game with a costly punishment option. In this game, participants interact with a different partner every round. In each round, they are asked whether they want to cooperate or to defect. They are informed beforehand that their partner simultaneously makes the same decision as they do and that, depending on the decisions of both players, the game results in different monetary outcomes, as summarized in the payoff matrix of the game (see Figure 1). The payoff matrix is construed in such a way that, at a collective level, it is

always more profitable to cooperate since mutual cooperation leads to a better outcome than mutual defection. However, at an individual level, there is an incentive to defect at the other's expense since unilateral defection yields the highest possible outcome. The matrix thereby incorporates a typical social dilemma in which the collective and the individual interests clash (Kollock, 1998). After participants have received feedback on their own and their partner's decision in the interaction along with the corresponding consequences for their account balances, they may choose to use some of their own endowment to deduct tenfold the amount from their partner's account balance as punishment.

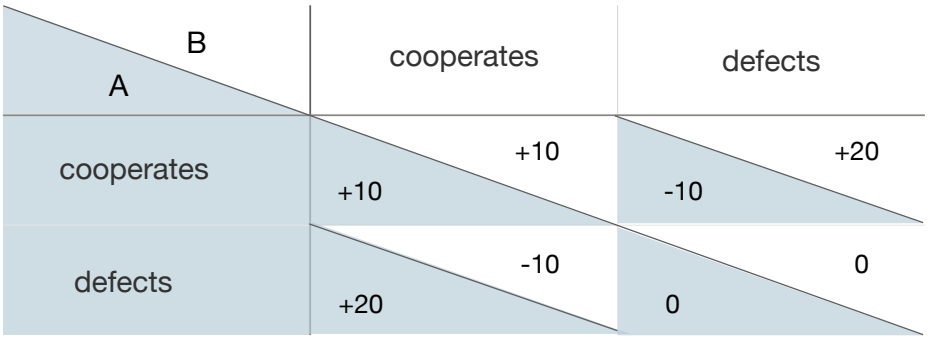

Figure 1. Payoff matrix of the Prisoner's Dilemma game. Shaded cells indicate the decision of and payoff to Player A.White cells indicate the decision of and payoff to Player B. Mutual cooperation yields the best outcome for both players collectively but, individually, defecting on a cooperating player resembles the best possible outcome, thereby capturing a typical social dilemma (Kollock, 1998).

When investigating the mechanisms behind punishment, it is important to acknowledge that moral punishment of defection is not the only type of punishment that might occur in an interaction. Punishment in social dilemma games is not only performed by cooperators but, sometimes, also by defectors and may be directed at both defectors and cooperators (e.g., Carpenter, 2007; Cinyabuguma et al., 2006; Falk et al., 2005; Gächter & Herrmann, 2009; Henrich et al., 2006; Herrmann et al., 2008; Nikiforakis, 2008; Sylwester et al., 2013). In order to clearly distinguish between different types of punishment, the multinomial cooperation-and-punishment model was used (cf. Mieth et al., 2021a, 2021b). Multinomial models help disambiguate

observable data by delineating the sequence of underlying, otherwise unobservable, latent cognitive processes contributing to a certain extent to the observed behavior (for a review see Batchelder & Riefer, 1999; Erdfelder et al., 2009). The probabilities with which these latent processes occur are estimated with easy-to-use computer programs such as *multitree* (Moshagen, 2010). Based on the observed cooperation and punishment behavior, the multinomial cooperation-and-punishment model allows to clearly differentiate between different types of punishment that may occur in an interaction. It further allows to dissect cooperation from moral punishment. In doing so, it provides a clear advantage over common behavioral measures used in other studies in which both components are entangled in only a single measure although, if separately measured, these components are essentially unrelated (Boyd et al., 2003; Mischkowski et al., 2018; Peysakhovich et al., 2014; Weber et al., 2018; Yamagishi et al., 2012). Another advantage of the model is that it yields an estimate of a *punishment bias*, that means an unspecific tendency to punish irrespective of the outcome of the game, which is distinguished from types of punishment that are contingent on the specific outcome of the game. This approach is parallel to how response bias is taken into account in other decision-making models (Batchelder & Riefer, 1990; Bayen et al., 1996; Buchner et al., 1995; Erdfelder et al., 2007; Menne et al., 2022).

The model is illustrated in Figure 2. The following description of the model is largely based on the model description in Philippsen et al. (2024b). The model incorporates two trees, one for each type of partner in the Prisoner's Dilemma game (defector or cooperator), as indicated by the rectangles on the left. Rectangles on the right describe the participant's observable responses in the game (cooperation or defection; punishment or no punishment). The letters along the branches of the trees denote the parameters of the model which represent the latent processes underlying the observable responses in the game. Parameter *C* describes a participant's probability to cooperate which is assumed to be independent of the individual partner's behavior since this is revealed only after the participant's own cooperation decision.

Therefore, the same parameter C can be used for both trees. The P. parameters along both trees reflect the conditional probabilities of different types of punishment that may be applied, depending on the participant's decision and the type of partner they encounter. To illustrate, if a participant decides to cooperate with probability C but encounters a defecting partner, they may apply *moral punishment* with probability P_{Moral}. Even if the participant does not apply moral punishment, which occurs with probability $1 - P_{\text{Moral}}$, they may still punish the partner due to an unspecific punishment bias with probability b. With the complementary probability $1 - b$, no punishment is applied in this case. If a participant decides to defect with probability $1 - C$ and encounters a defecting partner, they may apply *hypocritical punishment* with probability $P_{\text{Hypocritical}}$. Faithful to its name, this type of punishment may be motivated by the hypocritical intent to enforce a cooperation norm the participant themselves fail to follow (cf. Mieth et al., 2021b). If no hypocritical punishment is applied, which occurs with probability $1 - P_{\text{Hypocritical}}$, the participant may still punish the partner due to the unspecific punishment bias with probability b. With probability $1 - b$, no punishment is applied in this case. In turn, if a defecting participant encounters a cooperating partner, they may apply *antisocial punishment* with probability $P_{\text{Antisocial}}$. This type of punishment is considered antisocial in that it reflects an act of opposition towards cooperation. Even if no antisocial punishment is applied, which occurs with probability $1 - P_{\text{Antisocial}}$, the participant may still punish the partner due to the unspecific punishment bias with probability b. With probability $1 - b$, no punishment is applied in this case. Lastly, if a cooperating participant encounters a cooperating partner, there is no specific reason to punish the partner. Any punishment in this case is therefore assumed to be caused only by the unspecific punishment bias with probability b. With probability $1 - b$, no punishment is applied in this case.

By applying this model and thereby obtaining precise estimates of the probabilities of different types of punishment in a variety of different conditions, specific

mechanisms previously proposed to underlie moral punishment could be tested. This enabled a more profound insight into what actually drives moral punishment in one-shot interactions than analyzing overt behavior which must necessarily remain ambiguous because the same observed punishment can be brought about by different types of processes.

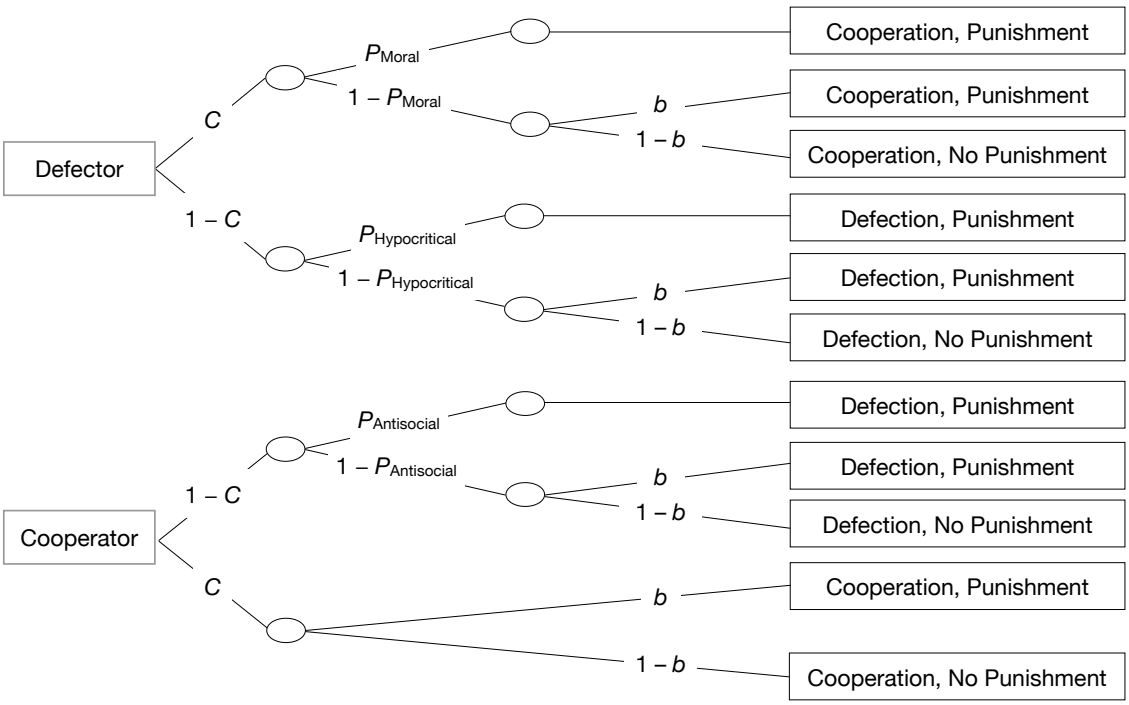

Figure 2. The multinomial cooperation-and-punishment model. Rectangles on the left indicate the types of partners that may be encountered in the Prisoner's Dilemma game (defector or cooperator); rectangles on the right describe the participant's observable responses in the game (cooperation or defection; punishment or no punishment). The letters along the branches of the trees reflect the parameters of the model, namely parameter *C* for cooperation, the *P*. parameters for the different types of punishment (P_{Moral}, $P_{Hypocritical}$ and $P_{Antisocial}$) and parameter *b* for the punishment bias.

The role of emotion

One of the first mechanisms proposed to underlie punishment of defection was anger caused by the perceived unfairness (Fehr & Gächter, 2002). This assumption is corroborated by a large number of studies indicating that the application of such punishment coincides with self-reported anger (Ambrus & Greiner, 2012; Bosman &

Van Winden, 2002; Dickinson & Masclet, 2015; Hopfensitz & Reuben, 2009; Pillutla & Chen, 1999; Seip et al., 2014) as well as physiological indices of anger (Ben-Shakhar et al., 2007; Civai et al., 2010; Gummerum et al., 2020; Joffily et al., 2014; Sanfey et al., 2003; Van't Wout et al., 2006). Seip et al. (2014) further delineated that anger is an essential prerequisite for an unfair interaction to trigger punishment. This is supported by other correlational studies reporting that anger mediates the relation between unfairness and punishment (Gummerum et al., 2020; Pillutla & Murnighan, 1996; Wang et al., 2009).

Overall, it is well established that anger evokes punishment. It is, however, less understood how expressing anger, in turn, affects punishment. Some studies indicate that a cathartic relief of anger, through a cooling-off period (Bolle et al., 2014; Dickinson & Masclet, 2015) or emotional ratings (Dickinson & Masclet, 2015), diminishes punishment of unfairness. In a different approach, Xiao and Houser (2005) demonstrated that offering players ways to communicate to their partner how they felt about the previous interaction other than costly punishment led to a decrease in rejection rates in an Ultimatum Game. In the Ultimatum Game, one player is endowed with a certain amount of money and asked to propose an offer on how to share the endowment to the other player. The other player may then choose to accept the offer, leading to the share being paid out accordingly, or to reject the offer, leading to both players receiving nothing. Rejecting an unfair offer in the Ultimatum Game may be interpreted as costly moral punishment because it implies relinquishing a personal gain to prevent an unfair distribution (Bolton & Zwick, 1995; Fehr & Fischbacher, 2003). Xiao and Houser's findings may therefore suggest that expressing one's emotions after an unfair interaction causes a cathartic relief of anger and thereby a reduction in moral punishment. Alternatively, it could be assumed that it was the *communication* of emotions which served as a valid alternative for moral punishment. If so, then it can be deduced that one function of moral punishment may be to communicate an emotional evaluation to the interaction partner. This

interpretation is reasonable given the crucial role of moral punishment in preserving cooperation and the multitude of literature linking punishment to the wish of making the offender understand their wrongdoing (e.g., Crockett et al., 2014; Gollwitzer & Denzler, 2009; Gollwitzer et al., 2011; Molnar et al., 2020) and of achieving a change within the punished person (Aharoni et al., 2022; Funk et al., 2014). The first set of experiments therefore served to test whether one function of moral punishment is indeed to communicate an emotional evaluation, possibly to deter from future free-riding.

Experiment 1.1: No expression versus communication of emotions

The aim of the first experiment was to conceptually replicate the findings of Xiao and Houser (2005) using a less ambiguous measure of moral punishment than observable behavior in the Ultimatum Game. While rejection rates in the Ultimatum Game are commonly interpreted as moral punishment, this interpretation has been challenged because it conflates moral punishment with cooperation, both of which are represented in only one decision (Mieth et al., 2021a; Yamagishi et al., 2012). A paradigm that enables to clearly dissect these two decisions is the Prisoner's Dilemma game with a costly punishment option, employed in all experiments of the present dissertation. In order to test whether offering participants a way to communicate how they feel about the previous interaction to their partner can serve as an alternative to moral punishment, participants were randomly assigned to one of two conditions: In the emotion-communication condition, participants were asked to rate their emotions along the Self-Assessment Manikin scales (Bradley & Lang, 1994) after each round. They had been informed beforehand that these ratings would be communicated to their partners. They then received the costly punishment option. In the no-emotion-expression condition, participants were not asked to rate their emotions prior to making their punishment decision. A detailed description of the manipulation as well as the results of the analyses performed in Experiments 1.1 and

1.2 can be found in Published Article 1, attached to this dissertation (Philippsen et al., 2023).

Providing participants with an alternative way to express and communicate their emotions in response to the interaction significantly reduced moral punishment compared to the no-emotion-expression condition, thereby conceptually replicating the results of Xiao and Houser (2005). As explained earlier, this may indicate that moral punishment is, at least partly, driven by the wish to communicate a negative emotional evaluation in response to being defected. Nonetheless, the reduction in moral punishment may also be explained by the *catharsis account*, suggesting that emotion expression may serve as a venting mechanism which does not necessitate a communicative component (cf. Dickinson & Masclet, 2015). The catharsis account therefore implies that privately expressing one's emotions without communicating them may already reduce anger and thereby moral punishment. Based on Experiment 1.1, in which an emotion-communication condition was compared solely to a no-emotion-expression condition, these two accounts could not be distinguished.

Experiment 1.2: Private expression versus communication of emotions

The aim of Experiment 1.2 was to further differentiate the effect reported in Experiment 1.1 and to test whether it was driven by the communication of emotions or merely by their expression. To do so and parallel to Experiment 1.1, an emotion-communication condition was contrasted to a no-emotion-expression condition. In addition, the emotion-communication condition was also contrasted to a private-emotion-expression condition in which participants were asked to express their emotions after each round but were informed that this information was kept private. If the reduction in moral punishment reported in Experiment 1.1 was merely due to a cathartic relief of anger caused by expressing one's emotions, then moral punishment should be diminished in both the emotion-communication and the private-emotion-expression condition compared to the no-emotion-expression condition. If, however,

a communicative function underlies the effect, then moral punishment should be diminished exclusively in the emotion-communication but not in the private-emotion-expression condition.

In line with the communication account, moral punishment was significantly reduced when emotions were communicated to the partner but not when they were only privately expressed. This indicates that the decline in moral punishment relative to the no-emotion-expression condition was not due to a cathartic relief enabled by the mere expression of emotions. Instead, the opportunity to communicate an emotional evaluation to the partner served as a valid alternative to moral punishment. Nonetheless, although slightly reduced in the emotion-communication condition, the probability of moral punishment remained above 0.60 in all three conditions.

As an additional result, both communicating and privately expressing one's emotions prior to making the punishment decision increased hypocritical punishment and decreased cooperation; privately expressing emotions even more so. Privately expressing emotions also increased the unspecific punishment bias, indicating that it led participants to punish their partners more randomly. These side findings, although not initially predicted, complement the main finding by further challenging the catharsis account of punishment (cf. Dickinson & Masclet, 2015). According to this account, expressing one's emotions and thereby venting one's frustration should have diminished the urge to punish the partner. An increase in hypocritical as well as random punishment, thus, cannot be reconciled with the catharsis account.

Discussion

In search of the mechanisms underlying the puzzling yet robust human behavior of sacrificing own resources to punish defecting strangers, anger is one of the most frequently mentioned candidates. Unfairness consistently causes anger which in turn

triggers the urge to punish the person causing that unfairness (Gummerum et al., 2020; Seip et al., 2014). Xiao and Houser (2005) argue that, in social dilemma games, punishment often constitutes participants' only resort to satisfy an intrinsic need to express their anger about the previous interaction. Accordingly, they found that offering participants other ways to communicate their emotions reduced rejection rates in an Ultimatum Game. This suggests that a communicative function underlies moral punishment. In two consecutive experiments, this effect could be conceptually replicated with the Prisoner's Dilemma game which allows for a less ambiguous measure of moral punishment than the Ultimatum game. Further, it could be demonstrated that it is not the mere expression of emotions and a resulting cathartic relief thereof that causes the reduction in moral punishment. Instead, it is the opportunity to communicate one's anger to the partner which partly serves as an alternative to moral punishment. Given the crucial role of moral punishment in maintaining cooperation (e.g., Boyd et al., 2003), it is sensible that this communicative component serves to signal socially inadequate behavior and to deter from pursuing this behavior, thereby enforcing cooperation. This is in line with literature emphasizing communication as a motive underlying punishment (cf. Funk et al., 2014).

Nonetheless, it is crucial to note that both experiments reported here consistently show that the attenuating effect of the communication of emotions on moral punishment is only small. A high level of moral punishment was observed even in the emotion-communication condition. This indicates that communicating emotions can partly substitute moral punishment but cannot replace it completely.

The role of conformity

Another mechanism that might explain why people engage in costly punishment is conformity. People strive for conformity in a number of different situations as it helps

to reduce ambiguity concerning the appropriate behavior in a given situation (Boyd & Richerson, 1988; Morgan et al., 2012). Instead of punishing specifically to discourage defection, as implied by a purely *moral account of punishment* (cf. Bone et al., 2014), people might actually punish the deviation from majority behavior, irrespective of its effect on cooperation, in order to enforce conformity (Carpenter & Matthews, 2012). Since cooperation is the majority behavior in most social settings, both a strictly moral account of punishment and a *conformity account of punishment* might explain why punishment of defection is the most prevalent form of punishment. A conformity account of punishment, however, has the unique advantage that it offers a reasonable explanation for why people sometimes antisocially punish cooperation which, by contrast, cannot be reconciled with a purely moral account of punishment (Horne & Irwin, 2016; Irwin & Horne, 2013).

Experiments 2.1 and 2.2

In two consecutive experiments, it was tested whether people primarily punish to discourage defection, as implied by an inherently moral account of punishment, or merely punish deviations from majority behavior. To do so, the cooperation base rates of the interaction partners were manipulated and participants were informed about the typical behavior in the game—that is cooperation or defection—before starting the game. In the cooperating-majority condition, partners were programmed to cooperate in 60 % and to defect in the other 40 % of the trials. In the defecting-majority condition, this ratio was reversed. A detailed description of the manipulation can be found in Published Article 2, attached to this dissertation (Philippsen et al., 2024a). The two experiments were identical in procedure with the exception that in Experiment 2.1, defecting participants were always morally punished by their partners whereas in Experiment 2.2, partners did not morally punish. The aim was to thereby test whether the results can be replicated when

participants cannot follow a punishment norm established by their interaction partners.

If punishment is, in fact, directed at discouraging defection, then moral punishment of defection should be the most prevalent form of punishment, irrespective of the manipulated base rates. If, on the other hand, punishment is driven by the motive to enforce conformity, then moral punishment of defection should be high when the majority of partners cooperates but low when the majority defects. Antisocial punishment of cooperation, in turn, should be high when the majority of partners defects and low when the majority cooperates. In fact, if punishment were exclusively driven by the motive to enforce conformity, then the probability of moral punishment of defection when the majority cooperates should be equal to the probability of antisocial punishment of cooperation when the majority defects because it would be the mere deviation from majority behavior not the type of behavior itself—specifically, the partner's unilateral defection—that drives punishment.

Results and Discussion

The results of the analyses performed in Experiment 2.1 and 2.2 are described in detail in Published Article 2, attached to this dissertation (Philippsen et al., 2024a). In line with the conformity account, in both experiments moral punishment of defection was higher when the majority cooperated. However, moral punishment of defection when the majority cooperated was much more likely than antisocial punishment of cooperation when the majority defected which should not be the case if punishment were exclusively driven by the motive to enforce conformity. Even more strikingly inconsistent with the conformity account is the fact that antisocial punishment of cooperation was *higher* when the majority cooperated than when the majority defected, directly refuting the assumption that people punish what is *uncommon*. Instead, it can be argued that people antisocially punish "moral do-gooders" for

raising a cooperation norm that they themselves disapprove of but might feel embarrassed for not adhering to (cf. Herrmann et al., 2008). This opposition towards cooperation is enhanced by a stronger normative pressure to cooperate and a stronger embarrassment caused by the higher prevalence of cooperators. Such an interpretation of antisocial punishment conforms to the results of studies from the field of do-gooder derogation, indicating more do-gooder derogation the more people belong to the alleged morally superior group (Loughnan & Piazza, 2018; Minson & Monin, 2012).

Further, it stands out that moral punishment, despite being sensitive to the manipulated cooperation rates, was remarkably high when the majority defected. This remained the case even when partners did not morally punish in Experiment 2.2 which indicates that the high levels of moral punishment were not the result of participants following a punishment norm established by their interactions partners. Overall, it can be concluded that people do not punish to blindly enforce conformity. Punishment is used primarily, yet not solely, in a moral fashion to discourage defection. Defectors may sometimes engage in antisocial punishment to oppose those raising a cooperation norm; some cooperators may slightly adjust their punishment behavior to the perceived strength of the cooperation norm but, overall, the effect of conformity on punishment seems rather weak.

The role of deliberation

According to the moral preference hypothesis, costly punishment of defection is driven by an internalized preference to do the right moral thing (Capraro & Perc, 2021). The four previously illustrated experiments support this hypothesis by indicating that moral punishment remains at a high level even in the presence of other cost-efficient ways to communicate one's emotions in response to the defection (Experiments 1.1 and 1.2) or when defection entails conforming to the majority

behavior (Experiments 2.1 and 2.2). A question that remains, however, is whether moral punishment is driven by people's intuition, as implied by an intuitive-morality account (Zaki & Mitchell, 2013), or whether it actually requires deliberation to show this economically irrational yet socially valuable behavior, as suggested by a deliberate-morality account (DeWall et al., 2008).

In order to stimulate intuitive behavior, researchers typically restrict cognitive resources through time pressure or a distractor task; in order to stimulate deliberate behavior, they impose a time delay (for a review see Capraro, 2024). In doing so, some researchers found that punishment was increased with restricted resources (Anderson & Dickinson, 2010; Balafoutas & Jaber-Lopez, 2018; Cappelletti et al., 2011; Halali et al., 2014; Liu et al., 2015; Sutter et al., 2003) and decreased when punishment decisions were delayed (Grimm & Mengel, 2011; Neo et al., 2013; Smith & Silberberg, 2010; Wang et al., 2011). These findings suggest that punishment occurs intuitively, thereby supporting an *intuitive-morality account of punishment*. However, these results are challenged by studies in which either no effect of manipulating the intuitive or deliberate processing mode was found (Achtziger et al., 2018; Artavia-Mora et al., 2017; Bosman et al., 2001; Cappelletti et al., 2011; Fraser & Nettle, 2020; Oechssler et al., 2015) or the exact opposite pattern was reported with decreased punishment when resources were restricted and increased punishment when decisions were delayed (Achtziger et al., 2016; Ferguson et al., 2014; Hochman et al., 2015). The latter findings suggest that punishment requires time to deliberate, possibly to overcome the selfish impulse of refraining from the costs of punishment and to instead act in accordance with moral goals. Such a *deliberate-morality account of punishment* is also underpinned by neuroimaging studies in which the application of punishment coincided with activation in areas of cognitive control (Knoch et al., 2008; Knoch et al., 2006).

Overall, the findings on the processing mode underlying moral punishment remain inconsistent. Further, they again stem from the Ultimatum Game which, as explained

earlier, only provides an ambiguous measure of moral punishment. In a previous study, Mieth et al. (2021a) therefore used the same paradigm and model as the experiments of the present dissertation to pursue the question of whether moral punishment relies on intuition or deliberation. They found that restricting cognitive resources by the means of a distractor task decreased moral punishment, providing support for a deliberate-morality account of punishment.

Experiments 3.1 and 3.2: Testing the deliberate-morality account of punishment

The aim of Experiment 3.1 was to conceptually replicate the findings by Mieth et al. (2021a) using time pressure instead of cognitive load to suppress deliberation (cf. Rand, 2016). Thus, punishment decisions had to be made within a five-second interval in the condition with time pressure. These decisions were then contrasted to the standard condition without time pressure. The aim of Experiment 3.2 was to test whether stimulating deliberation, in turn, has the reversed effect. In the condition with deliberation, punishment decisions were delayed by a 30-second pause in which participants were encouraged to deliberate their decision. This condition was again contrasted to the standard condition without deliberation in which punishment decisions were not delayed and participants were not specifically encouraged to deliberate their decision. A detailed description of the manipulations of the two experiments can be found in Published Article 3, attached to this dissertation (Philippsen et al., 2024b).

If moral punishment relies on intuition, as implied by an intuitive-morality account, then moral punishment should be increased with time pressure (Experiment 3.1) but decreased with deliberation (Experiment 3.2) compared to the corresponding standard control condition. If, on the other hand, moral punishment is facilitated through deliberation, as predicted by the deliberate-morality account, the opposite

pattern should be found: Moral punishment should be decreased with time pressure but increased when encouraged to deliberate one's punishment decisions.

The results of the analyses performed in Experiment 3.1 and 3.2 are described in detail in Published Article 3 (Philippsen et al., 2024b). Both experiments provided evidence in favor of the deliberate-morality account of punishment: Moral punishment was decreased with time pressure (Experiment 3.1) but increased when deliberation was encouraged (Experiment 3.2). The results of Experiments 3.1 and 3.2 therefore replicate and extend the findings of Mieth et al. (2021a) by showing that not only different methods to stimulate intuitive behavior decreases moral punishment but that, in turn, stimulating deliberate behavior increases moral punishment. While these results are consistent with some of the previously mentioned studies (e.g., Ferguson et al., 2014; Hochman et al., 2015), they are inconsistent with other studies in which the opposite pattern was found, thereby favoring an intuitive-morality account of punishment (e.g., Neo et al., 2013; Sutter et al., 2003). Such discrepancies might be partly explained by differences in the applied method and analyses (cf. Capraro, 2024). They might further allude to specific context factors modulating the influence of intuition and deliberation on punishment. In fact, in light of the vast inconsistencies on the effect of intuition and deliberation on moral behaviors, some researchers have suggested to refrain from a dichotomous classification of behaviors as either strictly intuitive or strictly deliberate behaviors and to consider them as contingent upon the context instead (Declerck & Boone, 2015; Isler et al., 2021). In that sense, instructing participants to deliberate their decision in Experiment 3.2 may have initiated them to deliberate the fairness of their decision by default since they are accustomed to considering fairness principles from everyday social interactions. Such unguided deliberation of fairness then increased moral punishment. However, according to the *process-based account* proposed by Declerck and Boone (2015), the effect of deliberation may depend on the type of deliberation that takes place. The aim of Experiment 3.3 was therefore to move beyond the effect of unguided

deliberation on punishment, and test whether this effect can be manipulated by varying the specific content of deliberation.

Experiment 3.3: Specific manipulation of the content of deliberation

The procedure was similar to the condition with deliberation of Experiment 3.2 with the exception that participants were encouraged to specifically deliberate either their self-interests or fairness during the time delay (for a detailed description of the manipulation as well as the analyses, see Philippsen et al., 2024b). The results revealed that moral punishment was significantly higher when participants deliberated fairness compared to when they deliberated their self-interests which demonstrates that the effect of deliberation does, in fact, depend on what is specifically deliberated. Similar to the previous experiments, it stands out that even in the condition which discouraged moral punishment—in this case, the condition in which participants deliberated self-interests—participants still morally punished defection with a probability of more than 0.60 although this entailed sacrificing own costs and thereby clearly went against their economic self-interest. This, yet again, demonstrates the persistence of moral punishment and thereby supports the moral preference hypothesis (Capraro & Perc, 2021). In contrast to moral punishment, antisocial punishment was significantly higher when participants deliberated self-interests compared to when they deliberated fairness which underscores that it serves to antisocially oppose the very fairness norms that moral punishment serves to uphold.

Discussion

It is well understood that moral punishment of defection helps enforce cooperation and is thereby integral to the functioning of human societies. What is less understood, however, is whether this behavior results from people's intuitive predisposition, as implied by an intuitive-morality account (Zaki & Mitchell, 2013),

or whether time and deliberation are required to show this economically irrational, yet socially valuable behavior, as predicted by a deliberate-morality account (DeWall et al., 2008). In two consecutive experiments, the results are more compatible with a deliberate-morality account: In Experiment 3.1, time pressure decreased moral punishment, suggesting that intuition favors selfish rather than moral behaviors in social dilemma games. In Experiment 3.2, time delay and the instruction to deliberate the decision, in turn, increased moral punishment. Nonetheless, this effect of deliberation could be modulated by manipulating the specific content of deliberation. When participants were encouraged to deliberate their self-interests in Experiment 3.3, moral punishment was lower than when they were encouraged to deliberate fairness. This supports the notion that deliberation does not always favor or inhibit moral behaviors as this effect depends on the context in which the decision is made (cf. Declerck & Boone, 2015). In that reasoning, Declerck and Boone intended to reconcile the prevailing inconsistencies regarding the effect of intuition and deliberation on punishment by proposing a more nuanced process-based account which incorporates context-specific modulations. Such an account is supported by evidence showing that aspects like the fairness of the offer (Ferguson et al., 2014; Halali et al., 2014), the length of the time delay (Oechssler et al., 2015), the specific game in question (Oechssler et al., 2015), the partner's social status (Harris et al., 2020) or group membership (Yudkin et al., 2016) modulate the effect of intuition and deliberation on punishment. With Experiment 3.3, these findings are extended by demonstrating that manipulating what is specifically deliberated during a time delay also affects moral punishment.

Despite being slightly reduced with time pressure or when encouraged to deliberate self-interests, moral punishment consistently remained at a high level. This corresponds to all of the previous experiments already demonstrating the robustness of moral punishment against a number of context-specific modulations. In line with the moral preference hypothesis (Capraro & Perc, 2021), the present findings again

show that moral punishment is persistent in many contexts, partly reduced but far from completely eliminated by time pressure and the instruction to deliberate self-interests.

General discussion

Human cooperation among anonymous strangers is exceptional within the animal kingdom. One mechanism to uphold such large-scale cooperation is punishment of defection (e.g., Boyd & Richerson, 1992; Hua & Liu, 2023). Over the years, it has frequently been demonstrated that people readily sacrifice own resources to punish those who do not cooperate (e.g., Barclay, 2006; Fehr & Gächter, 2000, 2002). Strangely, they do so even in one-shot interactions in which they do not benefit from forcing someone to cooperate since they cannot expect to interact with the person again. This raises the question: What drives people to sacrifice own resources to punish defectors in situations where there are no prospective benefits from doing so?

The aim of the present dissertation was therefore to find the mechanisms underlying people's puzzling yet robust inclination to sacrifice costs in order to morally punish defection in one-shot interactions. While previous studies have already identified a number of candidates, the findings remain inconsistent and are based on ambiguous measures which do not allow to distinguish moral punishment from cooperation or from an unspecific punishment bias. In the experiments presented in this dissertation, multinomial modeling was used to directly measure moral punishment. This allowed to more adequately test a number of proposed mechanisms and thereby further differentiate what specifically underlies moral punishment.

In a first step, it was demonstrated that an alternative way to communicate emotions can partly substitute moral punishment, thereby conceptually replicating the findings by Xiao and Houser (2005). Further, it could be delineated that it is not a cathartic relief of anger caused by the mere expression of emotions (as suggested by

Dickinson & Masclet, 2015) that underlies the effect. Instead, the findings suggest that moral punishment, at least partly, serves to communicate an emotional evaluation to the defecting partner (cf. Aharoni et al., 2022; Crockett et al., 2014; Funk et al., 2014; Nahmias & Aharoni, 2017).

In a second step, it could be demonstrated that people primarily punish defection and not merely deviations from majority behavior (as suggested by Carpenter & Matthews, 2012; Irwin & Horne, 2013; Li et al., 2021). While participants did adjust their punishment to the majority behavior, this effect was fairly small. Further, the effect on antisocial punishment was in the direction opposite to what a conformity account of punishment would have predicted, thereby clearly refuting the assumption that people punish what is uncommon.

Lastly, it was demonstrated that moral punishment is not the derivative of an intuitive predisposition but requires time and the possibility to deliberate fairness principles to overcome selfish, profit-maximizing impulses. Moral punishment was decreased with time pressure and, in turn, increased when deliberation was encouraged. Instructing participants to deliberate their self-interests (compared to deliberating fairness) reduced moral punishment, indicating that the effect of deliberation on punishment depends on the specific content that is deliberated.

Although different mechanisms were investigated within the three series of experiments presented in this dissertation, a similarity across all experiments is that in every condition in which the moral use of punishment was discouraged, moral punishment was remarkably persistent. Taken together, this demonstrates that moral punishment is driven by a strong internal preference robust against other more cost-efficient ways to communicate one's emotions, against conformity concerns, against time pressure or against the instruction to consider one's self-interests in the matter. These findings provide evidence supporting the moral preference hypothesis, proposed by Capraro and Perc (2021). According to this hypothesis, the persistence of

a variety of prosocial behaviors in different social dilemmas can be best explained by a "generalized moral preference" motivating them to act in accordance with what is perceived as morally right or wrong in that situation (Capraro & Perc, 2021; Capraro & Rand, 2018). Following this assumption, moral punishment may be driven by a strong internal preference to do the right moral thing, in this case punish the decision to defect rather than to cooperate, causing people to forgo economic self-interest in order to adhere to this preference.

A general moral preference to punish defection cannot, however, exhaustively explain people's punishment behavior. While the present experiments jointly suggest that such a preference underlies moral punishment, these experiments also show that participants consistently concede costs to engage in hypocritical, antisocial as well as unspecific punishment. These punishment types exhibited specific patterns of effects which highlights the necessity of clearly distinguishing between these types of punishment when trying to understand the mechanisms underlying costly punishment.

Hypocritical punishment was partly influenced in the same manner as moral punishment (Experiments 2.1, 2.2 and 3.2; cf. also Mieth et al., 2021b), suggesting that this form of punishment may be—faithful to its name—motivated by hypocritical outrage about the other's defection (cf. Mieth et al., 2021b) or by the attempt to feign support for a cooperation norm, as suggested by Experiments 2.1 and 2.2 (cf. Willer et al., 2009). However, hypocritical punishment also exhibited effects different to those relating to moral punishment (Experiments 1.1, 1.2, 3.1 and 3.3; cf. also Mieth et al., 2021a), suggesting that hypocritical punishment also relies on mechanisms distinct to the ones underlying moral punishment. Antisocial punishment seems to be in direct opposition to moral punishment, driven by a motivation to oppose the very cooperation norm that moral punishment serves to uphold. This opposition was increased when participants exhibited a stronger normative pressure to cooperate through a higher frequency of cooperating partners (Experiments 2.1 and 2.2) or

partners morally punishing defection (Mieth et al., 2016; Mieth et al., 2017, 2021b), possibly causing a stronger motivation to oppose that very cooperation norm. The instruction to deliberate self-interests (Experiment 3.3), highlighting the utility of defecting rather than cooperating, in turn caused a stronger urge to oppose partners who were not acting in line with their self-interests but sticking to the morally superior option of cooperating—in line with research assuming antisocial punishment to be driven by an aversion to moral do-gooders (cf. Herrmann et al., 2008). Finally, participants' unspecific punishment bias was increased when emotions were made salient (Experiment 1.2) or when cognitive resources were restricted (Experiment 3.1; cf. also Mieth et al., 2021a), suggesting that impairing deliberate processing increases people's unspecific bias to punish.

Even though these other forms of punishment have consistently been observed in the present as well as preceding experiments (Mieth et al., 2016; Mieth et al., 2017, 2021a, 2021b), moral punishment remains by far the most common type of punishment, in line with previous literature (e.g., Barclay, 2006; Falk et al., 2005; Fehr & Gächter, 2002). Looking at previous literature, it stands out that it is also the most commonly investigated form of punishment. Understanding what drives people to show this puzzling behavior is not only important from a theoretical but also from a practical point of view. Especially in modern times in which social interactions are more and more shifted to online one-shot interactions with strangers, investigating the mechanisms that underlie moral punishment and thereby contribute to sustaining cooperation in such one-shot interactions becomes increasingly important. The fact that we repeatedly find evidence for a strong internal preference to morally punish defection that is robust against various context modulations appears promising in that regard.

Still, in light of the present and previous indications that antisocial punishment directly undermines cooperation and thereby counteracts the effect of moral punishment, it seems crucial that future research does not focus solely on moral

punishment. Gaining a deeper understanding of what motivates people to antisocially punish those who contribute to the collective good and what effect this may have on people's willingness to cooperate as well as to morally punish defection may provide a valuable perspective when trying to understanding how prosocial behavior can be sustained.

Conclusion

In conclusion, all experiments of the present dissertation find compelling evidence that even in one-shot interactions, precluding the effects of strategic prosociality, participants do not act in accordance with their monetary self-interests. Instead, they were reliably willing to sacrifice own money to punish others at remarkable rates and they do so primarily, yet not exclusively, to morally punish defection. While moral punishment is slightly reduced when there are other, more cost-efficient ways to communicate one's emotions to the partner, when cooperation is the minority behavior, when decisions are made under time pressure and when they are instructed to deliberate self-interests in the matter, it is overall robust against manipulations discouraging the moral use of punishment. This indicates that people bear a strong, internal preference to morally punish defection even in situations where there are no prospective benefits from doing so. In light of current global crises, this appears promising since moral punishment plays a pivotal role in sustaining cooperation.

References

Achtziger, A., Alós-Ferrer, C., & Wagner, A. K. (2016). The impact of self-control depletion on social preferences in the ultimatum game. *Journal of Economic Psychology*, *53*, 1-16. https://doi.org/10.1016/j.joep.2015.12.005

Achtziger, A., Alós-Ferrer, C., & Wagner, A. K. (2018). Social preferences and self-control. *Journal of Behavioral and Experimental Economics*, *74*, 161-166. https://doi.org/10.1016/j.socec.2018.04.009

Aharoni, E., Simpson, D., Nahmias, E., & Gollwitzer, M. (2022). A painful message: Testing the effects of suffering and understanding on punishment judgments. *Zeitschrift für Psychologie*, *230*(2), 138-151. https://doi.org/10.1027/2151-2604/a000460

Ambrus, A., & Greiner, B. (2012). Imperfect public monitoring with costly punishment: An experimental study. *American Economic Review*, *102*(7), 3317-3332. https://doi.org/10.1257/aer.102.7.3317

Anderson, C., & Dickinson, D. L. (2010). Bargaining and trust: the effects of 36-h total sleep deprivation on socially interactive decisions. *Journal of Sleep Research*, *19*(1-Part-I), 54-63. https://doi.org/10.1111/j.1365-2869.2009.00767.x

Andreoni, J. (1995). Cooperation in public-goods experiments: Kindness or confusion? *American Economic Review*, *85*(4), 891-904. https://www.jstor.org/stable/2118238

Artavia-Mora, L., Bedi, A. S., & Rieger, M. (2016). Intuitive cooperation and punishment in the field. *IZA Discussion Paper No. 9871*. https://doi.org/10.2139/ssrn.2769179

Artavia-Mora, L., Bedi, A. S., & Rieger, M. (2017). Intuitive help and punishment in the field. *European Economic Review*, *92*, 133-145. https://doi.org/10.1016/j.euroecorev.2016.12.007

Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review*, *80*(4), 1095-1111. https://doi.org/10.2307/1960858

Balafoutas, L., & Jaber-Lopez, T. (2018). Impunity under pressure: On the role of emotions as a commitment device. *Economics Letters*, *168*, 112-114. https://doi.org/10.1016/j.econlet.2018.04.027

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

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, *97*(4), 548–564. https://doi.org/10.1037/0033-295X.97.4.548

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*(1), 57-86. https://doi.org/10.3758/BF03210812

Bayen, U. J., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 197-215. https://doi.org/10.1037/0278-7393.22.1.197

Ben-Shakhar, G., Bornstein, G., Hopfensitz, A., & Van Winden, F. (2007). Reciprocity and emotions in bargaining using physiological and self-report measures. *Journal of Economic Psychology*, *28*(3), 314-323. https://doi.org/10.1016/j.joep.2007.02.005

Bolle, F., Tan, J. H., & Zizzo, D. J. (2014). Vendettas. *American Economic Journal: Microeconomics*, *6*(2), 93-130. https://doi.org/10.1257/mic.6.2.93

Bolton, G. E., & Zwick, R. (1995). Anonymity versus punishment in ultimatum bargaining. *Games and Economic Behavior*, *10*(1), 95-121. https://doi.org/10.1006/game.1995.1026

Bone, J., Silva, A. S., & Raihani, N. J. (2014). Defectors, not norm violators, are punished by third-parties. *Biology Letters*, *10*(7), Article: 20140388. https://doi.org/10.1098/rsbl.2014.0388

Bosman, R., Sonnemans, J., & Zeelenberg, M. (2001). *Emotions, rejections, and cooling off in the ultimatum game* [Unpublished manuscript]. CREED/Department of Economics, University of Amsterdam. https://pure.uva.nl/ws/files/2263560/31_coolingoff.pdf

Bosman, R., & Van Winden, F. (2002). Emotional hazard in a power-to-take experiment. *The Economic Journal*, *112*(476), 147-169. https://doi.org/10.1111/1468-0297.0j677

Bowles, S., & Gintis, H. (2004). The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology*, *65*(1), 17-28. https://doi.org/10.1016/j.tpb.2003.07.001

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, *100*(6), 3531-3535. https://doi.org/10.1073/pnas.0630443100

Boyd, R., & Richerson, P. J. (1988). An evolutionary model of social learning: the effects of spatial and temporal variation. In R. Boyd & P. J. Richerson (Eds.), *Social learning* (pp. 29-48). Psychology Press.

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, *13*(3), 171-195. https://doi.org/10.1016/0162-3095(92)90032-Y

Boyd, R., & Richerson, P. J. (2005). *The origin and evolution of cultures* (1st ed.). Oxford University Press.

Bradley, M. M., & Lang, A.-G. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49-59. https://doi.org/10.1016/0005-7916(94)90063-9

Buchner, A., Erdfelder, E., & Vaterrodt-Plünnecke, B. (1995). Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *Journal of Experimental Psychology: General*, *124*(2), 137-160. https://doi.org/10.1037/0096-3445.124.2.137

Cappelletti, D., Güth, W., & Ploner, M. (2011). Being of two minds: Ultimatum offers under cognitive constraints. *Journal of Economic Psychology*, *32*(6), 940-950. https://doi.org/10.1016/j.joep.2011.08.001

Capraro, V. (2024). The dual-process approach to human sociality: Meta-analytic evidence for a theory of internalized heuristics for self-preservation. *Journal of Personality and Social Psychology*, *126*(5), 719–757. https://doi.org/10.1037/pspa0000375

Capraro, V., & Perc, M. (2021). Mathematical foundations of moral preferences. *Journal of the Royal Society Interface*, *18*(175), e20200880. https://doi.org/10.1098/rsif.2020.0880

Capraro, V., & Rand, D. G. (2018). Do the right thing: Experimental evidence that preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. *Judgment and Decision Making*, *13*(1), 99-111. https://doi.org/10.2139/ssrn.2965067

Carpenter, J. P. (2004). When in Rome: conformity and the provision of public goods. *The Journal of Socio-Economics*, *33*(4), 395-408. https://doi.org/10.1016/j.socec.2004.04.009

Carpenter, J. P. (2007). The demand for punishment. *Journal of Economic Behavior & Organization*, *62*(4), 522-542. https://doi.org/10.1016/j.jebo.2005.05.004

Carpenter, J. P., & Matthews, P. H. (2012). Norm enforcement: anger, indignation, or reciprocity? *Journal of the European Economic Association*, *10*(3), 555-572. https://doi.org/10.1111/j.1542-4774.2011.01059.x

Cinyabuguma, M., Page, T., & Putterman, L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics*, *9*(3), 265-279. https://doi.org/10.1007/s10683-006-9127-z

Civai, C., Corradi-Dell'Acqua, C., Gamer, M., & Rumiati, R. I. (2010). Are irrational reactions to unfairness truly emotionally-driven? Dissociated behavioural and emotional responses in the Ultimatum Game task. *Cognition*, *114*(1), 89-95. https://doi.org/10.1016/j.cognition.2009.09.001

Clutton-Brock, T. (2009). Cooperation between non-kin in animal societies. *Nature*, *462*(7269), 51-57. https://doi.org/10.1038/nature08366

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. https://doi.org/10.1037/xge0000018

Declerck, C., & Boone, C. (2015). *Neuroeconomics of prosocial behavior: The compassionate egoist*. Academic Press.

DeWall, C. N., Baumeister, R. F., Gailliot, M. T., & Maner, J. K. (2008). Depletion makes the heart grow less helpful: Helping as a function of self-regulatory energy and genetic relatedness. *Personality and Social Psychology Bulletin*, *34*(12), 1653-1662. https://doi.org/10.1177/0146167208323981

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

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie*, *217*(3), 108-124. https://doi.org/10.1027/0044-3409.217.3.108

Erdfelder, E., Cüpper, L., Auer, T.-S., & Undorf, M. (2007). The four-states model of memory retrieval experiences. *Zeitschrift für Psychologie*, *215*(1), 61-71. https://doi.org/10.1027/0044-3409.215.1.61

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, *73*(6), 2017-2030. https://doi.org/10.1111/j.1468-0262.2005.00644.x

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785-791. https://doi.org/10.1038/nature02043

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185-190. https://doi.org/10.1016/j.tics.2004.02.007

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980-994. https://doi.org/10.1257/aer.90.4.980

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

Ferguson, E., Maltby, J., Bibby, P. A., & Lawrence, C. (2014). Fast to forgive, slow to retaliate: Intuitive responses in the ultimatum game depend on the degree of unfairness. *Plos One*, *9*(5), e96344. https://doi.org/10.1371/journal.pone.0096344

Fraser, S., & Nettle, D. (2020). Hunger affects social decisions in a multi-round Public Goods Game but not a single-shot Ultimatum Game. *Adaptive Human Behavior and Physiology*, *6*, 334-355. https://doi.org/10.1007/s40750-020-00143-3

Funk, F., McGeer, V., & Gollwitzer, M. (2014). Get the message: Punishment is satisfying if the transgressor responds to its communicative intent. *Personality and Social Psychology Bulletin*, *40*(8), 986-997. https://doi.org/10.1177/0146167214533130

Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *364*(1518), 791-806. https://doi.org/10.1098/rstb.2008.0275

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

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

Grimm, V., & Mengel, F. (2011). Let me sleep on it: Delay reduces rejection rates in ultimatum games. *Economics Letters*, *111*(2), 113-115. https://doi.org/10.1016/j.econlet.2011.01.025

Gummerum, M., López-Pérez, B., Van Dijk, E., & Van Dillen, L. F. (2020). When punishment is emotion-driven: Children's, adolescents', and adults' costly punishment of unfair allocations. *Social Development*, *29*(1), 126-142. https://doi.org/10.1111/sode.12387

Halali, E., Bereby-Meyer, Y., & Meiran, N. (2014). Between self-interest and reciprocity: the social bright side of self-control failure. *Journal of Experimental Psychology: General*, *143*(2), 745. https://doi.org/10.1037/a0033824

Harris, A., Young, A., Hughson, L., Green, D., Doan, S. N., Hughson, E., & Reed, C. L. (2020). Perceived relative social status and cognitive load influence acceptance of unfair offers in the Ultimatum Game. *Plos One*, *15*(1), e0227717. https://doi.org/10.1371/journal.pone.0227717

Henrich, J., Bowles, S., Boyd, R. T., Hopfensitz, A., Richerson, P. J., Sigmund, K., Smith, E. A., Weissing, F. J., & Young, H. P. (2003). Group report: The cultural and genetic evolution of human cooperation. In P. Hammerstein (Ed.), *Genetic and Cultural Evolution of Cooperation* (pp. 445–468). Dahlem Workshop Reports. https://doi.org/10.7551/mitpress/3232.003.0025

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., & Henrich, N. (2006). Costly punishment across human societies. *Science*, *312*(5781), 1767-1770. https://doi.org/10.1126/science.1127333

Herrmann, B., Thoni, C., & Gachter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362-1367. https://doi.org/10.1126/science.1153808

Hill, K. (2002). Altruistic cooperation during foraging by the Ache, and the evolved human predisposition to cooperate. *Human Nature*, *13*, 105-128. https://doi.org/10.1007/s12110-002-1016-3

Hochman, G., Ayal, S., & Ariely, D. (2015). Fairness requires deliberation: The primacy of economic over social considerations. *Frontiers in Psychology*, *6*, Article: 747. https://doi.org/10.3389/fpsyg.2015.00747

Hopfensitz, A., & Reuben, E. (2009). The importance of emotions for the effectiveness of social punishment. *The Economic Journal*, *119*(540), 1534-1559. https://doi.org/10.1111/j.1468-0297.2009.02288.x

Horne, C., & Irwin, K. (2016). Metanorms and antisocial punishment. *Social Influence*, *11*(1), 7-21. https://doi.org/10.1080/15534510.2015.1132255

Hua, S., & Liu, L. (2023). Facilitating the evolution of cooperation through altruistic punishment with adaptive feedback. *Chaos, Solitons & Fractals*, *173*, Article: 113669. https://doi.org/10.1016/j.chaos.2023.113669

Irwin, K., & Horne, C. (2013). A normative explanation of antisocial punishment. *Social Science Research*, *42*(2), 562-570. https://doi.org/10.1016/j.ssresearch.2012.10.004

Isler, O., Gächter, S., Maule, A. J., & Starmer, C. (2021). Contextualised strong reciprocity explains selfless cooperation despite selfish intuitions and weak social heuristics. *Scientific Reports*, *11*(1), Article: 13868. https://doi.org/10.1038/s41598-021-93412-4.

Joffily, M., Masclet, D., Noussair, C. N., & Villeval, M. C. (2014). Emotions, sanctions, and cooperation. *Southern Economic Journal*, *80*(4), 1002-1027. https://doi.org/10.4284/0038-4038-2012.067

Knoch, D., Nitsche, M. A., Fischbacher, U., Eisenegger, C., Pascual-Leone, A., & Fehr, E. (2008). Studying the neurobiology of social interaction with transcranial direct current stimulation—the example of punishing unfairness. *Cerebral Cortex*, *18*(9), 1987-1990. https://doi.org/10.1093/cercor/bhm237

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, *314*(5800), 829-832. https://doi.org/10.1126/science.1129156

Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology*, *24*, 183-214. https://www.jstor.org/stable/223479

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

Li, X., Molleman, L., & van Dolder, D. (2021). Do descriptive social norms drive peer punishment? Conditional punishment strategies and their impact on cooperation. *Evolution and Human Behavior*, *42*(5), 469-479. https://doi.org/10.1016/j.evolhumbehav.2021.04.002

Liu, Y., He, N., & Dou, K. (2015). Ego-depletion promotes altruistic punishment. *Open Journal of Social Sciences*, *3*(11), 62–69. https://doi.org/10.4236/jss.2015.311009.

Loughnan, S., & Piazza, J. (2018). Thinking morally about animals. In K. Gray & J. Graham (Eds.), *Atlas of moral psychology* (pp. 165-174).

Menne, N. M., Winter, K., Bell, R., & Buchner, A. (2022). A validation of the two-high threshold eyewitness identification model by reanalyzing published data. *Scientific Reports*, *12*(1), Article: 13379. https://doi.org/10.1038/s41598-022-17400-y

Mieth, L., Bell, R., & Buchner, A. (2016). Facial likability and smiling enhance cooperation, but have no direct effect on moralistic punishment. *Journal of Experimental Psychology*, *63*(5), 263-277. https://doi.org/10.1027/1618-3169/a000338

Mieth, L., Buchner, A., & Bell, R. (2017). Effects of gender on costly punishment. *Journal of Behavioral Decision Making*, *30*(4), 899-912. https://doi.org/10.1002/bdm.2012

Mieth, L., Buchner, A., & Bell, R. (2021a). Cognitive load decreases cooperation and moral punishment in a Prisoner's Dilemma game with punishment option. *Scientific Reports*, *11*(1), 1-12. https://doi.org/10.1038/s41598-021-04217-4

Mieth, L., Buchner, A., & Bell, R. (2021b). Moral labels increase cooperation and costly punishment in a Prisoner's Dilemma game with punishment option. *Scientific Reports*, *11*(1), 1-13. https://doi.org/10.1038/s41598-021-89675-6

Minson, J. A., & Monin, B. (2012). Do-gooder derogation: Disparaging morally motivated minorities to defuse anticipated reproach. *Social Psychological and Personality Science*, *3*(2), 200-207. https://doi.org/10.1177/1948550611415695

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

Molnar, A., Chaudhry, S., & Loewenstein, G. (2020). *'It's not about the money. It's about sending a message!': Unpacking the components of revenge* [Working Paper No. 8102]. Center for Economic Studies.

Morgan, T. J., Rendell, L. E., Ehn, M., Hoppitt, W., & Laland, K. N. (2012). The evolutionary basis of human social learning. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1729), 653-662. https://doi.org/10.1098/rspb.2011.1172

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, *42*(1), 42-54. https://doi.org/10.3758/BRM.42.1.42

Nahmias, E., & Aharoni, E. (2017). Communicative theories of punishment and the impact of apology. In C. Suprenant (Ed.), *Rethinking punishment in the era of mass incarceration* (pp. 144-161). Routledge.

Neo, W. S., Yu, M., Weber, R. A., & Gonzalez, C. (2013). The effects of time delay in reciprocity games. *Journal of Economic Psychology*, *34*, 20-35. https://doi.org/10.1016/j.joep.2012.11.001

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. https://doi.org/10.1016/j.jpubeco.2007.04.008

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560-1563. https://doi.org/10.1126/science.1133755

Oechssler, J., Roider, A., & Schmitz, P. W. (2015). Cooling off in negotiations: Does it work? *Journal of Institutional and Theoretical Economics*, 565-588. https://doi.org/10.1628/093245615X14307212950056

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *American Political Science Review*, *86*(2), 404-417. https://doi.org/10.2307/1964229

Peysakhovich, A., Nowak, M. A., & Rand, D. G. (2014). Humans display a 'cooperative phenotype'cthat is domain general and temporally stable. *Nature Communications*, *5*(1), Article: 4939. https://doi.org/10.1038/ncomms5939

Philippsen, A., Mieth, L., Buchner, A., & Bell, R. (2023). Communicating emotions, but not expressing them privately, reduces moral punishment in a Prisoner's Dilemma game. *Scientific Reports*, *13*(1), Article: 14693. https://doi.org/10.1038/s41598-023-41886-9

Philippsen, A., Mieth, L., Buchner, A., & Bell, R. (2024a). People punish defection, not failures to conform to the majority. *Scientific Reports*, *14*(1), Article: 1211. https://doi.org/10.1038/s41598-023-50414-8

Philippsen, A., Mieth, L., Buchner, A., & Bell, R. (2024b). Time pressure and deliberation affect moral punishment. *Scientific Reports*, *14*(1), Article: 16378. https://doi.org/10.1038/s41598-024-67268-3

Pillutla, M. M., & Chen, X.-P. (1999). Social norms and cooperation in social dilemmas: The effects of context and feedback. *Organizational Behavior and Human Decision Processes*, *78*(2), 81-103. https://doi.org/10.1006/obhd.1999.2825

Pillutla, M. M., & Murnighan, J. K. (1996). Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational Behavior and Human Decision Processes*, *68*(3), 208-224. https://doi.org/10.1006/obhd.1996.0100

Price, M. E. (2005). Punitive sentiment among the Shuar and in industrialized societies: Cross-cultural similarities. *Evolution and Human Behavior*, *26*(3), 279-287. https://doi.org/10.1016/j.evolhumbehav.2004.08.009

Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science*, *27*(9), 1192-1206. https://doi.org/10.1177/0956797616654455

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, *300*(5626), 1755-1758. https://doi.org/10.1126/science.1082976

Seip, E. C., Van Dijk, W. W., & Rotteveel, M. (2014). Anger motivates costly punishment of unfair behavior. *Motivation and Emotion*, *38*(4), 578-588. https://doi.org/10.1007/s11031-014-9395-4

Smith, P., & Silberberg, A. (2010). Rational maximizing by humans (Homo sapiens) in an ultimatum game. *Animal Cognition*, *13*, 671-677. https://doi.org/10.1007/s10071-010-0310-4

Sutter, M., Kocher, M., & Strauß, S. (2003). Bargaining under time pressure in an experimental ultimatum game. *Economics Letters*, *81*(3), 341-347. https://doi.org/10.1016/S0165-1765(03)00215-5

Sylwester, K., Herrmann, B., & Bryson, J. J. (2013). Homo homini lupus? Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics*, *6*(3), 167-188. https://doi.org/10.1037/npe0000009

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). In search of the uniquely human. *Behavioral and Brain Sciences*, *28*(5), 721-735. https://doi.org/10.1017/S0140525X05540123

Van't Wout, M., Kahn, R. S., Sanfey, A. G., & Aleman, A. (2006). Affective state and decision-making in the ultimatum game. *Experimental Brain Research*, *169*(4), 564-568. https://doi.org/10.1007/s00221-006-0346-5

Walker, J. M., & Halloran, M. A. (2004). Rewards and sanctions and the provision of public goods in one-shot settings. *Experimental Economics*, *7*, 235-247. https://doi.org/10.1023/B:EXEC.0000040559.08652.51

Wang, C. S., Galinsky, A. D., & Murnighan, J. K. (2009). Bad drives psychological reactions, but good propels behavior: Responses to honesty and deception. *Psychological Science*, *20*(5), 634-644. https://doi.org/10.1111/j.1467-9280.2009.02344.x

Wang, C. S., Sivanathan, N., Narayanan, J., Ganegoda, D. B., Bauer, M., Bodenhausen, G. V., & Murnighan, K. (2011). Retribution and emotional regulation: The effects of time delay in angry economic interactions. *Organizational Behavior and Human Decision Processes*, *116*(1), 46-54. https://doi.org/10.1016/j.obhdp.2011.05.007

Weber, T. O., Weisel, O., & Gächter, S. (2018). Dispositional free riders do not free ride on punishment. *Nature Communications*, *9*(1), 1-9. https://doi.org/10.1038/s41467-018-04775-8

Willer, R., Kuwabara, K., & Macy, M. W. (2009). The false enforcement of unpopular norms. *American Journal of Sociology*, *115*(2), 451-490. https://doi.org/10.1086/599250

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

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, *51*(1), 110-116. https://doi.org/10.1037/0022-3514.51.1.110

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

Yudkin, D. A., Rothmund, T., Twardawski, M., Thalla, N., & Van Bavel, J. J. (2016). Reflexive intergroup bias in third-party punishment. *Journal of Experimental Psychology: General*, *145*(11), 1448. https://doi.org/10.1037/xge0000190

Zaki, J., & Mitchell, J. P. (2013). Intuitive prosociality. *Current Directions in Psychological Science*, *22*(6), 466-470. https://doi.org/10.1177/0963721413492764

Published articles

Published Article 1

The article includes Experiments 1.1 and 1.2.

Philippsen, A., Mieth, L., Buchner, A., & Bell, R. (2023). Communicating emotions, but not expressing them privately, reduces moral punishment in a Prisoner's Dilemma game. *Scientific Reports*, *13*(1), Article: 14693. https://doi.org/10.1038/s41598-023-41886-9

scientific reports

Check for updates

OPEN

Communicating emotions, but not expressing them privately, reduces moral punishment in a Prisoner's Dilemma game

Ana Philippsen✉, Laura Mieth, Axel Buchner & Raoul Bell

The existence of moral punishment, that is, the fact that cooperative people sacrifice resources to punish defecting partners requires an explanation. Potential explanations are that people punish defecting partners to privately express or to communicate their negative emotions in response to the experienced unfairness. If so, then providing participants with alternative ways to privately express or to communicate their emotions should reduce moral punishment. In two experiments, participants interacted with cooperating and defecting partners in a Prisoner's Dilemma game. After each round, participants communicated their emotions to their partners (Experiments 1 and 2) or only expressed them privately (Experiment 2). Each trial concluded with a costly punishment option. Compared to a no-expression control group, moral punishment was reduced when emotions were communicated to the defecting partner but not when emotions were privately expressed. Moral punishment may thus serve to communicate emotions to defecting partners. However, moral punishment was only reduced but far from being eliminated, suggesting that the communication of emotions does not come close to replacing moral punishment. Furthermore, prompting participants to focus on their emotions had undesirable side-effects: Privately expressing emotions diminished cooperation, enhanced hypocritical punishment (i.e., punishment of defecting partners by defecting participants), and induced an unspecific bias to punish the partners irrespective of their actions.

People have a strong tendency to punish unfair behaviors and are even willing to accept costs to do so[1–7]. As a potential explanation for this puzzling behavior, it has been suggested that moral punishment of unfair behaviors of others is driven by the negative emotional reaction to the perceived unfairness[8–12]. This explanation is supported by evidence showing that the experience of negative emotions can increase punishment rates[13–15]. Conversely, offering players of social dilemma games alternative ways to vent their frustration about an experienced unfairness can reduce punishment rates[16–18]. Moral punishment may thus serve to express or to communicate emotional evaluations of the other's behavior, hence forth referred to as *emotion communication*. If so, then providing participants with alternative ways to privately express or communicate their emotions should reduce moral punishment. Here, we test these hypotheses by examining how emotion expression and communication affects moral punishment in a Prisoner's Dilemma game.

Cooperation implies accepting costs for the benefit of others. On average, humans have a rather strong disposition to cooperate with others that prevails even when interacting with strangers in one-shot interactions, albeit the strength of this disposition varies among individuals. An overall strong propensity for cooperation is essential to the evolutionary success of human groups and cultures[19,20]. Free riders, on the other hand, may exploit the cooperation of others without reciprocating. If too many group members free-ride, cooperation declines and eventually collapses[4,19,21]. Hence, a dilemma arises: If each group member only follows their selfish interests, the outcome for the group as a whole is worse than it could have been if every group member had cooperated.

This dilemma between what is in the immediate interest of the individual and what is best for the group is captured in the Prisoner's Dilemma game[22]. Two players are endowed with a certain amount of money. They can either decide to cooperate by sacrificing part of their endowment for the collective good or they can decide to defect by refraining from cooperation at the other player's expense. The Prisoner's Dilemma is defined by its payoff structure (Fig. 1). A defecting player who benefits from a cooperating partner receives the highest payoff.

Department of Experimental Psychology, Heinrich Heine University Düsseldorf, Universitätsstrasse 1, 40225 Düsseldorf, Germany. ✉email: Ana.Philippsen@hhu.de

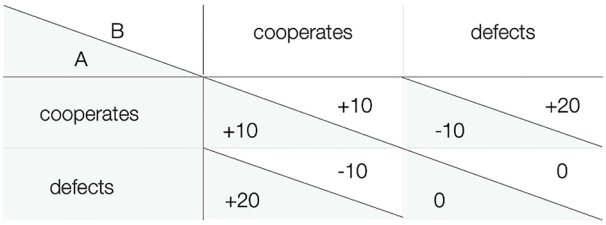

Figure 1. Payoff structure of the Prisoner's Dilemma game. Values in shaded cells indicate the payoff to Player A, values in white cells indicate the payoff to Player B. The payoffs are displayed as a function of both players' decisions in the Prisoner's Dilemma game.

Mutual cooperation leads to a better outcome for both players than mutual defection. A cooperating player who interacts with a defecting partner receives the lowest outcome. At a collective level, cooperation is desirable because mutual cooperation leads to a better outcome for both players than mutual defection. However, at an individual level, it is always more profitable to defect, irrespective of what the other player does. This payoff structure thereby captures the basic dilemma of cooperation. From this clash of individual and collective interests, the free-rider problem arises. Cooperation can only be maintained at a high level in groups and societies if the free rider problem is solved.

One solution to the free rider problem is moral punishment as it removes the incentive of defection and thereby effectively enforces cooperation[19,23,24]. The threat of moral punishment can greatly enhance cooperation rates by reducing free riding[3,25–32]. However, moral punishment incurs costs to the punisher[4,33,34]. Participants have to invest some of their own endowment to deduce points or money from the other player's account. In repeated interactions, punishers can build a reputation[35–41] and can thereby directly benefit from their punishment, forcing others to cooperate[42,43]. However, people use costly punishment even in one-shot interactions in which they cannot build a reputation or benefit in another way from forcing others to cooperate in subsequent rounds[3,5,6,44]. This raises the question: What drives people to punish free riders in one-shot interactions?

As a proximate mechanism, Fehr and Gächter[3] have proposed that moral punishment is driven by strong negative emotions. Unfair decisions have been found to cause anger in a variety of different social dilemma games[3,8,45–47]. People may use costly punishment to express their negative emotional response to the experience of unfairness. Consistent with this view, people who use costly punishment often report anger[8,14,18,31,48,49] and show evidence of emotional arousal in physiological measures[9,11,46,50–52]. Further, Seip et al.[14] showed that it is not the mere perception of unfairness per se but rather the anger in response to the perception of unfairness that triggers moral punishment. Accordingly, anger has been demonstrated to mediate punishment in several correlation-based analyses[8,11,53]. A causal role of anger in moral punishment is supported by findings showing that experimentally induced anger increases punishment rates[13–15,54].

In line with the basic tenets of classical catharsis theory[55,56], some researchers have suggested that offering participants alternative ways to relieve their anger, referred to as *venting*, may reduce costly punishment[17,18]. Dickinson and Masclet[18] demonstrated that different venting methods could significantly reduce subsequent punishment rates in the Public Good game, a multi-player variant of the Prisoner's Dilemma game. Offering participants the opportunity to rate their experienced emotions (anger, joy and surprise) during a cooling-off waiting phase diminished punishment rates even more than a cooling-off waiting period without emotional ratings but not as much as a high venting condition that included a combination of different venting strategies. These findings suggest that the opportunity to express one's current emotional state may, to some degree, decrease the need for moral punishment.

In economic games, applying costly punishment is often the only option for players to express their negative emotions. Xiao and Houser[16] therefore examined how the expression of emotions—as a potential alternative to costly punishment—affected the participants' behavior in an Ultimatum Game. In the Ultimatum Game, one player, *the proposer,* is endowed with a certain amount of money and can decide to send any proportion of this endowment, ranging from 0 to 100%, to the other player. The *responder* can then decide to accept, allowing for the monetary shares to be paid out as proposed, or to reject the offer, causing both players to receive nothing. Rejection in the Ultimatum Game can be interpreted as costly moral punishment as the responder sacrifices money to deny the proposer an unfair share[8,57,58]. When players of a one-shot Ultimatum Game were given the option to send written messages to the other player to communicate how they felt, the rejection of unfair offers declined, suggesting that the communication of emotions partly replaced moral punishment. This interpretation, however, is not straightforward because the rejection of offers in the Ultimatum Game could also be interpreted as a decline in cooperation. It thus seems interesting to examine how the communication of emotions affects moral punishment in a paradigm that allows to more clearly distinguish between cooperation and moral punishment.

In the present study, a one-shot simultaneous Prisoner's Dilemma game was combined with a costly punishment option[59–62]. To assess cooperation and different types of punishment, we used the multinomial cooperation-and-punishment model that has been successfully applied and validated in previous studies[61,62]. The model serves to distinguish among cooperation, three types of punishment (moral, hypocritical and antisocial punishment) and a punishment bias. Within the model, *moral punishment* is defined as the type of punishment that is exclusively triggered by an unfair interaction in which the participant's cooperation is exploited by the partner's defection. This type of punishment can be considered moral because it implies sacrificing resources to enforce norms of cooperation[62]. While our hypotheses mainly pertain to moral punishment, other types of punishment

occur in social dilemma games as well[1, 5, 44, 63–71]. Within the model, *hypocritical punishment* is defined as the type of punishment that is exclusively triggered by an interaction in which both partners defect. The purpose of this type of punishment is to enforce a cooperative norm the participants themselves fail to follow[62]. *Antisocial punishment* is defined as the type of punishment that is exclusively triggered by the mismatch between the participant's defection and the partner's cooperation. This type of punishment serves to oppose the normative pressure toward cooperation. Finally, people may show an unspecific bias for punishment implying that they punish their partners regardless of the outcome of the Prisoner's Dilemma game, for instance, when they are distracted from the task[61], so that a proper measurement model of punishment has to take bias into account.

The aim of the present experiments was to test how the expression and communication of emotions about unfair interactions affects moral punishment. Experiment 1 serves to test whether moral punishment is reduced when participants have an alternative route for communicating their negative emotions about unfair interactions to their partners. To this end, we manipulated between subjects whether participants could communicate their emotional response to the outcome of the Prisoner's Dilemma game. If moral punishment serves to communicate emotions, moral punishment should be reduced in a condition in which participants can communicate their negative emotions about the outcome of the Prisoner's Dilemma game to their partners in comparison to a condition in which there is no way of communicating emotions to the partner other than the act of punishment. However, a reduction of moral punishment may also be predicted based on the catharsis account, according to which the mere expression of emotions is already sufficient and communication is not necessary to cause a reduction in punishment. Experiment 2 was designed to test the catharsis account by comparing the effects of an emotion-communication condition to the effects of a private-emotion-expression condition in which emotions were only privately expressed. To anticipate, both experiments demonstrate that moral punishment is reduced—but only by a rather small amount—in the emotion-communication condition in comparison to the control condition. Privately expressing emotions without communicating them had no effect on moral punishment, favoring the emotion-communication account over the catharsis account. No specific hypotheses were derived about the non-moral types of punishment and the punishment bias. Nevertheless, it is interesting to explore whether the effects of emotion expression are closely restricted to moral punishment or whether there are potential harmful side effects of prompting participants to focus on their emotions.

Experiment 1

Method. *Sample.* We aimed at collecting at least 200 valid data sets (100 per group) and stopped data collection at the end of the week in which this criterion was reached. The final sample consisted of 203 participants (130 female, 73 male) with a mean age of 22 ($SD = 4$) who were randomly assigned to one of two conditions. One group of participants had the option to communicate their emotions about the Prisoner's Dilemma game to the partner ($n = 101$) and one group of participants did not have this option ($n = 102$). A sensitivity analysis showed that, with this sample size and 20 punishment decisions per participant, it was possible to detect small effects of $w = 0.06$ with a statistical power of $1 - \beta = 0.95$ at an α level of 0.05 when comparing the moral-punishment parameters between the two conditions[72]. The study was advertised on campus and in social media. Participants received either course credit or a small honorarium as a compensation for participation.

Materials and procedure. After consenting to the study, participants were informed that their task was to interact with different partners during a game. At the start of the experiment, they were endowed with 4 € (displayed as 400 cents) which they could invest into the game. Participants were informed that they would play for real money and that they would receive the money in their account at the end of the experiment. Participants played 26 trials of a simultaneous one-shot Prisoner's Dilemma game with a costly punishment option. The first six trials were training trials. Participants interacted with partners whose responses were determined by a computer program to ensure that half of the partners cooperated and half defected while still presenting trials in random order. The experimental manipulation of the behavior of the partners is a common approach in Experimental Psychology to gain control over confounding factors that may influence an individual's behavior. The same procedure has been used in many previous studies applying the same task[59–62, 73] and is similar to the procedures used in other studies on the psychological underpinnings of social cooperation[44, 46, 74, 75].

In each round, participants saw a silhouette on the left side of the screen representing themselves (see Fig. 2). To emphasize the social nature of the game, a color photograph (640×480 pixels) of a different partner was displayed on the right side of the screen in each round of the game. The photograph was randomly drawn from a pool of 90 female and 90 male faces of the Chicago Face Database[76], matching the participants' gender. These photographs showed the faces of young white adults. All faces had a neutral expression and were shown from a frontal view.

Participants could decide whether they wanted to cooperate or to defect. Cooperating meant to invest 30 cents in a joint business venture while defecting meant to invest nothing. To emphasize the social implications of the behavior in the Prisoner's Dilemma game (cf.[62]), the two options "I cooperate" and "I defect" were displayed as buttons above both the silhouette and the partner's photograph. Participants knew that they would make their decision at the same time as their partner and that these decisions affected their payoff.

When the participants had made their decision, the selected option was highlighted. At the same time, the partner's choice was displayed. The corresponding investments of both partners were displayed in arrows moving from each side to the center of the screen within 750 ms. With a delay of 750 ms, the sum of investments was displayed in-between the two arrows. After another 750 ms, a bonus—corresponding to one third of the sum of investments—was displayed. After 750 ms, the bonus was added to the sum of investments and the resulting total sum in the shared account was revealed. Both partners received half of the money in the shared account irrespective of their investments. After 750 ms, the two shares were represented by two arrows moving away

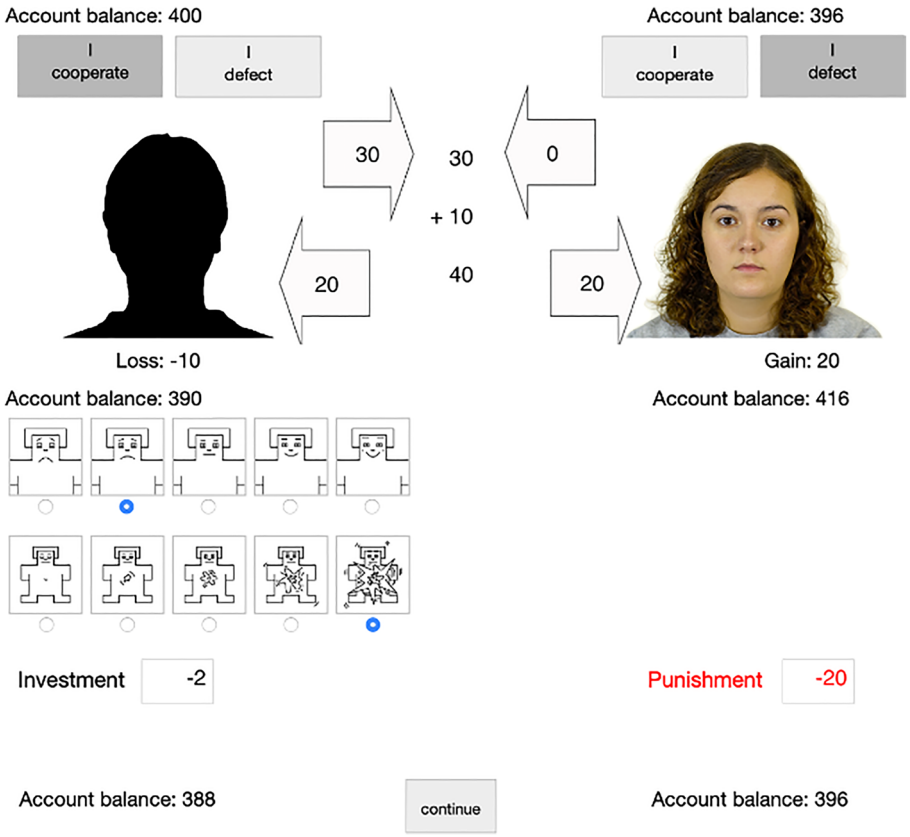

Figure 2. Example trial of the Prisoner's Dilemma game with costly punishment in Experiment 1. In this example trial, the participant in the emotion-communication condition chose to cooperate while the partner defected, resulting in a loss of 10 cents for the participant (left) and a gain of 20 cents for the partner (right). The participant then used the valence and arousal scales of the Self-Assessment Manikin[77] to communicate their emotions to the partner. Afterwards, the participant invested two cents to deduce 20 cents from the partner's account as punishment. The partner's photograph was randomly selected from the Chicago Face Database[76].

from the center to each side of the screen within 750 ms. Finally, the individual gains or losses as well as the updated account balances were displayed. The game was thus associated with the following outcomes: If both players cooperated by investing 30 cents, a bonus of 20 cents was added to the invested 60 cents so that both the participant and the partner received 40 cents from the total sum of 80 cents, resulting in a gain of 10 cents for each player. If both players defected, none of them gained or lost any money. If one of the players cooperated by investing 30 cents and the other player defected by investing nothing, a bonus of 10 cents was added to the shared account of 30 cents so that both players received 20 cents from the total sum of 40 cents. This resulted in a gain of 20 cents for the defecting player and in a loss of 10 cents for the cooperating player. The payoff structure thus corresponds to that of a typical Prisoner's Dilemma in which the collective incentive to cooperate clashes with the individual incentive to defect[22].

After each round of the Prisoner's Dilemma game, participants were provided with a costly punishment option displayed at the bottom of the screen. Participants were asked to indicate whether they wanted to punish the partner. They could invest either 0 cents if they did not want to punish their partner or 1 to 9 cents from their own account to subtract 10 to 90 cents from their partner's account. This 1:10 ratio was chosen to facilitate the use of the punishment option. The same ratio has been frequently applied in previous investigations using the multinomial cooperation-and-punishment model[59–62]. One second after participants had confirmed their punishment decision by clicking a "Punishment" button, the updated account balances were shown at the bottom of the screen. Participants could then press a "Continue" button to start the next trial.

At the end of the experiment, participants were thanked, compensated for their participation and debriefed that they had interacted with preprogrammed partners. They were reminded that they could withdraw their consent to the storage and processing of their data without having to accept any detriments but no participant did so. The amount of money paid to the participants varied between 3.50 and 6.60 € ($M = 5.21$, $SD = 0.78$). The experiment took about 12 min on average.

Communication of emotions. To examine the effect of emotion communication on punishment, participants were randomly assigned to one of two conditions. One group of participants had no opportunity to express their emotions (no emotion expression). Another group of participants was asked to communicate to their partner

how they felt about the interaction in the Prisoner's Dilemma game before they decided whether to punish the partner (emotion communication). The instructions in the emotion-communication condition explained that participants would be able to communicate their emotions to their partners:

> [You] have the opportunity to communicate to your partner how you have felt. Two rating scales are available for this purpose. Please indicate on the first scale how happy or unhappy you have felt. Please indicate on the second scale how relaxed or aroused you have felt. These scales will be displayed to your partner as soon as you confirm your response using the 'Continue' button.

The participants rated their emotions on the valence and arousal scales of the *Self-Assessment Manikin*[77]. The two non-verbal scales consist of five pictograms each, one for valence (from 1 = unhappy to 5 = happy) and one for arousal (from 1 = calm to 5 = aroused). Average valence scores were $M = 2.63$ ($SD = 0.68$) in response to defecting partners and $M = 3.82$ ($SD = 0.64$) in response to cooperating partners. Average arousal scores were $M = 2.36$ ($SD = 0.94$) in response to defecting partners and $M = 2.27$ ($SD = 0.96$) in response to cooperating partners.

The cooperation-and-punishment model. Multinomial models are useful tools to disambiguate observable categorical data by distinguishing among underlying latent processes and their contributions to observed response frequencies in terms of parameter probabilities[78–81]. Computer programs such as *multiTree*[82] have been developed to test how well a model fits the data, to estimate the model parameters and to test hypotheses directly at the level of the model parameters. Hypotheses tests are performed by introducing parameter restrictions and testing whether the restrictions lead to a significant decrease in how well the model fits the data. The multinomial model used here has already been successfully applied and validated in previous studies[61,62]. A graphical illustration of the model is shown in Fig. 3. The upper tree in Fig. 3 refers to interactions with defecting partners, the lower tree refers to interactions with cooperating partners.

The first latent process specified in each tree is that participants may choose to cooperate with probability C or to defect with the complementary probability $1 - C$. Given that the participant and the partner decide simultaneously whether to cooperate or to defect, the participant's cooperation is assumed to be independent of the partner's cooperation. Therefore, the same parameter C is used in both trees of Fig. 3. If the participant's cooperation is met with the partner's defection, the participant uses moral punishment with probability P_{Moral}. When no moral punishment is applied with probability $1 - P_{Moral}$, punishment may still occur due to an unspecific punishment bias with probability b. With probability $1 - b$, no punishment is applied. Following mutual defection, hypocritical punishment may be applied with probability $P_{Hypocritical}$. Even if no hypocritical punishment is applied with probability $1 - P_{Hypocritical}$, punishment may still occur due to the punishment bias with

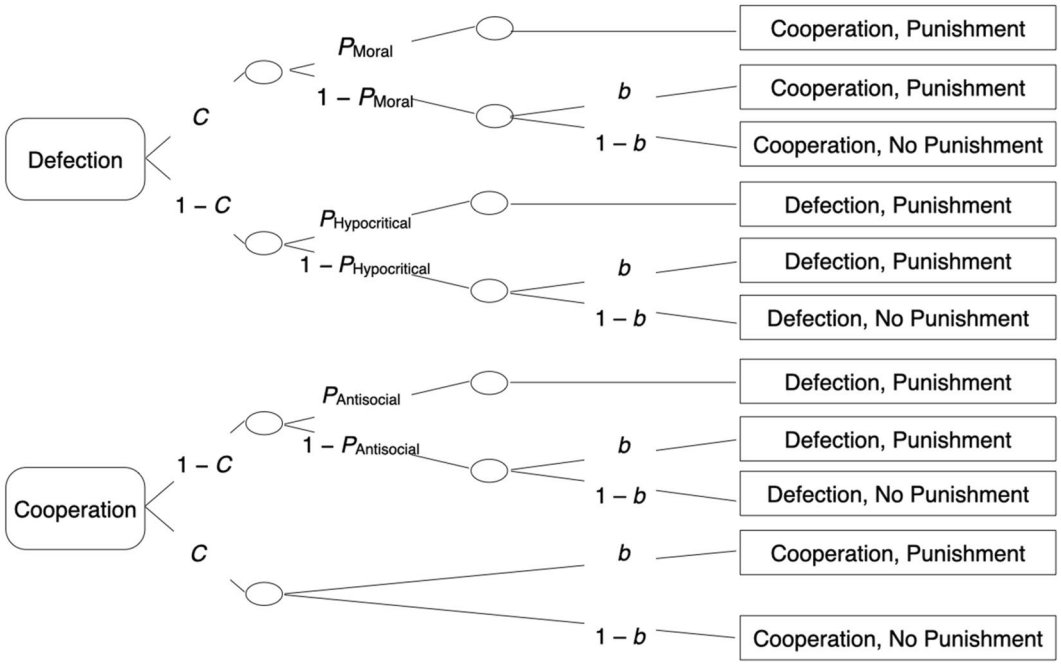

Figure 3. Multinomial cooperation-and-punishment model. Rounded rectangles on the left represent the partner behavior, rectangles on the right represent the participant behavior in a one-shot Prisoner's Dilemma game with a costly punishment option. Letters along the branches denote the parameters of the model (C = cooperative behavior in the Prisoner's Dilemma game, P_{Moral} = moral punishment after unilateral cooperation, $P_{Hypocritical}$ = hypocritical punishment after mutual defection, $P_{Antisocial}$ = antisocial punishment after unilateral defection; b = unspecific punishment bias).

probability b. With probability $1 - b$, no punishment is applied. If participant's defection is met with a partner's cooperation, antisocial punishment may be applied with probability $P_{\text{Antisocial}}$. If no antisocial punishment is applied with probability $1 - P_{\text{Antisocial}}$, punishment may still occur due to the punishment bias with probability b. With probability $1 - b$, no punishment is applied. Mutual cooperation does not provide any specific reason to punish the partner but punishment may still occur due to the punishment bias b (cf.[61,62]). With probability $1 - b$, no punishment is applied.

Results. The data were analyzed using the multinomial cooperation-and-punishment model (see Fig. 3) to assess how the communication of emotions affects cooperation and punishment. Two instances of the model are needed to analyze the results of Experiment 1, one instance for each condition (no emotion expression, emotion communication). The base model fit the data, $G^2(2) = 0.99$, $p = 0.610$. The estimates of the cooperation parameter C are shown in Fig. 4. Cooperation did not significantly differ between conditions, $\Delta G^2(1) = 0.37$, $p = 0.543$, $w = 0.01$.

Figure 5 displays the results pertaining to costly punishment. The left panel shows the estimates of the parameters representing moral, hypocritical and antisocial punishment. As predicted, moral punishment was significantly reduced when participants had the opportunity to communicate their emotions to their partners in comparison to when they were not given this opportunity, $\Delta G^2(1) = 5.11$, $p = 0.024$, $w = 0.04$. Neither hypocritical, $\Delta G^2(1) = 0.51$, $p = 0.476$, $w = 0.01$, nor antisocial punishment, $\Delta G^2(1) = 0.02$, $p = 0.881$, $w < 0.01$, differed significantly between conditions. The right panel shows the estimates of the punishment-bias parameter. At a descriptive level, the punishment bias was enhanced in the emotion-communication condition in comparison to the no-emotion-expression condition but this difference was not significant, $\Delta G^2(1) = 3.64$, $p = 0.056$, $w = 0.03$.

Discussion. The aim of Experiment 1 was to test the prediction of the emotion-communication account of moral punishment according to which moral punishment should be reduced when participants have an alternative route of communicating their emotions to their partners after each round of the Prisoner's Dilemma game. In line with this prediction, the communication of emotions significantly reduced moral punishment relative to a condition in which participants had no opportunity to express their emotions other than punishment. This finding suggests that one function of moral punishment is the communication of negative emotions in response to the partner's unilateral defection. However, moral punishment was only reduced, but far from being eliminated in the emotion-communication condition. The communication of emotions thus reduces the need for moral punishment but does not come close to replacing it. The effect of emotion communication was specific to moral punishment. Hypocritical punishment of defecting partners was not affected by communicating one's emotions prior to punishment decisions. Similarly, the communication of emotions did not affect antisocial punishment of cooperating partners. Within the limits of the sensitivity of the statistical tests used in Experiment 1, these findings suggest that the communication of emotions does not seem to be a relevant factor in hypocritical and antisocial punishment.

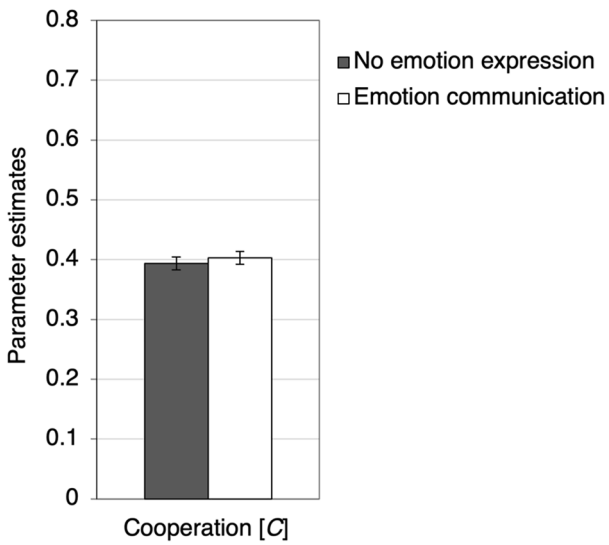

Figure 4. Estimates of the cooperation parameter C as a function of whether emotions could be communicated after each round of the Prisoner's Dilemma game (no emotion expression, emotion communication). The error bars represent standard errors.

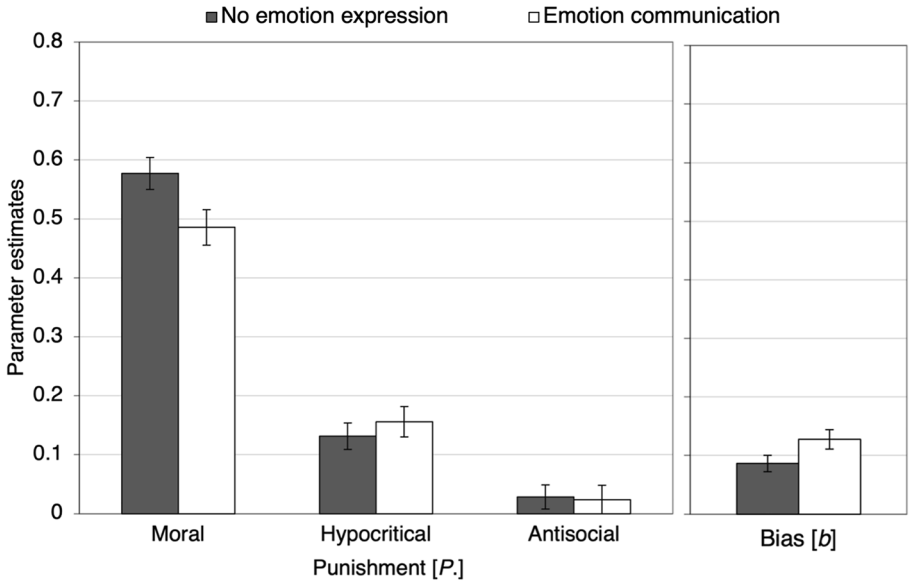

Figure 5. Estimates of the parameters representing moral, hypocritical and antisocial punishment (left panel) and punishment bias (right panel) as a function of whether participants could communicate their emotions about the Prisoner's Dilemma game before punishing the partners (no emotion expression, emotion communication). The error bars represent standard errors.

Experiment 2

Given that the effect of communicating emotions on moral punishment was relatively small in Experiment 1, an important aim of Experiment 2 was to test whether this effect can be replicated. The most important aim of Experiment 2, however, was to test whether this effect, if replicated, was due to the communication of negative emotions to the partner, as implied by the emotion-communication account, or whether it was simply due to the mere expression of negative emotions, as implied by the catharsis account. Given that, in Experiment 1, an emotion-communication condition was contrasted with a control condition in which one's emotions could not be expressed, it is impossible to distinguish between these two accounts. The results of Experiment 1 are thus not only consistent with the emotion-communication account; the results are equally compatible with the catharsis account according to which the mere expression of emotions suffices to reduce the need for moral punishment.

Both accounts have received some support in previous studies. For instance, Xiao and Houser[16] found that offering responders an opportunity to communicate their emotions to proposers in an Ultimatum Game diminished rejection rates. This finding supports the assumption that one function of moral punishment is to communicate to the proposers that their behavior was inadequate under a cooperative norm. According to this account, the component of moral punishment in question is directed at regulating *others'* behaviors. By contrast, Dickinson and Masclet[18] found a significant effect on punishment rates in a Public Goods game when participants expressed their emotions privately in written messages they knew would never be sent, supporting the assumption that a private self-centered process underlies the application of costly punishment with the intention to regulate one's *own* emotions. Experiment 2 was designed to distinguish between these contrasting accounts. To this end, the emotion-communication condition was compared to a no-emotion-expression and to a private-emotion-expression condition. If the effects of emotional responses on moral punishment are due to a self-centered venting process, moral punishment should be reduced in both the emotion-communication condition and the private-emotion-expression condition in comparison to the control condition without emotion expression. If it is the communication of emotions that underlies the effect, then moral punishment should be reduced only in the emotion-communication condition but not in the private-emotion-expression condition in comparison to the no-emotion-expression control condition. Experiment 2 also addresses another limitation of Experiment 1, namely that punishment was unilateral. Participants could punish their partners but did not receive any punishment from the partners. Previous studies[59–62] have shown that antisocial punishment rates are quite low under these conditions suggesting that the experience of being punished increases antisocial punishment which is in line with the idea that antisocial punishment serves to oppose the normative pressure toward cooperation. Consistent with these previous findings, antisocial punishment occurred only with a comparatively low probability in the present Experiment 1. It thus seemed interesting to explore whether the expression of emotions has an effect on antisocial punishment when antisocial punishment occurs with a higher probability which represents more favorable conditions for finding an effect if it exists. Also, we deliberately increased the sample size and, thus, the sensitivity of the statistical tests, in Experiment 2 relative to the sample size of Experiment 1 such that it was possible to detect even small effects on antisocial punishment, hypocritical punishment and the punishment bias. This seemed important given that, at a descriptive level, hypocritical punishment and the punishment bias were increased in the emotion-communication condition compared to the no-emotion-expression

condition in Experiment 1. As a more ecologically valid form of communicating emotions, we asked participants to communicate their emotions in the way they were used to from everyday life, that is, by using the emojis in the design determined by their computer's operating system rather than the Self-Assessment Manikin that we had used in Experiment 1.

Method. *Sample.* Given that the aim of Experiment 2 was to further dissect the effect of emotion-expression on moral punishment, we presumed the population effect size of interest to be half the size of the sample effect observed in Experiment 1 ($w = 0.04$). An a-priori power analysis in G*Power[72] showed that with an α level of 0.05 and 20 punishment decisions in the Prisoner's Dilemma game, a sample size of $n = 1625$ was necessary to detect an effect of emotion-expression on moral punishment of $w = 0.02$ with a statistical power of $1 - \beta = 0.95$. To achieve such a large sample size, the experiment was performed online. Participants were recruited via the online research panel provider *mingle*. Of those participants who had started the Prisoner's Dilemma game, 199 withdrew from the experiment, 44 data files were incomplete and 121 data files had to be removed due to double participation. The final sample consisted of 1681 participants (720 female, 957 male and 4 diverse), aged 18 to 87 years ($M = 53$, $SD = 16$).

Materials and procedure. Switching from the laboratory setting of Experiment 1 to an online format required a few adjustments to the procedure. Participants of the online panel provider *mingle* are used to being compensated with points that can be exchanged for vouchers, charity donations or bank transfers. Therefore, participants were informed that they were playing for points (1 point = 0.01 € or 1 cent) that they would receive from *mingle* after the study was completed. At the start of the game, participants were endowed with 100 points (corresponding to 1 € or 100 cents). As in Experiment 1, participants played 26 rounds of the one-shot simultaneous Prisoner's Dilemma game (Fig. 6). The first six trials were training trials. Photographs of 26 white adult faces were selected from the Chicago Face Database[76]. Half of the faces were female. The partner's behavior (cooperation, defection) was again randomly determined. The payoff structure was the same as in Experiment 1.

The procedure was similar to the procedure of Experiment 1 but was adjusted to the online environment to ensure that the information would be displayed smoothly on the participants' personal computers. Each trial started with the presentation of the participant's current account balance in the middle of the browser window. After having clicked on a "Continue" button, participants saw the partner's photograph (266 × 186 pixels) in the middle of the screen. The photograph had a blue frame (4 pixels). The photograph remained visible on screen until the end of the trial. Participants selected whether to cooperate or to defect. Upon clicking a "Continue" button, the participants received written feedback about their own decision (e.g., "You cooperate.") and their partner's decision (e.g., "Your partner defects.") and were informed about the monetary consequences of these decisions for both players (e.g., "You lose 10 points." And "Your partner gains 20 points."). Statements referring to the participant were shown in black while statements referring to the partner were shown in blue, corresponding to the blue frame around the partner's photograph.

The feedback about the outcomes of the Prisoner's Dilemma game remained visible until the end of the trial. Upon clicking a "Continue" button, participants were asked to make a punishment decision. To simplify the procedure for the online environment, a maximum of up to three points could be invested to subtract between 10 and 30 points from the partner's account. Other than in Experiment 1, punishment was not unilateral. Participants were informed that the partners had the same punishment option as the participants. The participant's and the partner's punishment decisions were displayed simultaneously. To approximate the typical behavior of real players, the partners were programmed to punish the unilateral defection of the participants by investing a random amount between 1 and 3 points to deduce between 10 and 30 points from the participant's account. Participants received immediate feedback about their own punishment decision (e.g., "You invest 2 points to punish your partner.") and its effect on the partner's account (e.g., "20 points are deducted from your partner's account.") as well as about the partner's punishment decision (e.g., "Your partner does not punish you.") and its effect on the participant's account (e.g., "No fine will be deducted from your account."). Again, statements referring to the participant were shown in black while statements referring to the partner were shown in blue.

Participants could then start the next trial by clicking a "Continue" button. On average, participants acquired a final account balance of 78 ($SD = 32$) points. The experiment took about 18 min.

Emotion expression and communication. Participants were randomly assigned to one of three groups: Participants in the *private-emotion-expression* condition ($n = 576$) privately expressed their emotional state without communicating with their partner. Participants in the *emotion-communication* condition ($n = 541$) sent a message about their emotional state to their partners. Participants in the control condition ($n = 564$) had no opportunity to express or to communicate their emotions. In the private-emotion-expression condition, participants were asked to express how they felt after each Prisoner's Dilemma interaction. They were reassured that this information would not be shared with their partners. Participants selected one of seven emojis, ordered in the same random horizontal array in each trial, and submitted their answer using a button labeled "save emotional state". After having confirmed their response, the selected emoji was displayed along with a statement confirming that the emotional state had been saved. In the emotion-communication condition, participants were instructed to communicate to their partner how they felt after the interaction. As in the private-emotion-expression condition, participants selected one of the seven emojis, ordered in the same random horizontal array in each trial, and submitted their answer using a button labeled "send message". The subsequent feedback confirmed that the message had been sent.

In order to express or to communicate their emotions, participants could choose among seven emojis expressing *anger, sadness, joy, surprise, fear, schadenfreude* and a *neutral* state. The emojis were selected based on a

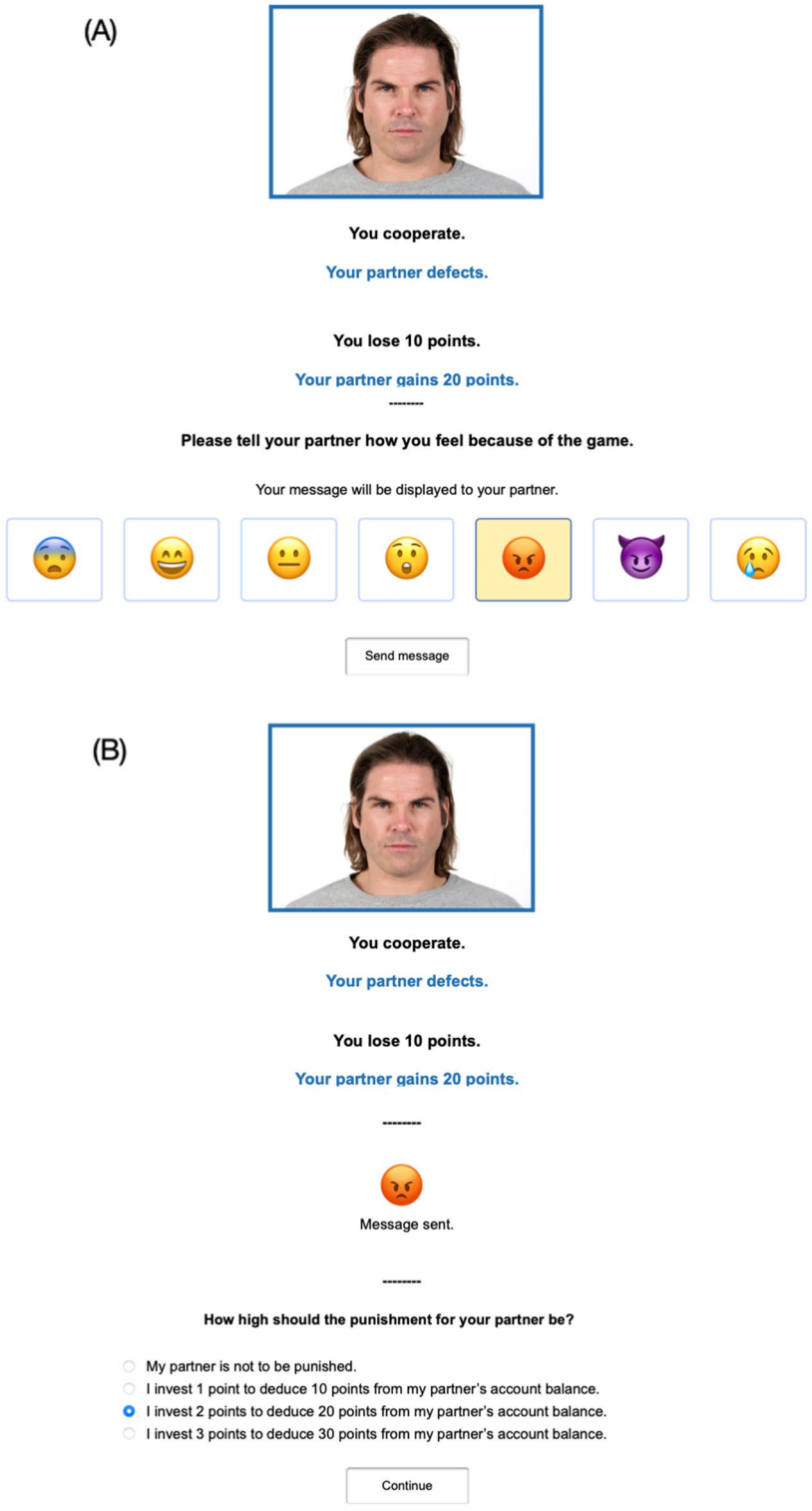

Figure 6. Example trial of the Prisoner's Dilemma game with costly punishment in the emotion-communication condition in Experiment 2. The cooperating participant chose to send the angry emoji to the defecting partner (**A**) and then decided to invest two points to deduce 20 points from the partner's account (**B**). The partner's photograph was randomly selected from the Chicago Face Database[76].

norming study ($N = 16$) in which participants were asked to choose, from a selection of suitable emojis, the ones that best expressed these emotions. The emojis that were displayed were those offered by the participant's individual operating system which allowed participants to use the same emojis as in their everyday online communication. Participants were instructed either to privately express their emotions or to communicate their emotions to their partner but they were not instructed about the particular type of emotion they were supposed to express. Instead, they could use the available emojis in any way they wanted in order to express or communicate their emotions. The most frequently selected emoji after an interaction with a defecting partner was the one expressing a neutral state (25%), closely followed by the angry emoji (20%). After having interacted with a cooperating partner, the majority of the participants (52%) selected the happy emoji. The selected emoji remained visible during the punishment decision until the end of the trial.

Ethical approval and consent to participate. Both experiments reported here were conducted in accordance with the guidelines laid down in the Declaration of Helsinki and by the German Research Foundation (DFG) including confidentiality of data and personal conduct. Informed consent was obtained prior to participation. For the noninvasive, purely behavioral research reported in the present series of experiments which carried no risk for the participants, a formal approval by the institution's ethical board is not legally required in Germany (see: https://www.dfg.de/en/research_funding/faq/faq_humanities_social_science/index.html).

Results. As in Experiment 1, we used the cooperation-and-punishment model illustrated in Fig. 3 to disentangle cooperation, the different types of punishment and the punishment bias. To analyze the results of Experiment 2 we needed three instances of the model illustrated in Fig. 3, one instance for each condition (no emotion expression, private emotion expression, emotion communication). The base model fit the data, $G^2(3) = 2.58$, $p = 0.460$. The estimates of the cooperation parameter C are depicted in Fig. 7. In both the private-emotion-expression condition, $\Delta G^2(1) = 25.73$, $p < 0.001$, $w = 0.03$, and the emotion-communication condition, $\Delta G^2(1) = 4.37$, $p = 0.037$, $w = 0.01$, cooperation decreased in comparison to the no-emotion-expression control condition, suggesting that requiring participants to focus on their emotions about the Prisoner's Dilemma game had detrimental effects on cooperation. Private emotion expression reduced cooperation even further than emotion communication, $\Delta G^2(1) = 8.52$, $p = 0.004$, $w = 0.02$.

Figure 8 shows the results pertaining to costly punishment. In line with the main finding of Experiment 1, moral punishment was significantly decreased when participants had the opportunity to communicate their emotions to their partners in comparison to both the no-emotion-expression control condition, $\Delta G^2(1) = 7.19$, $p = 0.007$, $w = 0.01$ and the private-emotion-expression condition, $\Delta G^2(1) = 4.32$, $p = 0.038$, $w = 0.01$. By contrast, privately expressing emotions did not reduce moral punishment relative to the no-emotion-expression control condition, $\Delta G^2(1) = 0.32$, $p = 0.573$, $w < 0.01$. Hypocritical punishment was enhanced in both the private-emotion-expression condition, $\Delta G^2(1) = 14.01$, $p < 0.001$, $w = 0.02$, and the emotion-communication condition, $\Delta G^2(1) = 7.55$, $p = 0.006$, $w = 0.01$, relative to the no-emotion-expression control condition, while there was no difference between the private-emotion expression condition and the emotion-communication condition, $\Delta G^2(1) = 1.00$, $p = 0.318$, $w = 0.01$.

In line with the results of Experiment 1, there was no effect of emotion expression or communication on antisocial punishment, $\Delta G^2(2) = 1.22$, $p = 0.543$, $w = 0.01$. Communicating one's emotions did not significantly affect the punishment bias compared to the no-emotion-expression control condition, $\Delta G^2(1) = 1.60$, $p = 0.206$,

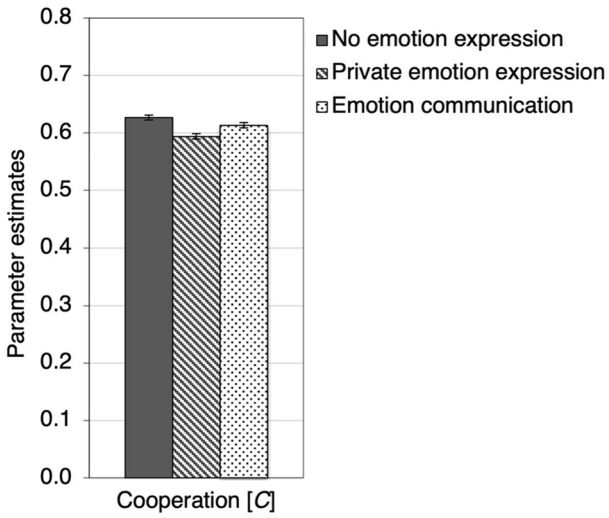

Figure 7. Estimates of the cooperation parameter C which specifies the probability of cooperation as a function of whether emotions were expressed or communicated after each round of the Prisoner's Dilemma game (no emotion expression, private emotion expression, emotion communication). Error bars represent standard errors.

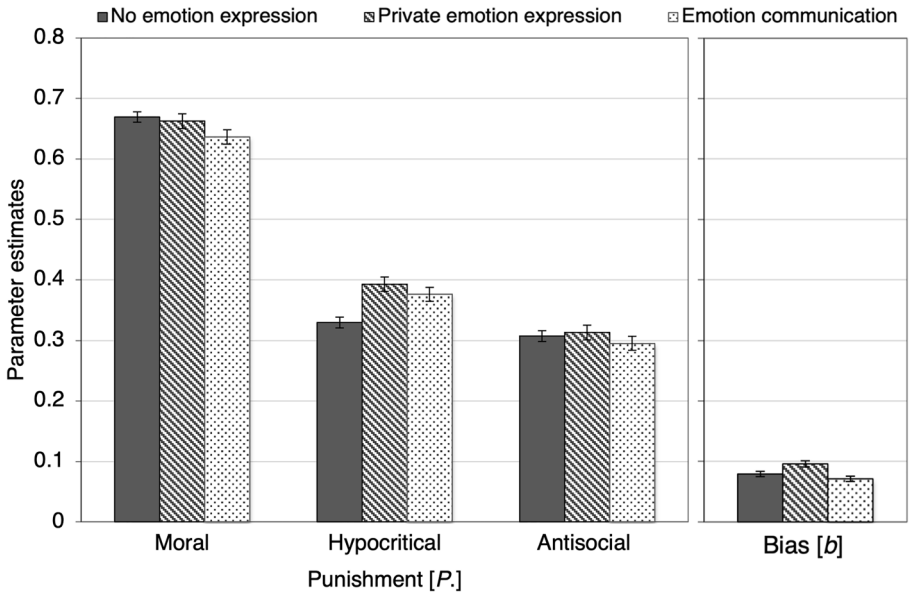

Figure 8. Estimates of parameters representing moral, hypocritical and antisocial punishment (left panel) and punishment bias (right panel) as a function of whether participants could express or communicate their emotions about the Prisoner's Dilemma game before punishing the partners (no emotion expression, private emotion expression, emotion communication). The error bars represent standard errors.

$w = 0.01$. However, expressing one's emotions privately significantly increased the punishment bias compared to the no-emotion-expression control condition, $\Delta G^2(1) = 6.23$, $p = 0.013$, $w = 0.01$, and the emotion-communication condition, $\Delta G^2(1) = 13.81$, $p < 0.001$, $w = 0.02$.

Discussion. The main result of Experiment 1 was replicated in Experiment 2. Communicating one's emotions to the partner prior to the punishment decision attenuated moral punishment. Extending the results of Experiment 1, Experiment 2 demonstrates that moral punishment is attenuated only when the emotions can be *communicated* to the interaction partner. The fact that moral punishment was not attenuated when participants expressed their emotions privately indicates that the emotional function of moral punishment is other- rather than self-directed.

The pattern of findings is thus consistent with the emotion-communication account according to which moral punishment serves to communicate one's emotional state in response to the outcome of the Prisoner's Dilemma game to the partner. The fact that privately expressing emotions had no effect on moral punishment provides evidence against the catharsis account. It seems important to note that moral punishment was only reduced and far from being abolished in the emotion-communication condition. This aspect of the present results suggests that there are other functions of moral punishment in addition to communicating one's emotions—such as that of reducing the unfair payoff imbalance (cf.[83, 84])—that are very likely much more important than the communication of emotions.

The overall level of cooperation was considerably higher in Experiment 2 than in Experiment 1. The most likely explanation of this difference is that participants could punish their partners but did not receive any punishment from the partners in Experiment 1. In Experiment 2, by contrast, the partners punished the participants for unilateral defection. It is thus reasonable to assume that the moral punishment of the participants' unilateral defection enforced more cooperation from participants in Experiment 2, consistent with the role of moral punishment in enforcing cooperation[59, 60, 62].

By increasing the sample size, we were better able to detect even smaller effects of the communication or expression of emotions in Experiment 2 compared to Experiment 1. This increased sensitivity of the statistical tests may be the reason why some of the effects that were absent or only present at a descriptive level in Experiment 1 reached statistical significance in Experiment 2. First of all, both, expressing one's emotions privately and communicating one's emotions, reduced cooperation in the Prisoner's Dilemma game relative to the no-emotion-expression control condition. Secondly, there was an increase in hypocritical punishment in both the private-emotion-expression condition and the emotion-communication condition relative to the no-emotion-expression control condition. Thirdly, privately expressing one's emotions also increased the punishment bias relative to the no-emotion-expression control condition, suggesting that privately focusing on one's emotions may increase the likelihood of punishment regardless of the outcome of the preceding interaction. These findings highlight that prompting participants to focus on emotions and thereby increasing the emotional saliency of the outcomes of the Prisoner's Dilemma game does not only have desirable effects on social interactions.

General discussion

People accept costs to punish uncooperative individuals even in one-shot interactions in which punishment cannot yield any direct personal benefits (e.g.,[5,65]). The question is why people show this economically irrational behavior. Fehr and Gächter[3] have proposed that this moral form of punishment is driven by a negative emotional response to perceived unfairness. Therefore, negative emotions are assumed to represent a key driving force behind moral punishment. Accordingly, a bulk of studies has linked the application of punishment to self-reports (e.g.,[8,14]) or physiological indices[50,52] of anger. The main aim of the present study was to test whether the expression or communication of emotions may reduce moral punishment. Both experiments confirm that the communication of emotions reduces moral punishment.

Two accounts were distinguished. According to the *emotion-communication account*, moral punishment serves to communicate one's discontent with a social interaction to the interaction partner. According to the emotion-communication account, providing participants with an alternative way of communicating their emotions should reduce moral punishment. By contrast, privately expressing emotions should be ineffective in reducing moral punishment. According to the *catharsis account*, a reduction of moral punishment relative to the control condition should already be observed when participants privately vent their emotions. Therefore, the catharsis account implies that moral punishment should be reduced in both the emotion-communication condition and the private-emotion-expression condition relative to the no-emotion-expression condition. In line with the emotion-communication account, moral punishment was attenuated when participants were provided with an option to *communicate* their emotions about the Prisoner's Dilemma game to their partners in Experiments 1 and 2. By contrast, privately expressing one's emotions did not significantly attenuate moral punishment in Experiment 2 despite the large sample size and, thus, the considerable sensitivity of the relevant statistical test. The pattern of results thus provides evidence against the catharsis account according to which punishment serves to regulate one's own emotions by venting frustration (cf.[18]) and in favor of the assumption that punishment is instrumental in regulating the other's behavior, possibly by signaling a negative emotional evaluation of unilateral defection with the purpose to enforce cooperation even if one cannot directly benefit from it[16,85].

It is also important to note that both experiments reported here consistently show that the effect of the communication of emotions on moral punishment is only small. A high level of moral punishment remained even in the emotion-communication condition. This indicates that an opportunity to communicate emotions does not come close to replacing moral punishment. However, while emotion communication appears to play only a minor role for moral punishment, this does not entail that emotions per se play a minor role. As previous studies clearly demonstrate, emotions, particularly anger, constitute a strong motivator of costly moral punishment (e.g.,[11,14,52]). However, the mere communication of emotions, in contrast to moral punishment, is not effective in removing the unfair payoff differences resulting from the Prisoner's Dilemma game and might therefore not represent a valid alternative for this emotionally driven form of punishment.

An interesting side finding is that paying attention to emotions, particularly expressing emotions privately without being able to communicate them, had a range of undesirable effects on cooperation and punishment in Experiment 2. Focusing on one's emotions about the outcomes of the Prisoner's Dilemma game—especially when they were only privately expressed—had a negative impact on one's willingness to engage in cooperation, possibly because the focus on emotions emphasized the fact that the payoff structure of the Prisoner's Dilemma game is constructed in such a way that the individual outcome of the game is always more negative after cooperation (see Fig. 1). Furthermore, expressing and communicating emotions increased hypocritical punishment and privately expressing emotions induced a punishment bias. These findings indicate that integrating emotional responses into the game paradigm may sometimes amplify rather than defuse negative social interactions. These findings go against the predictions of catharsis theory[55,56] according to which an opportunity to vent one's emotional frustration by expressing it should have had attenuating effects on punishment (cf.[18,86]). This suggests that prompting participants to focus on their emotions—which might otherwise have been less salient—may have social costs. It is a fascinating avenue for future research to further explore the possibility that making emotions salient may have detrimental effects on social interactions.

In line with the research methods of Experimental Psychology[44,46,59–62,73,74], the focus of the present study lies on the individual's cognition and behavior. Therefore, the behavior of the partner is seen as an extraneous influence that is experimentally controlled by factoring it in the design. Furthermore, in the present experiments, participants were not only informed about the raw incentive structure of the game but were also presented with a situation that was rich in social cues (including, for example, the partners' faces). This is different from Experimental Economics in which the focus is often on how the incentive structure affects the interactions of dyads or groups of individuals interacting with each other. The fact that the participants cooperate with their partners and punish their partners even when cooperation and punishment go against their financial interests, as well as the fact that the communication of emotions to the partners but not the private expression of emotions affected moral punishment, suggests that the present paradigm taps into mechanisms of social interactions. This is in line with recent experimental findings demonstrating that beliefs about the human versus preprogrammed nature of partners has surprisingly little effects on the behavior in economic games[87,88]. Nevertheless, it is of course an important goal for future research to explore whether the present conclusions hold in more ecologically realistic settings.

Conclusion

To summarize, two experiments were performed to test the effects of the expression and communication of emotions on moral punishment. In line with the emotion-communication account, moral punishment was attenuated when participants were provided with an alternative way of communicating their negative emotions in response to the outcomes of the Prisoner's Dilemma game. The effects on moral punishment were only present when the

emotions were communicated to the partner and absent when the emotions were merely privately expressed, emphasizing the communicative role of moral punishment. Still, moral punishment was attenuated but did not come close to being abolished by the communication of emotions, suggesting that communicating emotions is only a minor component of moral punishment.

Data availability

We provide the data used in our analyses via the Open Science Framework. The data are publicly available at https://osf.io/z8fwh/.

Received: 13 January 2023; Accepted: 1 September 2023
Published online: 06 September 2023

References

1. Ostrom, E., Walker, J. & Gardner, R. Covenants with and without a sword: Self-governance is possible. *Am. Polit. Sci. Rev.* **86**, 404–417. https://doi.org/10.2307/1964229 (1992).
2. Yamagishi, T. The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.* **51**, 110–116. https://doi.org/10.1037/0022-3514.51.1.110 (1986).
3. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140. https://doi.org/10.1038/415137a (2002).
4. Fehr, E. & Fischbacher, U. Social norms and human cooperation. *Trends Cogn. Sci.* **8**, 185–190. https://doi.org/10.1016/j.tics.2004.02.007 (2004).
5. Falk, A., Fehr, E. & Fischbacher, U. Driving forces behind informal sanctions. *Econometrica* **73**, 2017–2030. https://doi.org/10.1111/j.1468-0262.2005.00644.x (2005).
6. Barclay, P. Reputational benefits for altruistic punishment. *Evol. Hum. Behav.* **27**, 325–344. https://doi.org/10.1016/j.evolhumbehav.2006.01.003 (2006).
7. Henrich, J. *et al.* Costly punishment across human societies. *Science* **312**, 1767–1770. https://doi.org/10.1126/science.1127333 (2006).
8. Pillutla, M. M. & Murnighan, J. K. Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organ. Behav. Hum. Decis. Process.* **68**, 208–224. https://doi.org/10.1006/obhd.1996.0100 (1996).
9. Van't Wout, M., Kahn, R. S., Sanfey, A. G. & Aleman, A. Affective state and decision-making in the ultimatum game. *Exp. Brain Res.* **169**, 564–568. https://doi.org/10.1007/s00221-006-0346-5 (2006).
10. Mischkowski, D., Glöckner, A. & Lewisch, P. From spontaneous cooperation to spontaneous punishment–Distinguishing the underlying motives driving spontaneous behavior in first and second order public good games. *Organ. Behav. Hum. Decis. Process.* **149**, 59–72. https://doi.org/10.1016/j.obhdp.2018.07.001 (2018).
11. Gummerum, M., López-Pérez, B., Van Dijk, E. & Van Dillen, L. F. When punishment is emotion-driven: Children's, adolescents', and adults' costly punishment of unfair allocations. *Soc. Dev.* **29**, 126–142. https://doi.org/10.1111/sode.12387 (2020).
12. Gummerum, M., López-Pérez, B., Van Dijk, E. & Van Dillen, L. F. Ire and punishment: incidental anger and costly punishment in children, adolescents, and adults. *J. Exp. Child Psychol.* **218**, 105376. https://doi.org/10.1016/j.jecp.2022.105376 (2022).
13. Nelissen, R. M. & Zeelenberg, M. Moral emotions as determinants of third-party punishment: Anger, guilt and the functions of altruistic sanctions. *Judgm. Decis. Mak.* **4**, 543–553 (2009).
14. Seip, E. C., Van Dijk, W. W. & Rotteveel, M. Anger motivates costly punishment of unfair behavior. *Motiv. Emot.* **38**, 578–588. https://doi.org/10.1007/s11031-014-9395-4 (2014).
15. Gummerum, M., Van Dillen, L. F., Van Dijk, E. & López-Pérez, B. Costly third-party interventions: The role of incidental anger and attention focus in punishment of the perpetrator and compensation of the victim. *J. Exp. Soc. Psychol.* **65**, 94–104. https://doi.org/10.1016/j.jesp.2016.04.004 (2016).
16. Xiao, E. & Houser, D. Emotion expression in human punishment behavior. *Proc. Natl. Acad. Sci.* **102**, 7398–7401. https://doi.org/10.1073/pnas.0502399102 (2005).
17. Bolle, F., Tan, J. H. & Zizzo, D. J. Vendettas. *Am. Econ. J. Microecon.* **6**, 93–130. https://doi.org/10.1257/mic.6.2.93 (2014).
18. Dickinson, D. L. & Masclet, D. Emotion venting and punishment in public good experiments. *J. Public Econ.* **122**, 55–67. https://doi.org/10.1016/j.jpubeco.2014.10.008 (2015).
19. Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. The evolution of altruistic punishment. *Proc. Natl. Acad. Sci.* **100**, 3531–3535. https://doi.org/10.1073/pnas.0630443100 (2003).
20. Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560–1563. https://doi.org/10.1126/science.113375 (2006).
21. Andreoni, J. Cooperation in public-goods experiments: Kindness or confusion?. *Am. Econ. Rev.* **85**, 891–904 (1995).
22. Kollock, P. Social dilemmas: The anatomy of cooperation. *Annu. Rev. Sociol.* **24**, 183–214 (1998).
23. Axelrod, R. An evolutionary approach to norms. *Am. Polit. Sci. Rev.* **80**, 1095–1111. https://doi.org/10.2307/1960858 (1986).
24. Heckathorn, D. D. Collective action and the second-order free-rider problem. *Ration. Soc.* **1**, 78–100. https://doi.org/10.1177/1043463189001001006 (1989).
25. Yamagishi, T. Seriousness of social dilemmas and the provision of a sanctioning system. *Soc. Psychol. Q.* **51**, 32–42. https://doi.org/10.2307/2786982 (1988).
26. Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**, 980–994. https://doi.org/10.1257/aer.90.4.980 (2000).
27. Masclet, D., Noussair, C., Tucker, S. & Villeval, M.-C. Monetary and nonmonetary punishment in the voluntary contributions mechanism. *Am. Econ. Rev.* **93**, 366–380. https://doi.org/10.1257/000282803321455359 (2003).
28. Bochet, O., Page, T. & Putterman, L. Communication and punishment in voluntary contribution experiments. *J. Econ. Behav. Organ.* **60**, 11–26. https://doi.org/10.1016/j.jebo.2003.06.006 (2006).
29. Gurerk, O., Irlenbusch, B. & Rockenbach, B. The competitive advantage of sanctioning institutions. *Science* **312**, 108–111. https://doi.org/10.1126/science.1123633 (2006).
30. Fudenberg, D. & Pathak, P. A. Unobserved punishment supports cooperation. *J. Public Econ.* **94**, 78–86. https://doi.org/10.1016/j.jpubeco.2009.10.007 (2010).
31. Ambrus, A. & Greiner, B. Imperfect public monitoring with costly punishment: An experimental study. *Am. Econ. Rev.* **102**, 3317–3332. https://doi.org/10.1257/aer.102.7.3317 (2012).
32. Fischer, S., Grechenig, K. & Meier, N. Monopolizing sanctioning power under noise eliminates perverse punishment but does not increase cooperation. *Front. Behav. Neurosci.* https://doi.org/10.3389/fnbeh.2016.00180 (2016).
33. Úbeda, F. & Duéñez-Guzmán, E. A. Power and corruption. *Evolution* **65**, 1127–1139. https://doi.org/10.1111/j.1558-5646.2010.01194.x (2011).
34. van den Berg, P., Molleman, L. & Weissing, F. J. The social costs of punishment. *Behav. Brain Sci.* **35**, 42–43. https://doi.org/10.1017/s0140525x11001348 (2012).

35. Sigmund, K., Hauert, C. & Nowak, M. A. Reward and punishment. *Proc. Natl. Acad. Sci.* **98**, 10757–10762. https://doi.org/10.1073/pnas.161155698 (2001).
36. Santos, M. D., Rankin, D. J. & Wedekind, C. The evolution of punishment through reputation. *Proc. R. Soc. B. Biol. Sci.* **278**, 371–377. https://doi.org/10.1098/rspb.2010.1275 (2011).
37. dos Santos, M., Rankin, D. J. & Wedekind, C. Human cooperation based on punishment reputation. *Evolution* **67**, 2446–2450. https://doi.org/10.1111/evo.12108 (2013).
38. Roos, P., Gelfand, M., Nau, D. & Carr, R. High strength-of-ties and low mobility enable the evolution of third-party punishment. *Proc. R. Soc. B. Biol. Sci.* **281**, 20132661. https://doi.org/10.1098/rspb.2013.2661 (2014).
39. Raihani, N. J. & Bshary, R. The reputation of punishers. *Trends Ecol. Evol.* **30**, 98–103. https://doi.org/10.1016/j.tree.2014.12.003 (2015).
40. dos Santos, M. & Wedekind, C. Reputation based on punishment rather than generosity allows for evolution of cooperation in sizable groups. *Evol. Hum. Behav.* **36**, 59–64. https://doi.org/10.1016/j.evolhumbehav.2014.09.001 (2015).
41. Jordan, J. J. & Rand, D. G. Third-party punishment as a costly signal of high continuation probabilities in repeated games. *J. Theor. Biol.* **421**, 189–202. https://doi.org/10.1016/j.jtbi.2017.04.004 (2017).
42. Clutton-Brock, T. H. & Parker, G. A. Punishment in animal societies. *Nature* **373**, 209–216. https://doi.org/10.1038/373209a0 (1995).
43. Bowles, S. & Gintis, H. The evolution of strong reciprocity: Cooperation in heterogeneous populations. *Theor. Popul. Biol.* **65**, 17–28. https://doi.org/10.1016/j.tpb.2003.07.001 (2004).
44. Irwin, K. & Horne, C. A normative explanation of antisocial punishment. *Soc. Sci. Res.* **42**, 562–570. https://doi.org/10.1016/j.ssresearch.2012.10.004 (2013).
45. Frank, R. H. *Passions Within Reason: The Strategic Role of the Emotions* (W W Norton & Co, 1988).
46. Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E. & Cohen, J. D. The neural basis of economic decision-making in the ultimatum game. *Science* **300**, 1755–1758. https://doi.org/10.1126/science.1082976 (2003).
47. Camerer, C. F. *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton University Press, 2011).
48. Bosman, R. & Van Winden, F. Emotional hazard in a power-to-take experiment. *Econ. J.* **112**, 147–169. https://doi.org/10.1111/1468-0297.0j677 (2002).
49. Hopfensitz, A. & Reuben, E. The importance of emotions for the effectiveness of social punishment. *Econ. J.* **119**, 1534–1559. https://doi.org/10.1111/j.1468-0297.2009.02288.x (2009).
50. Ben-Shakhar, G., Bornstein, G., Hopfensitz, A. & Van Winden, F. Reciprocity and emotions in bargaining using physiological and self-report measures. *J. Econ. Psychol.* **28**, 314–323. https://doi.org/10.1016/j.joep.2007.02.005 (2007).
51. Civai, C., Corradi-Dell'Acqua, C., Gamer, M. & Rumiati, R. I. Are irrational reactions to unfairness truly emotionally-driven? Dissociated behavioural and emotional responses in the Ultimatum Game task. *Cognition* **114**, 89–95. https://doi.org/10.1016/j.cognition.2009.09.001 (2010).
52. Joffily, M., Masclet, D., Noussair, C. N. & Villeval, M. C. Emotions, sanctions, and cooperation. *South. Econ. J.* **80**, 1002–1027. https://doi.org/10.4284/0038-4038-2012.067 (2014).
53. Wang, C. S., Galinsky, A. D. & Murnighan, J. K. Bad drives psychological reactions, but good propels behavior: Responses to honesty and deception. *Psychol. Sci.* **20**, 634–644. https://doi.org/10.1111/j.1467-9280.2009.02344.x (2009).
54. Drouvelis, M. & Grosskopf, B. The effects of induced emotions on pro-social behaviour. *J. Public Econ.* **134**, 1–8. https://doi.org/10.1016/j.jpubeco.2015.12.012 (2016).
55. Feshbach, S. & Singer, R. D. *Television and Aggression* (Jossey-Bass Inc., 1971).
56. Lee, J. *Facing the fire: Experiencing and expressing anger appropriately*. (Bantam Books, 2011).
57. Fehr, E. & Fischbacher, U. The nature of human altruism. *Nature* **425**, 785–791. https://doi.org/10.1038/nature02043 (2003).
58. Gintis, H., Bowles, S., Boyd, R. & Fehr, E. Explaining altruistic behavior in humans. *Evol. Hum. Behav.* **24**, 153–172. https://doi.org/10.1016/S1090-5138(02)00157-5 (2003).
59. Mieth, L., Bell, R. & Buchner, A. Facial likability and smiling enhance cooperation, but have no direct effect on moralistic punishment. *J. Exp. Psychol.* **63**, 263–277. https://doi.org/10.1027/1618-3169/a000338 (2016).
60. Mieth, L., Buchner, A. & Bell, R. Effects of gender on costly punishment. *J. Behav. Decis. Mak.* **30**, 899–912. https://doi.org/10.1002/bdm.2012 (2017).
61. Mieth, L., Buchner, A. & Bell, R. Cognitive load decreases cooperation and moral punishment in a Prisoner's Dilemma game with punishment option. *Sci. Rep.* **11**, 1–12. https://doi.org/10.1038/s41598-021-04217-4 (2021).
62. Mieth, L., Buchner, A. & Bell, R. Moral labels increase cooperation and costly punishment in a Prisoner's Dilemma game with punishment option. *Sci. Rep.* **11**, 1–13. https://doi.org/10.1038/s41598-021-89675-6 (2021).
63. Anderson, C. M. & Putterman, L. Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games. Econ. Behav.* **54**, 1–24. https://doi.org/10.1016/j.geb.2004.08.007 (2006).
64. Cinyabuguma, M., Page, T. & Putterman, L. Can second-order punishment deter perverse punishment?. *Exp. Econ.* **9**, 265–279. https://doi.org/10.1007/s10683-006-9127-z (2006).
65. Carpenter, J. P. The demand for punishment. *J. Econ. Behav. Organ.* **62**, 522–542. https://doi.org/10.1016/j.jebo.2005.05.004 (2007).
66. Denant-Boemont, L., Masclet, D. & Noussair, C. N. Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Econ. Theory* **33**, 145–167. https://doi.org/10.1007/s00199-007-0212-0 (2007).
67. Herrmann, B., Thoni, C. & Gachter, S. Antisocial punishment across societies. *Science* **319**, 1362–1367. https://doi.org/10.1126/science.115380 (2008).
68. Sylwester, K., Herrmann, B. & Bryson, J. J. Homo homini lupus? Explaining antisocial punishment. *J. Neurosci. Psychol. Econ.* **6**, 167–188. https://doi.org/10.1037/npe0000009 (2013).
69. de Melo, G. & Piaggio, M. The perils of peer punishment: Evidence from a common pool resource framed field experiment. *Ecol. Econ.* **120**, 376–393. https://doi.org/10.1016/j.ecolecon.2015.05.011 (2015).
70. Pfattheicher, S., Keller, J. & Knezevic, G. Sadism, the intuitive system, and antisocial punishment in the public goods game. *Pers. Soc. Psychol. Bull.* **43**, 337–346. https://doi.org/10.1177/0146167216684134 (2017).
71. Pleasant, A. & Barclay, P. Why hate the good guy? Antisocial punishment of high cooperators is greater when people compete to be chosen. *Psychol. Sci.* **29**, 868–876. https://doi.org/10.1177/0956797617752642 (2018).
72. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191. https://doi.org/10.3758/BF03193146 (2007).
73. Bell, R., Mieth, L. & Buchner, A. Separating conditional and unconditional cooperation in a sequential Prisoner's Dilemma game. *PLoS ONE* **12**, e0187952. https://doi.org/10.1371/journal.pone.0187952 (2017).
74. Parks, C. D. & Stone, A. B. The desire to expel unselfish members from the group. *J. Pers. Soc. Psychol.* **99**, 303–310. https://doi.org/10.1037/a0018403 (2010).
75. Bürhan, P. & Alici, T. Enhanced source memory for cheaters with higher resemblance to own-culture typical faces. *Psychon. Bull. Rev.* **30**, 700–711. https://doi.org/10.3758/s13423-022-02177-y (2022).
76. Ma, D. S., Correll, J. & Wittenbrink, B. The Chicago face database: A free stimulus set of faces and norming data. *Behav. Res. Methods* **47**, 1122–1135. https://doi.org/10.3758/s13428-014-0532-5 (2015).
77. Bradley, M. M. & Lang, A.-G. Measuring emotion: The self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* **25**, 49–59. https://doi.org/10.1016/0005-7916(94)90063-9 (1994).

78. Riefer, D. M. & Batchelder, W. H. Multinomial modeling and the measurement of cognitive processes. *Psychol. Rev.* **95**, 318–339. https://doi.org/10.1037/0033-295X.95.3.318 (1988).
79. Batchelder, W. H. & Riefer, D. M. Theoretical and empirical review of multinomial process tree modeling. *Psychon. Bull. Rev.* **6**, 57–86. https://doi.org/10.3758/BF03210812 (1999).
80. Erdfelder, E. *et al.* Multinomial processing tree models: A review of the literature. *Z. Psychol.* **217**, 108–124. https://doi.org/10.1027/0044-3409.217.3.108 (2009).
81. Schmidt, O., Erdfelder, E. & Heck, D. W. Tutorial on multinomial processing tree modeling: How to develop, test, and extend MPT models. *Psychol. Methods* https://doi.org/10.1037/met0000561 (2022).
82. Moshagen, M. multiTree: A computer program for the analysis of multinomial processing tree models. *Behav. Res. Methods* **42**, 42–54. https://doi.org/10.3758/BRM.42.1.42 (2010).
83. Fehr, E. & Schmidt, K. M. A theory of fairness, competition, and cooperation. *Q. J. Econ* **114**, 817–868. https://doi.org/10.1162/003355399556151 (1999).
84. Raihani, N. J. & Bshary, R. Punishment: One tool, many uses. *Evol. Hum. Sci.* **1**, E12. https://doi.org/10.1017/ehs.2019.12 (2019).
85. Czap, H. J., Czap, N. V., Khachaturyan, M., Burbach, M. E. & Lynne, G. D. *Agricultural and Applied Economics Association (AAEA) Conferences* (Pittsburgh, Pennsylvania, 2011).
86. Bushman, B. J. Does venting anger feed or extinguish the flame? Catharsis, rumination, distraction, anger, and aggressive responding. *Pers. Soc. Psychol. Bull.* **28**, 724–731 (2002).
87. Nielsen, Y. A., Pfattheicher, S. & Keijsers, M. Prosocial behavior toward machines. *Curr. Opin. Psychol.* **43**, 260–265. https://doi.org/10.1016/j.copsyc.2021.08.004 (2022).
88. Krasnow, M. M., Howard, R. M. & Eisenbruch, A. B. The importance of being honest? Evidence that deception may not pollute social science subject pools after all. *Behav. Res. Methods* **52**, 1175–1188. https://doi.org/10.3758/s13428-019-01309-y (2020).

Author contributions

A.P., L.M., A.B., and R.B. contributed to the study conception and design. Material preparation, data collection and analysis were performed by A.P. and L.M. All authors contributed through discussion and interpretation of the results. A.P. wrote the manuscript with subsequent input and final approval from all co-authors.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

Published Article 2

The article includes Experiments 2.1 and 2.2.

Philippsen, A., Mieth, L., Buchner, A., & Bell, R. (2024). People punish defection, not failures to conform to the majority. *Scientific Reports*, *14*(1), Article: 1211. https://doi.org/10.1038/s41598-023-50414-8

scientific reports

Check for updates

OPEN

People punish defection, not failures to conform to the majority

Ana Philippsen✉, Laura Mieth, Axel Buchner & Raoul Bell

Do people punish others for defecting or for failing to conform to the majority? In two experiments, we manipulated whether the participants' partners cooperated or defected in the majority of the trials of a Prisoner's Dilemma game. The effects of this base-rate manipulation on cooperation and punishment were assessed using a multinomial processing tree model. High compared to low cooperation rates of the partners increased participants' cooperation. When participants' cooperation was not enforced through partner punishment, the participants' cooperation was closely aligned to the cooperation rates of the partners. Moral punishment of defection increased when cooperation rates were high compared to when defection rates were high. However, antisocial punishment of cooperation when defection rates were high was much less likely than moral punishment of defection when cooperation rates were high. In addition, antisocial punishment was increased when cooperation rates were high compared to when defection rates were high. The latter two results contradict the assumption that people punish conformity-violating behavior regardless of whether the behavior supports or disrupts cooperation. Punishment is thus sensitive to the rates of cooperation and defection but, overall, the results are inconsistent with the idea that punishment primarily, let alone exclusively, serves to enforce conformity with the majority.

The capacity for large-scale cooperation has crucially fostered human evolution and the establishment of societies as we know them today. As cooperation implies accepting personal costs for achieving a long-term collective benefit, there is often an incentive to free ride on the other's cooperation. This clash of individual and collective interests creates a social dilemma [cf.[1]]. The free-rider problem poses a threat: If too many people free ride, cooperation continuously loses its appeal, declines and the system collapses[2–5]. Cooperation levels vary strongly between groups as a function of a number of different factors and may fall above or below 50%, depending on the situation[6,7]. One factor that is often believed to support the maintenance of cooperation is the punishment of people who refuse to cooperate and instead defect[8–10]. While punishment of defection in repeated interactions can obviously benefit the punishing individuals by enforcing cooperation of their partners in future interactions, punishment of defection in one-shot interactions is more challenging to explain. Irrespective of this, it is a fact that people punish defectors even in one-shot interactions in which there are no obvious incentives for doing so. This is evident not only in the lab[11–13] but also in everyday social interactions. For example, in a one-time interaction on an online shopping site, buyers who feel they were treated unfairly (e.g., because they ordered goods that later turn out to be of poorer quality than advertised) may spend time and effort to write negative reviews to punish the seller. It is thus important to gain a better understanding of this puzzling yet socially tangible behavior. Two possible explanations can be distinguished for why people punish defection in one-shot interactions. One possibility is that cooperating individuals punish others specifically for their defection[14]. Another possibility is that people punish behavior to enforce conformity with the majority regardless of whether it supports or disrupts cooperation[15–18]. Here we test these accounts by examining how a manipulation of the proportions of cooperation and defection affects costly punishment in a Prisoner's Dilemma game.

The Prisoner's Dilemma game is a classical paradigm for studying cooperation. In this game, two players simultaneously decide to either cooperate or defect which leads to different possible outcomes, as determined by the game's payoff structure (see Fig. 1). A defecting player who interacts with a cooperating partner receives the highest outcome. A cooperating player who interacts with a defecting partner receives the lowest outcome. At an individual level, it is therefore more profitable to defect. At a collective level, however, cooperation is desirable

Department of Experimental Psychology, Heinrich Heine University Düsseldorf, Universitätsstrasse 1, 40225 Düsseldorf, Germany. ✉email: Ana.Philippsen@hhu.de

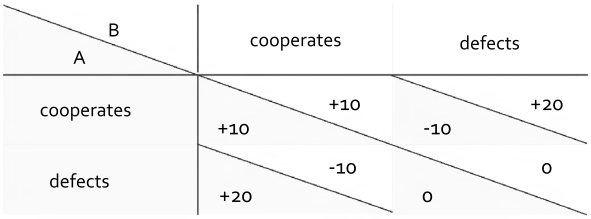

Figure 1. Examples of a payoff structure of the Prisoner's Dilemma game. The payoffs are displayed as a function of both players' decisions in the Prisoner's Dilemma game. Shaded cells denote the payoff to Player A, white cells denote the payoff to Player B.

because mutual cooperation leads to a better outcome for both interactants combined than mutual defection. This payoff structure thereby captures the basic dilemma of cooperation [cf.[1]].

People often strive to achieve mutual cooperation but try to avoid being cheated by a defecting partner who does not reciprocate cooperation. Therefore, it comes as no surprise that cooperation in economic dilemmas is often conditioned on the perceived or proclaimed prevalence of cooperation[16,19–23]. For example, Engel et al.[23] provided their participants with selective information about the behavior of either very cooperative or very uncooperative groups before participating in an economic game. Participants were more likely to cooperate when they had received information about the behavior of cooperative groups than when they had received information about the behavior of uncooperative groups. These findings indicate that a person's propensity to cooperate is influenced by the assumed prevalence of cooperation.

A factor that has been shown to crucially contribute to the maintenance of cooperation within groups is moral punishment [cf.[2]]. Here, the term *moral punishment* is used to specifically refer to the punishment of defecting partners by cooperating individuals. Defection becomes unattractive when a significant proportion of people punish defection because punishment decreases the payoffs of defecting partners. Moral punishment can thus help to solve the free-rider problem by disincentivizing defection, thereby increasing the level of cooperation[8,11,13,24,25]. However, moral punishment often entails personal costs to the punisher. Therefore, moral punishment can be considered a second-order cooperative act[5,11,13,26,27]. Given the importance of moral punishment for the establishment and maintenance of cooperation, it is crucial to understand the factors that drive people to punish others for defection.

Two broad accounts can be distinguished with regard to how the proportion of cooperation or defection should affect people's punishment behavior. One possibility is that punishment is primarily used to discourage defection, regardless of the prevalence of defection [e.g.,[14]]. This seems reasonable as punishment in economic games is, as a rule, mainly directed at defectors. However, in a small proportion of cases, people may prefer not to cooperate, sometimes leading to *antisocial punishment* of cooperative acts[28–30]. This type of punishment is termed antisocial as it undermines cooperation[31,32]. A possible explanation is that antisocial punishers are motivated by their disapproval of the normative pressure towards cooperation, exerted by individuals who are perceived as moral "do-gooders"[33–35]. While it may, at first glance, seem obvious that people should punish behaviors they disapprove of—which would explain the prevalence of both moral *and* antisocial punishment—it has been suggested that people do at least sometimes punish others for failing to conform to the majority regardless of their own private preferences[36]. The *conformity account* implies that punishment is directed at behavior that deviates from what is typical[15–18,36]. People may punish atypical behaviors to enforce conformity as conformity may reduce the costs that result from conflicts arising from uncertainty about the appropriate behavior. Furthermore, people may engage in punishment when they think that the punishment is justified by the fact that others approve of their punishment which also keeps the costs of punishment low[17]. Considering the high prevalence of cooperation in human groups and societies, punishment will often be directed at defectors who fail to contribute to the collective benefit. However, there is a dark side to enforcing conformity irrespective of the consequences of the behavior: People may antisocially punish atypical behavior even when it is promoting the collective good[16,17] simply because it violates expectations.

Here we examine how the proportion of cooperation and defection affects costly punishment in the Prisoner's Dilemma game. This study follows a previous study by Li et al.[37] in which participants had to decide between cooperation and defection in a Prisoner's Dilemma game. Prior to making their punishment decision, participants received information about eleven possible scenarios regarding how many other players had previously chosen to cooperate (ranging from less than 5% to more than 95%). Each participant was then asked to make a punishment decision for defecting partners in every one of these hypothetical scenarios. Punishment increased with the percent of cooperation in the reference group. Apart from the fact that conceptual replications of important findings are always useful, there are several additional reasons to expand on the previous findings. First, Li et al.[37] asked participants to respond to a list of eleven scenarios with different hypothetical base rates which may have accentuated the impact of the base rates on behavior. It is thus interesting to examine whether moral punishment increases with the proportion of cooperation when participants interact with partners directly. Second, Li et al.[37] concentrated only on moral punishment by requiring participants to provide punishment decisions only for defecting partners. Here, we allow participants to make punishment decisions regardless of the outcome of the Prisoner's Dilemma game which gives us the opportunity to distinguish between different types of punishment.

To allow to cleanly distinguish between different types of punishment and a bias towards punishing, the *multinomial cooperation-and-punishment model* has been developed. The model belongs to the class of multinomial

processing tree models. These models have become increasingly popular to measure the components of human decision making [for a review see[38]]. Multinomial models are flexible and accessible measurement models for which easy-to-read tutorials[39] and user-friendly software[40] exists. They disambiguate observable behavior by enabling the measurement of the processes underlying overt behavior such as different strategies in decision-making tasks[41–43]. The relationship between observable behavioral categories and the underlying processes can be visualized in a tree-like structure. Here, we use the multinomial cooperation-and-punishment model (see Fig. 3) which has been successfully applied and validated in previous studies[44–46]. Besides the cooperation parameter *C*, representing the participants' propensity to cooperate, the model entails that specific types of punishment have to be distinguished from a general punishment bias. *Moral punishment* is defined as the type of punishment that is specifically provoked when the participant's cooperation is met with the partner's defection. This type of punishment can be viewed as moral because it is aimed at retaliating the perceived violation of a cooperation norm. To illustrate, moral punishment is enhanced when the labels of the behavioral options in the Prisoner's Dilemma game facilitate a moral interpretation of the behaviors relative to when the labels are neutral[44]. *Hypocritical punishment* is the type of punishment that is specifically provoked by an interaction in which both the participant and the partner chose to defect. This type of punishment can be viewed as hypocritical because participants punish behavior in others which they themselves have shown. *Antisocial punishment* is specifically provoked by an interaction in which the participant's defection is met with a partner's cooperation. This type of punishment can be labeled as antisocial in the sense that it reflects an opposition against cooperation norms. To illustrate, previous studies[44,46] have shown that antisocial punishment is increased when participants experience normative pressure to cooperate through the moral punishment exerted by the partners. Furthermore, a proper measurement model of punishment has to take an unspecific bias to punish into account. This allows us to test whether the observed effects are distinct for the different punishment types or reflect a general increase in the willingness to punish, for example, as a way to vent frustration about factors that are unrelated to the outcome of the immediate interaction[46].

To test whether people primarily punish others for violating conformity, we manipulated the proportion of cooperating and defecting partners in the Prisoner's Dilemma game between groups. In the *cooperating-majority condition*, partners cooperated in 60% of the trials and defected in the other 40%, thereby making cooperation the dominant behavior. In the *defecting-majority condition*, this ratio was reversed, thereby making defection the dominant behavior. To ensure that the behavior of the majority was correctly represented, the participants were truthfully informed prior to the start of the game whether most partners would cooperate or defect. If punishment primarily serves to discourage defection, moral punishment should prevail irrespective of the base-rate manipulation. If punishment primarily serves to enforce conformity, punishment should be highly susceptible to the base-rate manipulation. Specifically, moral punishment should be high in the cooperating-majority condition but low or even absent in the defecting-majority condition[37]. Based on the idea that people may enforce conformity with the majority behavior regardless of their own preferences[36], hypocritical punishment should follow the same pattern as moral punishment. Hypocritical punishment should thus be increased in the cooperating-majority condition in comparison to the defecting-majority condition. If people punish to enforce conformity with the dominant behavior in the Prisoner's Dilemma game, antisocial punishment, that is, the punishment of cooperation by defecting participants, should be high in the defecting-majority condition but low or even absent in the cooperating-majority condition[16,17]. In fact, if punishment were exclusively determined by the goal to enforce conformity, then the probability that cooperating participants use moral punishment to punish a deviation from a cooperating majority should be identical to the probability that defecting participants use antisocial punishment to punish a deviation from a defecting majority.

Experiment 1

Method

Sample

We aimed to obtain about 500 valid data sets in each of the two experiments with the help of the online panel provider *mingle*. Of the data files of those participants who started the Prisoner's Dilemma game, 54 data files had to be removed because the participants did not complete the experiment and 70 data files had to be excluded due to double participation. The final sample consisted of 544 participants (305 female, 239 male) aged 18–88 ($M = 49$, $SD = 15$) years. A sensitivity analysis showed that with a sample size of $N = 544$ and 25 decisions per participant it was possible to detect effects of the base-rate manipulation on the cooperation and punishment parameters of the multinomial cooperation-and-punishment model (see below) of the size $w = 0.03$ with a statistical power of $1 - \beta = 0.95$ at an α level of 0.05[47].

Base-rate manipulation

At the start of the experiment, participants were assigned to either the cooperating-majority condition ($n = 278$) or the defecting-majority condition ($n = 266$). Depending on the assigned condition, participants were instructed either that most people would cooperate and only some would defect or that most people would defect and only some would cooperate. These instructions were used to ensure that participants formed a correct representation about the majority behavior even before the Prisoner's Dilemma game started.

The fact that the partners' responses were determined by a computer program then allowed us to manipulate the proportion of cooperating and defecting partners in line with these instructions. Experimentally manipulating the partner behavior is a common approach in Experimental Psychology to generate varying base rates while maintaining control over confounding factors that may otherwise influence partner behavior[16,48–54]. In the cooperating-majority condition, partners were programmed to cooperate in 60% of the trials and to defect in 40% of the trials. In the defecting-majority condition, this ratio was reversed.

Prisoner's Dilemma game

Materials and procedure of the Prisoner's Dilemma game were parallel to those of a previous online study examining costly punishment in the Prisoner's Dilemma game[46]. After giving their informed consent and answering demographic questions, participants received the instructions for the Prisoner's Dilemma game. Participants of the online panel provider mingle are compensated with points that can be exchanged for online vouchers, charity donations or money (with 1 point corresponding to 1 Euro cent). Participants were thus informed that they were playing for points which they would be awarded by mingle at the end of the study in addition to the points they would receive for participating in the study. At the start of the experiment, participants were endowed with 150 points. Participants played 30 trials, five of which were training trials, of a simultaneous one-shot Prisoner's Dilemma game with a costly punishment option.

Each trial of the Prisoner's Dilemma started with the display of the participant's current account balance in the middle of the screen. Participants knew that they would interact with a different partner in every trial. Upon clicking a "Continue" button, the interaction partner was shown. To emphasize the social nature of the game, participants saw a color photograph (266 × 186 pixels) of a different partner in each trial. To this end, photographs of 30 white adult faces, half of which were female and half of which were male, were randomly drawn from the Chicago Face Database[55]. All faces had a neutral expression and were shown from a frontal view. The partner's photograph was centered on-screen and surrounded by a blue frame (4 pixels, see Fig. 2).

Beneath the photograph, participants could choose to cooperate or to defect by clicking the corresponding button and submitting their choice with a "Continue" button. Participants had been instructed that they and their partner would see their decisions to cooperate or to defect simultaneously. There were four different outcomes depending on both partners' decisions, as illustrated by the payoff matrix in Fig. 1. Participants knew that mutual cooperation would lead to a gain of 10 points for each partner while mutual defection would lead to no gain or loss. They also knew that a defecting partner would gain 20 points when interacting with a cooperating partner who would in return lose 10 points. Participants received feedback about their own decision (e.g., "You cooperate.") and their partner's decision (e.g., "Your partner defects.") and how these decisions affected each players' account balance (e.g., "You lose 10 points.", "Your partner gains 20 points."). Feedback regarding the participant's decision and outcome was displayed in black font color whereas feedback on the partner's decision and outcome was shown in blue font color, corresponding to the blue frame around the partner's photograph. The photograph and the feedback of the interaction outcome remained visible on the screen until the end of each trial.

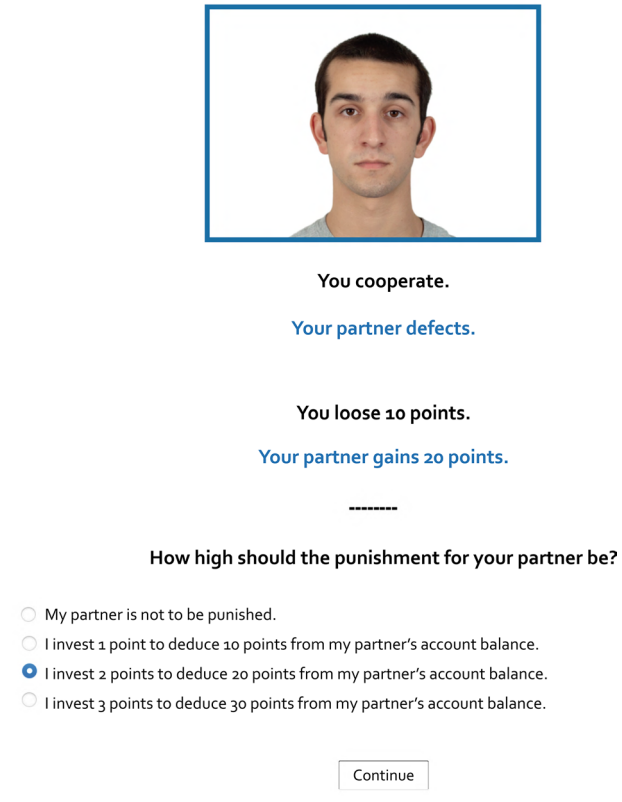

Figure 2. Example trial of the Prisoner's Dilemma game with costly punishment. In this example trial, the participant cooperated while the partner defected which led to a loss of 10 points for the participant and a gain of 20 points for the partner. The participant then chose to morally punish the partner by investing 2 points so that 20 points were subtracted from the partner's account balance. The partner's photograph was randomly selected from the Chicago Face Database[55].

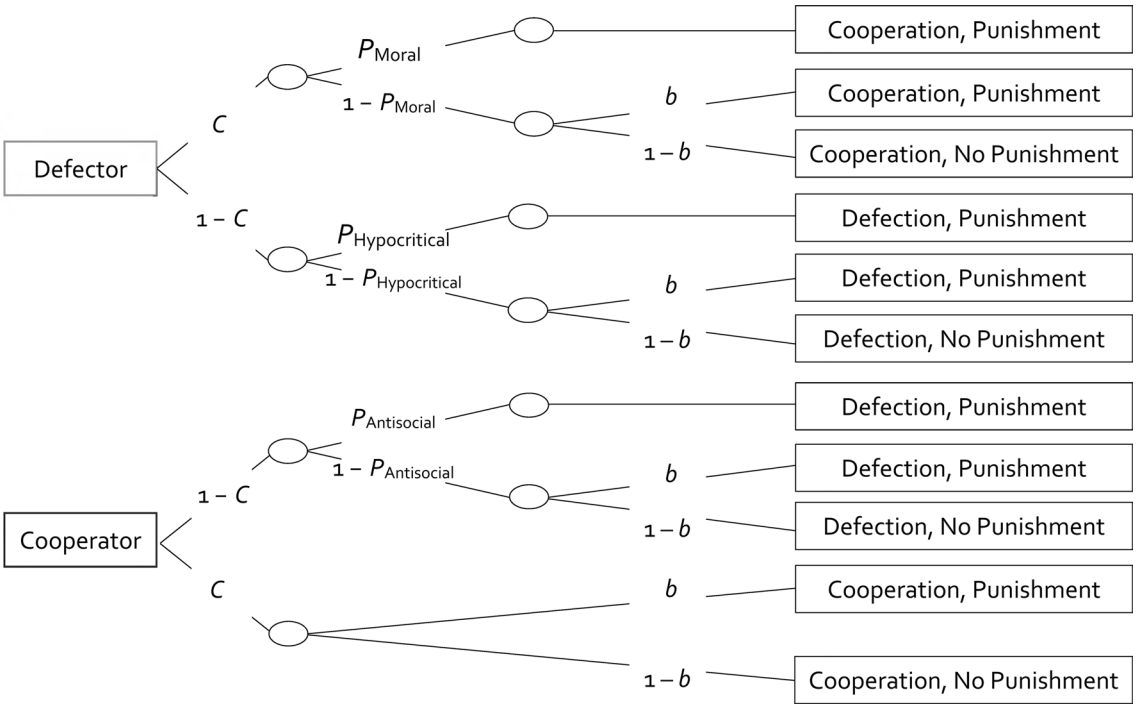

Figure 3. Multinomial cooperation-and-punishment model. Rectangles on the left represent the partner's behavior. Rectangles on the right represent the participant's behavior. Letters along the branches indicate the parameters of the model (C = cooperation, P_{Moral} = moral punishment of unilateral defection, $P_{Hypocritical}$ = hypocritical punishment following mutual defection, $P_{Antisocial}$ = antisocial punishment of unilateral cooperation; b = unspecific punishment bias).

Costly-punishment option

After each interaction in the Prisoner's Dilemma, participants were offered a costly punishment option. Participants could decide either not to punish their partner or to invest 1, 2 or 3 points to deduce 10, 20 or 30 points, respectively, from their partner's account balance. Participants were informed beforehand that their partners would simultaneously make their decision to punish the participants. As in a previous experiment[46], the partners were programmed to always punish unilateral defection of the participants by deducing a randomly determined amount of 10, 20 or 30 points from the participants' account. Upon clicking a "Continue" button, participants received feedback about their own punishment decision (e.g., "You invest 2 points to punish your partner.") and its effect on the partner's account balance (e.g., "20 points will be deducted from your partner's account balance."). Participants simultaneously learned about their partner's punishment decision (e.g., "Your partner does not punish you.") and its effect on their own account balance (e.g., "No fine will be deducted from your account balance."). With a "Continue" button, participants could then start the next trial. The average final account balance was 128 (SD = 54) points.

The cooperation-and-punishment model

Multinomial models have become increasingly popular as they allow to estimate the latent cognitive processes that underlie observable categorical behavioral data [e.g.,[42,43,56,57]]. The cooperation-and-punishment model used here has been successfully used to measure cooperation and punishment in previous studies[44–46]. It is illustrated in Fig. 3. The model incorporates two trees, one for the defecting partners and one for the cooperating partners. The first latent process specified in both trees is the participant's propensity to cooperate which is assumed to be independent of the individual partner's behavior that is revealed only after the participant's decision. Therefore, the same parameter C can be used for both trees: Participants may choose to cooperate with probability C or to defect with probability 1–C. Depending on whether the partner cooperates or defects, distinct types of punishment may occur. If the participant's cooperation is met with the partner's defection, the participant may apply moral punishment with probability P_{Moral}. Even if the participant does not apply moral punishment with probability $1 - P_{Moral}$, the participant may still punish the partner because of an unspecific punishment bias with probability b. With probability $1 - b$, no punishment is applied. After the mutual defection of both players, hypocritical punishment may be applied with probability $P_{Hypocritical}$. Even if no hypocritical punishment is applied with probability $1 - P_{Hypocritical}$, punishment may still occur due to the unspecific punishment bias with probability b. With probability $1 - b$, no punishment is applied. If the participant's defection mismatches with the cooperation of the partner, the participant may apply antisocial punishment with probability $P_{Antisocial}$. If the participant does not apply antisocial punishment with probability $1–P_{Antisocial}$, punishment may still occur due to the unspecific punishment bias with probability b. With probability $1 - b$, no punishment is applied. Mutual cooperation does not provide any specific reason to punish the partner. Any punishment in this case is therefore used to estimate the punishment bias b which reflects an unspecific tendency to punish the partner irrespective of the outcome

of the interaction. To illustrate, if an emotion-centered processing focus induces feelings of frustration, this may well result in the indiscriminate punishment of partners irrespective of the outcomes of the Prisoner's Dilemma game which is then reflected in the punishment bias b[46]. The model implies that this punishment bias has to be distinguished from types of punishment that discriminate between different partner behaviors in a parallel way to how response bias has to be distinguished from more specific responses in other decision-making models[58–62]. With probability 1–b, no punishment is applied.

Results

When using multinomial models to test substantive hypotheses it is ideal to begin with a base model that fits the data. A multinomial model fits the data if the goodness-of-fit test assessing the discrepancy between the observed responses and the responses predicted by the model is non-significant, as indicated by a p-value larger than the α-level (usually 0.05). The corresponding goodness-of-fit statistic G^2 is chi-square distributed with degrees of freedom indicated in parentheses. To analyze the present data, two sets of the trees of the multinomial cooperation-and-punishment model depicted in Fig. 3 are needed for the base model, one set for the cooperating-majority condition and one for the defecting-majority condition. This base model fit the data, $G^2(2) = 3.46$, $p = 0.177$.

Multinomial models allow hypothesis tests to be performed directly at the level of the parameters representing the cognitive processes assumed to underly observed behavior. For example, the hypothesis that the participants' propensity to cooperate is significantly higher in the cooperating-majority condition than in the defecting-majority condition can be tested by restricting the C parameters of the two conditions to be equal. If this equality restriction significantly worsens the fit of the restricted model compared to the base model, as indicated by the ΔG^2 statistic which is chi-square distributed with degrees of freedom displayed in parentheses, it can be concluded that the participants' propensity to cooperate differs between the two conditions. Figure 4 displays the estimates of the cooperation parameter C. Cooperation was indeed significantly higher in the cooperating-majority condition than in the defecting-majority condition, $\Delta G^2(1) = 272.57$, $p < 0.001$, $w = 0.14$.

Estimates of the punishment parameters are shown in Fig. 5. In line with the conformity account, moral punishment was significantly higher in the cooperating-majority condition than in the defecting-majority condition, $\Delta G^2(1) = 19.79$, $p < 0.001$, $w = 0.04$. Also consistent with the conformity account, a high base rate of cooperation in comparison to defection led to an increase in hypocritical punishment, $\Delta G^2(1) = 7.88$, $p = 0.005$, $w = 0.02$. So far, the data seem compatible with the conformity account. However, participants were much more likely to use moral punishment to punish a deviation from the cooperating-majority group than to use antisocial punishment to punish a deviation from the defecting-majority group, $\Delta G^2(1) = 557.29$, $p < 0.001$, $w = 0.20$, which provides evidence against the assumption that punishment is exclusively determined by the goal to enforce conformity. Also, in direct opposition to the prediction of the conformity account, antisocial punishment was enhanced in the cooperating-majority condition compared to the defecting-majority condition, $\Delta G^2(1) = 11.55$, $p = 0.001$, $w = 0.03$. Finally, the punishment bias did not differ between the cooperating-majority condition and the defecting-majority condition, $\Delta G^2(1) = 2.10$, $p = 0.147$, $w = 0.01$.

Discussion

The aim of the experiment was to test two accounts of punishment. If punishment serves to enforce conformity, then punishment should be directed at punishing *any* deviation from the majority and should therefore be affected by the proportion of cooperating and defecting partners. Specifically, moral punishment should be increased when cooperation is the dominant behavior whereas antisocial punishment should be increased when defection is the dominant behavior. In line with the conformity account, moral punishment was significantly

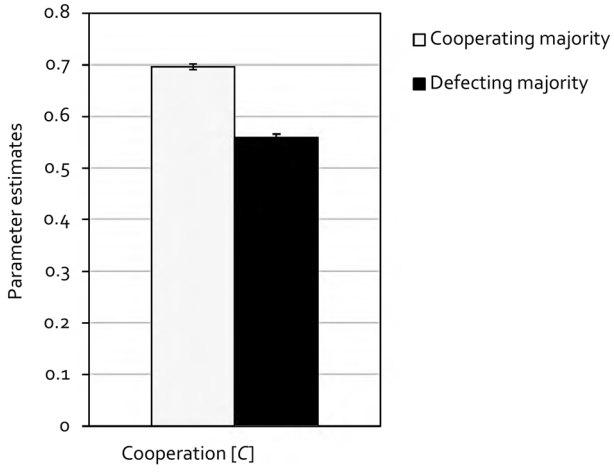

Figure 4. Estimates of the cooperation parameter C as a function of cooperation base rates in Experiment 1 (with partner punishment). In the cooperating-majority condition, partners cooperated in 60% of the trials and defected in the other 40%. In the defecting-majority condition, this ratio was reversed. Error bars represent standard errors.

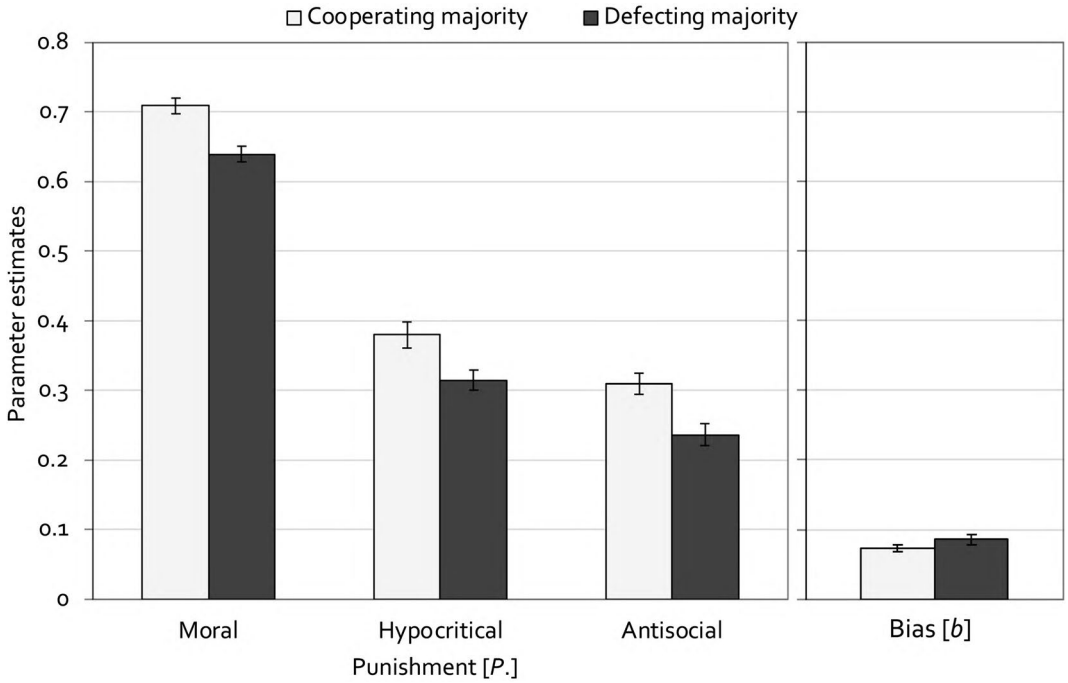

Figure 5. Estimates of the parameters representing moral, hypocritical and antisocial punishment (left panel) and the punishment bias (right panel) in Experiment 1 (with partner punishment). In the cooperating-majority condition, partners cooperated in 60% of the trials and defected in the other 40%. In the defecting-majority condition, this ratio was reversed. Error bars represent standard errors.

higher in the cooperating-majority condition compared to the defecting-majority condition. In line with the idea that people punish to enforce conformity regardless of their own preferences[36], hypocritical punishment was also higher in the cooperating-majority condition compared to the defecting-majority condition. However, if punishment were exclusively determined by the goal to enforce conformity, then the probability that participants use moral punishment to punish a deviation from the cooperating majority should be identical to the probability that they use antisocial punishment to punish a deviation from the defecting majority. This was clearly not the case. In addition, antisocial punishment was enhanced in the cooperating-majority condition compared to the defecting-majority condition which is also not compatible with the conformity account. In other words, these results clearly rule out that people punish what is uncommon without regard to the type of behavior that is shown. The fact that moral punishment was much more likely than antisocial punishment regardless of the proportion of cooperation and defection strongly suggests that, while punishment is affected by the base rates of cooperation and defection, punishment primarily serves to discourage defection[14]. Finally, it seems noteworthy that the punishment bias was not affected by the base-rate manipulation, suggesting that a high proportion of defection did not generally decrease the propensity to punish.

Similar to the punishment parameters, the probability of cooperation was significantly higher in the cooperating-majority condition than in the defecting-majority condition. This is in line with a bulk of studies reporting how participants condition their own cooperation on the perceived or proclaimed cooperation rates of others[16,19–21,23]. Interestingly, while being clearly influenced by the prevailing cooperation rates, participants' propensity to cooperate still exceeded the base rate of cooperation in both the cooperating-majority condition and the defecting-majority condition: When partners cooperated in 60% of the trials in the cooperating-majority condition, participants cooperated in 70% of the trials, whereas when partners cooperated in 40% of the trials in the defecting-majority condition, participants nevertheless cooperated in 56% of the trials.

Cooperation rates may have been elevated in Experiment 1 because the partners reliably punished the unilateral defection of the participants and thereby discouraged defection. This could potentially also explain why moral punishment remained at a high level in the cooperating-majority condition as well as the defecting-majority condition in that it seems conceivable that participants may have followed the example of their partners when deciding to apply moral punishment [cf.[37,63–65]]. It thus is necessary to test how the proportion of cooperation and defection affects moral punishment when participants cannot base their own punishment decisions on the example set by the partners. Therefore, we tested in Experiment 2 how the proportion of cooperating and defecting partners affects moral punishment when punishment is unilaterally available to the participants but not to the partners, as in previous experiments[44,66,67]. If the effects of the base-rate manipulation are independent of the presence or absence of partner punishment, the pattern of results from Experiment 1 should be replicated. To the degree that the effects of the base rate manipulation depend on the presence or absence of the partners' moral punishment, the effects should differ between Experiments 1 and 2.

Experiment 2

Method

Parallel to Experiment 1, we aimed at recruiting about 500 valid data sets with the help of the online panel provider *mingle*. Of those participants who had started the game, 54 data files had to be excluded because the participants did not complete the experiment; 48 data files had to be excluded due to double participation. The final sample consisted of $N = 495$ participants (209 female, 284 male, 2 non-binary) aged 18–90 years with a mean age of 49 ($SD = 16$) years. The slightly smaller sample size relative to that of Experiment 1 ($n = 544$) did not substantially affect the sensitivity of the statistical tests. It was still possible to detect effects of $w = 0.03$ with a statistical power of $1 - \beta = 0.95$ at an α level of 0.05 when comparing the cooperation and punishment parameters between the cooperating-majority condition ($n = 250$) and the defecting-majority condition ($n = 245$)[47].

Materials and procedure were identical to those of Experiment 1 with the exception that the punishment option was unilaterally available to the participants, implying that the partners did not punish participants' defection. Participants therefore only received feedback about their own punishment decision and its effect on the partner's account balance. The average final account balance was 275 ($SD = 100$) points.

Results

As in Experiment 1, the data were analyzed using the multinomial cooperation-and-punishment model (see Fig. 3). The goodness-of-fit test showed that the base model provided a good fit to the data, $G^2(2) = 0.40$, $p = 0.819$. The estimates of the cooperation parameter C are shown in Fig. 6. Replicating the results of Experiment 1, cooperation was significantly higher in the cooperating-majority condition in comparison to the defecting-majority condition, $\Delta G^2(1) = 188.35$, $p < 0.001$, $w = 0.12$.

Figure 7 displays the estimates of the punishment parameters (left panel) and the punishment bias (right panel). In line with Experiment 1, moral punishment was significantly higher in the cooperating-majority condition than in the defecting-majority condition, $\Delta G^2(1) = 10.01$, $p = 0.002$, $w = 0.03$. Also consistent with Experiment 1, a high base rate of cooperation in comparison to defection led to an increase in hypocritical punishment, $\Delta G^2(1) = 8.70$, $p = 0.003$, $w = 0.03$. Further replicating Experiment 1 and in direct opposition to the prediction of the conformity account, moral punishment in the cooperating-majority group was much more likely than antisocial punishment in the defecting-majority group, $\Delta G^2(1) = 486.20$, $p < 0.001$, $w = 0.20$, which is evidence against the assumption that these types of punishment are exclusively determined by the goal to enforce conformity. Parallel to the results of Experiment 1, and further disconfirming the conformity account, antisocial punishment was enhanced in the cooperating-majority condition compared to the defecting-majority condition, $\Delta G^2(1) = 4.87$, $p = 0.027$, $w = 0.02$. Finally, the punishment bias was significantly higher in the defecting-majority condition than in the cooperating-majority condition, $\Delta G^2(1) = 11.46$, $p = 0.001$, $w = 0.03$.

Discussion

The aim of Experiment 2 was to test whether the effects of Experiment 1 can be replicated when partners do not morally punish defection. Replicating the main findings of Experiment 1, moral, hypocritical and antisocial punishment were significantly higher in the cooperating-majority condition in comparison to the defecting-majority condition in Experiment 2. While the effects of the base-rate manipulation on moral and hypocritical punishment

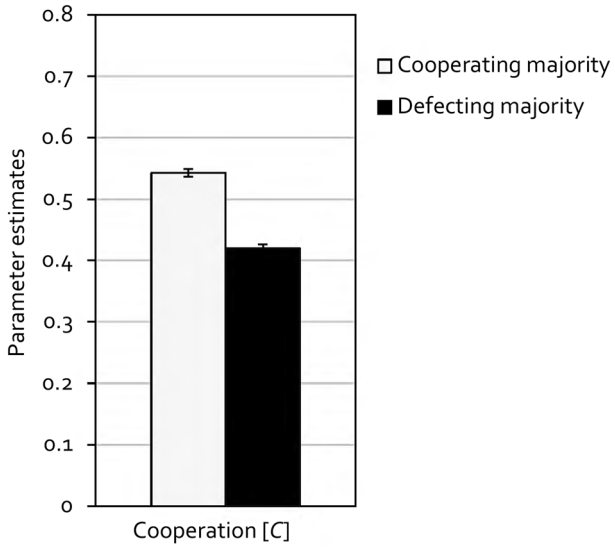

Figure 6. Estimates of the cooperation parameter C as a function of cooperation base rates in Experiment 2 (without partner punishment). In the cooperating-majority condition, partners cooperated in 60% of the trials and defected in the other 40%. In the defecting-majority condition, this ratio was reversed. Error bars represent standard errors.

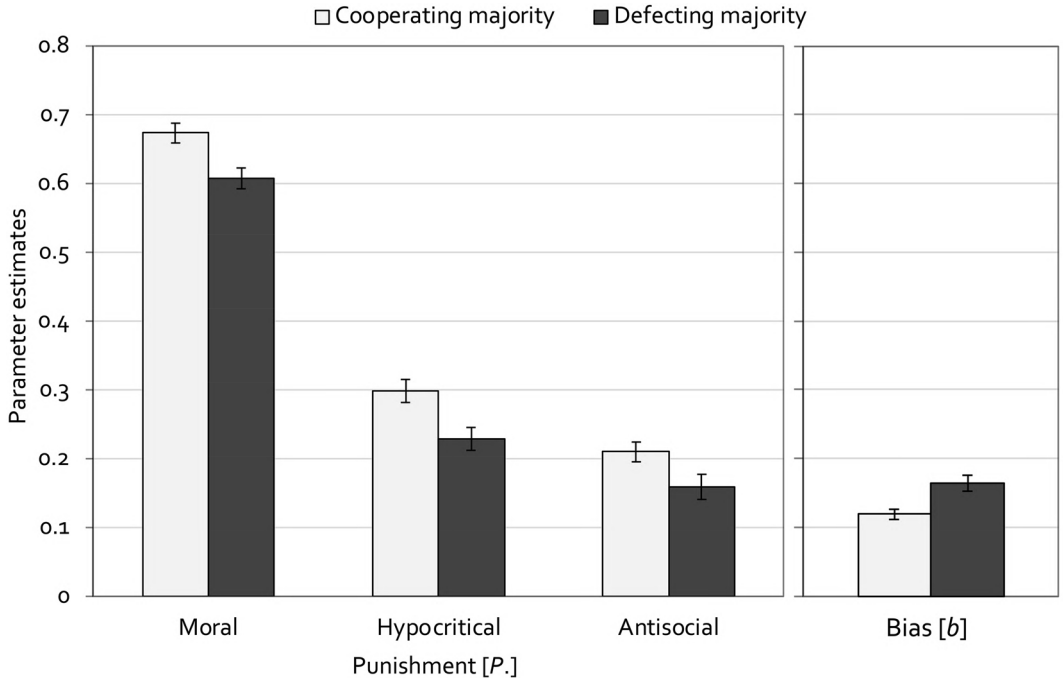

Figure 7. Estimates of the parameters representing moral, hypocritical, and antisocial punishment (left panel) and the punishment bias (right panel) in Experiment 2 (without partner punishment). In the cooperating-majority condition, partners cooperated in 60% of the trials and defected in the other 40%. In the defecting-majority condition, this ratio was reversed. Error bars represent standard errors.

are partly in line with the conformity account, the effect on antisocial punishment is in direct opposition to what the conformity account implies, as is the fact that moral punishment in the cooperating-majority group was much more likely than antisocial punishment in the defecting-majority group. This necessarily leads to the conclusion that people do not punish behavior only because it deviates from what the majority does. Interestingly, moral punishment rates still remained at a high level even though, in contrast to Experiment 1, participants could not follow their partners' example when deciding whether to use moral punishment. This supports the conclusion that when applying moral punishment people are not merely conforming to the observed punishment behavior of their partners. Instead, there seems to be an intrinsic motive for punishing defection. The present results thereby nicely fit with the recently proposed moral preference hypothesis according to which costly punishment of defection is driven by an internalized preference to act in a way that is typically considered moral[68,69]. Other than in Experiment 1, the punishment bias was increased in the defecting-majority condition in Experiment 2. This suggests that when mainly interacting with defecting partners, participants tend to randomly punish their partners more frequently, possibly as a way to vent frustration about the high prevalence of defection. As in Experiment 1, it can be concluded that there was no general reluctance to punish in the defecting-majority condition.

The effect of the base-rate manipulation on the participants' own inclination to cooperate was replicated in Experiment 2. Moreover, when cooperation was not enforced by moral punishment, participants' own cooperation rates aligned more closely with the manipulated base rates than participants' cooperation rates in Experiment 1. This points to a conformist motive behind cooperation, in line with the previous literature[16,21,23].

General discussion

The moral punishment of defection is integral to enforcing and maintaining cooperation in the light of the free-rider problem e.g.,[8,13]. It is therefore important to understand what drives people to accept the costs associated with punishing others. If punishment primarily serves to discourage defection[14], people should use the punishment option primarily to morally punish unilateral defection while antisocial punishment should occur with a comparatively smaller probability regardless of whether the majority of the partners cooperates or defects. If punishment primarily serves to enforce conformity[15–18,36], people should punish all behaviors that do not conform to what the majority does regardless of the specific type of behavior in question. Both accounts predict that people will primarily use moral punishment when most people cooperate. However, the conformity account makes the unique prediction that moral punishment should become less prevalent when most people defect. The present study followed a previous study by Li et al.[37] who found that moral punishment indeed decreases with decreasing cooperation rates. A limitation of the previous study was that participants conditioned their responses on instructed hypothetical base rates of cooperative behavior without experiencing them directly. In the present study, we used a Prisoner's Dilemma game with a costly punishment option and manipulated whether the participants' partners cooperated or defected in the majority (60%) of trials. In line with the study by Li et al.[37], we consistently found across two experiments that moral punishment was more prevalent in the

cooperating-majority condition than in the defecting-majority condition. Extending the previous study, we found across both experiments that hypocritical punishment was also more prevalent when the base rate of cooperation was high compared to when it was low. This pattern is consistent with the idea that people may enforce conformity with the majority even when they do not share the preferences of the majority[36].

So far, the results seem to support the conformity account. However, there are several aspects of the results that are inconsistent with this account. First, moral punishment of defection in the cooperating-majority group was much more likely than antisocial punishment of cooperation in the defecting-majority group which is inconsistent with the assumption that these types of punishment are exclusively determined by the goal to enforce conformity. If that were the case, then the probability of antisocial punishment in the defecting-majority condition should be as high as the probability of moral punishment in the cooperating-majority condition. This prediction is clearly contradicted by the data we observed. Another important prediction of the conformity account is that people should be more likely to use antisocial punishment to punish cooperation in the defecting-majority condition than in the cooperating-majority condition. However, antisocial punishment was lower in the defecting-majority condition than in the cooperating-majority condition, in direct opposition to the prediction of the conformity account.

Overall, the results are thus most compatible with an integrative account according to which people primarily use punishment to discourage defection[14] but still adjust the punishment to the perceived cooperation levels. A high prevalence of cooperation is often believed to create or strengthen a cooperative norm[22,23,70]. Therefore, defection in a cooperative environment may be perceived as being more deviant and thus more deserving of punishment than defection in an environment in which defection is common[37,71]. Hypocritical punishment may be used to make up for one's own failure to adhere to the cooperative norm as it has been observed that people tend to use punishment to feign sincere support of the majority group behavior despite their actual disapproval[36]. Antisocial punishment may be assumed to be driven by an opposition to the normative pressure towards cooperation that is not shared. For instance, antisocial punishment has often been attributed to an aversion to morally superior "do-gooders"[31,33–35]. People may use antisocial punishment as a retaliation for the embarrassment evoked by one's unilateral defection. When cooperation is more prevalent, the embarrassment that is caused by the norm violation could well be amplified, causing a stronger urge to harm or devaluate the opponent for causing the embarrassment. In fact, increased levels of do-gooder derogation have been reported when the perceived number of people belonging to the morally superior group was high because a strong conformist pressure created a stronger threat to one's moral identity [72,73]. It thus is psychologically plausible that antisocial punishment increases rather than decreases with a strong normative pressure towards cooperation as it may reflect a direct opposition towards cooperation.

Given that the present results suggest that high cooperation levels lead to more antisocial punishment, the question arises as to why the prevalence of antisocial punishment is often negatively related to the prevalence of cooperation in cross-cultural comparisons[28,31] in which participants from societies with low cooperation rates usually experience more antisocial punishment. Here it must be kept in mind that such findings are only correlational and the low cooperation levels might be a consequence of the detrimental effect of antisocial punishment on cooperation instead of the cause for the high antisocial punishment. In the present study, we used an experimental manipulation of the proportion of cooperation and defection to identify its effects on the different types of punishment without having to second-guess the direction of the effects. It also seems striking that most evidence in favor of the conformity account of costly punishment comes from the Public Goods game that examines cooperation within larger groups[16,17], but see[23]. It thus seems conceivable that the requirement to find a balance between individual and collective interests in larger group settings may create stronger conformist pressures than the dyadic interactions in the Prisoner's Dilemma game.

Finally, it seems noteworthy that a conformity effect was not only observed with respect to punishment but also with respect to cooperation. Participants' willingness to cooperate was clearly affected by whether the majority of the partners cooperated or defected. This is in line with a bulk of studies on how participants condition their cooperation on perceived or proclaimed cooperation rates of others[16,19–23]. Interestingly, cooperation rates clearly exceeded the manipulated base rate when the partners applied moral punishment to discourage defection (Experiment 1). Without partner punishment (Experiment 2), participants lacked an economic incentive to cooperate. As a result, the participants' propensity to cooperate aligned more closely with the manipulated base rates which therefore points to a conformist motive behind cooperation.

The aim of the present experiments was to test whether costly punishment is affected by the prevalence of cooperation. By varying the cooperation rates of simulated interaction partners in a between-groups design we were able to experimentally manipulate the base rates of cooperation and defection while maintaining experimental control over extraneous factors that may otherwise influence the players' behaviors. This approach differs from what is common practice in Experimental Economics but conforms to research traditions in Experimental Psychology [e.g.,[16,48,50,52,54]]. In this context, two observations seem worth noting. First, participants readily cooperated with, and even punished, their partners even though this implied sacrificing some of their own money. Second, the punishment rates observed in the highly controlled experiments presented here are comparable to the punishment rates reported in studies using real interaction partners [e.g.,[12,27]]. Taken together, these observations suggest that the present experimental paradigm reliably activated mechanisms of social interactions. Still, it is of course an intriguing avenue for future research to test whether the present conclusions generalize to different settings in which, for instance, participants interact in human dyads.

Conclusion

Do we punish others for failing to conform to the majority irrespective of the specific type of behavior in question? The present results clearly demonstrate that people do not punish a specific behavior only because it is

uncommon. Regardless of the prevalence of cooperation or defection, participants primarily used moral punishment to express their disapproval of a partner's unilateral defection. This indicates that punishment is primarily used to discourage defection and not to enforce blind conformity with the majority. Nevertheless, there were several ways in which participants' behaviors were sensitive to the proportion of cooperation and defection they experienced. The present results corroborate previous findings [cf.[37]] suggesting that moral punishment increases with the proportion of cooperating partners in the Prisoner's Dilemma game. In other words, defecting behavior that deviates from what the majority does is punished more. The same was found for hypocritical punishment. Nevertheless, moral punishment of deviations from a cooperating majority was much higher than antisocial punishment of deviations from a defecting majority which should not be the case if these types of punishment were exclusively determined by the goal to enforce conformity. Furthermore, antisocial punishment was increased when the prevalence of cooperation was high which suggests that antisocial punishment increases with the perceived pressure towards cooperation. Punishment is thus sensitive to the rates of cooperation and defection but, overall, the results are inconsistent with the idea that punishment primarily, let alone exclusively, serves to enforce conformity.

Ethics approval and consent to participate

The study was conducted in accordance with the guidelines laid down in the Declaration of Helsinki and by the German Research Foundation (DFG) including confidentiality of data and personal conduct. Informed consent was obtained prior to participation. For the noninvasive, purely behavioral research reported in the present series of experiments which carried no risk for the participants, a formal approval by the institution's ethical board is not legally required in Germany (see: https://www.dfg.de/en/research_funding/faq/faq_humanities_social_science/index.html).

Data availability

We provide the data used in our analyses via the Open Science Framework. The data is publicly available at https://osf.io/fycg3/.

Received: 25 August 2023; Accepted: 19 December 2023
Published online: 12 January 2024

References

1. Kollock, P. Social dilemmas: The anatomy of cooperation. *Annu. Rev. Sociol.* **24**, 183–214 (1998).
2. Fehr, E. & Fischbacher, U. Social norms and human cooperation. *Trends Cognit. Sci.* **8**, 185–190. https://doi.org/10.1016/j.tics.2004.02.007 (2004).
3. Andreoni, J. Cooperation in public-goods experiments: Kindness or confusion?. *Am. Econ. Rev.* **85**, 891–904 (1995).
4. Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560–1563. https://doi.org/10.1126/science.113375 (2006).
5. Boyd, R. & Richerson, P. J. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195. https://doi.org/10.1016/0162-3095(92)90032-Y (1992).
6. Rapoport, A. & Chammah, A. M. *Prisoner's Dilemma: A Study in Conflict and Cooperation* Vol. 165 (University of Michigan Press, 1965).
7. Chen, X., Szolnoki, A. & Perc, M. Competition and cooperation among different punishing strategies in the spatial public goods game. *Phys. Rev. E.* **92**, 012819. https://doi.org/10.1103/PhysRevE.92.012819 (2015).
8. Ostrom, E., Walker, J. & Gardner, R. Covenants with and without a sword: Self-governance is possible. *Am. Polit. Sci. Rev.* **86**, 404–417. https://doi.org/10.2307/1964229 (1992).
9. Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. The evolution of altruistic punishment. *Proc. Natl. Acad. Sci.* **100**, 3531–3535. https://doi.org/10.1073/pnas.0630443100 (2003).
10. Hua, S. & Liu, L. Facilitating the evolution of cooperation through altruistic punishment with adaptive feedback. *Chaos Solitons Fractals* **173**, 113669 (2023).
11. Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**, 980–994. https://doi.org/10.1257/aer.90.4.980 (2000).
12. Falk, A., Fehr, E. & Fischbacher, U. Driving forces behind informal sanctions. *Econometrica* **73**, 2017–2030. https://doi.org/10.1111/j.1468-0262.2005.00644.x (2005).
13. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140. https://doi.org/10.1038/415137a (2002).
14. Bone, J., Silva, A. S. & Raihani, N. J. Defectors, not norm violators, are punished by third-parties. *Biol. Lett.* **10**, 20140388. https://doi.org/10.1098/rsbl.2014.0388 (2014).
15. Horne, C. *The Rewards of Punishment: A Relational Theory of Norm Enforcement* (Stanford University Press, 2009).
16. Irwin, K. & Horne, C. A normative explanation of antisocial punishment. *Soc. Sci. Res.* **42**, 562–570. https://doi.org/10.1016/j.ssresearch.2012.10.004 (2013).
17. Horne, C. & Irwin, K. Metanorms and antisocial punishment. *Soc. Influ.* **11**, 7–21. https://doi.org/10.1080/15534510.2015.1132255 (2016).
18. Carpenter, J. P. & Matthews, P. H. Norm enforcement: Anger, indignation, or reciprocity?. *J. Eur. Econ. Assoc.* **10**, 555–572. https://doi.org/10.1111/j.1542-4774.2011.01059.x (2012).
19. Fischbacher, U., Gächter, S. & Fehr, E. Are people conditionally cooperative? Evidence from a public goods experiment. *Econ. Lett.* **71**, 397–404. https://doi.org/10.1016/S0165-1765(01)00394-9 (2001).
20. Kocher, M. G., Cherry, T., Kroll, S., Netzer, R. J. & Sutter, M. Conditional cooperation on three continents. *Econ. Lett.* **101**, 175–178. https://doi.org/10.1016/j.econlet.2008.07.015 (2008).
21. Chaudhuri, A. Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Exp. Econ.* **14**, 47–83. https://doi.org/10.1007/s10683-010-9257-1 (2011).
22. Fowler, J. H. & Christakis, N. A. Cooperative behavior cascades in human social networks. *Proc. Natl. Acad. Sci.* **107**, 5334–5338. https://doi.org/10.1073/pnas.0913149107 (2010).
23. Engel, C., Kube, S. & Kurschilgen, M. Managing expectations: How selective information affects cooperation and punishment in social dilemma games. *J. Econ. Behav. Organ.* **187**, 111–136. https://doi.org/10.1016/j.jebo.2021.04.029 (2021).
24. Clutton-Brock, T. H. & Parker, G. A. Punishment in animal societies. *Nature* **373**, 209–216. https://doi.org/10.1038/373209a0 (1995).

25. Gurerk, O., Irlenbusch, B. & Rockenbach, B. The competitive advantage of sanctioning institutions. *Science* **312**, 108–111. https://doi.org/10.1126/science.1123633 (2006).
26. Przepiorka, W. & Diekmann, A. Individual heterogeneity and costly punishment: A volunteer's dilemma. *Proc. R. Soc. B. Biol. Sci.* **280**, 20130247. https://doi.org/10.1098/rspb.2013.0247 (2013).
27. Carpenter, J. P. The demand for punishment. *J. Econ. Behav. Organ.* **62**, 522–542. https://doi.org/10.1016/j.jebo.2005.05.004 (2007).
28. Henrich, J. *et al.* Costly punishment across human societies. *Science* **312**, 1767–1770. https://doi.org/10.1126/science.1127333 (2006).
29. Cinyabuguma, M., Page, T. & Putterman, L. Can second-order punishment deter perverse punishment?. *Exp. Econ.* **9**, 265–279. https://doi.org/10.1007/s10683-006-9127-z (2006).
30. Pfattheicher, S., Keller, J. & Knezevic, G. Sadism, the intuitive system, and antisocial punishment in the public goods game. *Pers. Soc. Psychol. Bull.* **43**, 337–346. https://doi.org/10.1177/0146167216684134 (2017).
31. Herrmann, B., Thoni, C. & Gachter, S. Antisocial punishment across societies. *Science* **319**, 1362–1367. https://doi.org/10.1126/science.115380 (2008).
32. Sylwester, K., Herrmann, B. & Bryson, J. J. Homo homini lupus? Explaining antisocial punishment. *J. Neurosci. Psychol. Econ.* **6**, 167–188. https://doi.org/10.1037/npe0000009 (2013).
33. Monin, B. Holier than me? Threatening social comparison in the moral domain. *Int. Rev. Soc. Psychol.* **20**, 53–68 (2007).
34. Gächter, S. & Herrmann, B. Reciprocity, culture and human cooperation: Previous insights and a new cross-cultural experiment. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 791–806. https://doi.org/10.1098/rstb.2008.0275 (2009).
35. Pleasant, A. & Barclay, P. Why hate the good guy? Antisocial punishment of high cooperators is greater when people compete to be chosen. *Psychol. Sci.* **29**, 868–876. https://doi.org/10.1177/0956797617752642 (2018).
36. Willer, R., Kuwabara, K. & Macy, M. W. The false enforcement of unpopular norms. *Am. J. Sociol.* **115**, 451–490. https://doi.org/10.1086/599250 (2009).
37. Li, X., Molleman, L. & van Dolder, D. Do descriptive social norms drive peer punishment? Conditional punishment strategies and their impact on cooperation. *Evol. Hum. Behav.* **42**, 469–479. https://doi.org/10.1016/j.evolhumbehav.2021.04.002 (2021).
38. Erdfelder, E. *et al.* Multinomial processing tree models: A review of the literature. *Z. für Psychologie/J. Psychol.* **217**, 108–124 (2009).
39. Schmidt, O., Erdfelder, E. & Heck, D. W. How to develop, test, and extend multinomial processing tree models: A tutorial. *Psychol. Methods* https://doi.org/10.1037/met0000561 (2023).
40. Moshagen, M. multiTree: A computer program for the analysis of multinomial processing tree models. *Behav. Res. Methods* **42**, 42–54. https://doi.org/10.3758/BRM.42.1.42 (2010).
41. Castela, M., Kellen, D., Erdfelder, E. & Hilbig, B. E. The impact of subjective recognition experiences on recognition heuristic use: A multinomial processing tree approach. *Psychon. Bull. Rev.* **21**, 1131–1138. https://doi.org/10.3758/s13423-014-0587-4 (2014).
42. Klauer, K. C., Stahl, C. & Erdfelder, E. The abstract selection task: New data and an almost comprehensive model. *J. Exp. Psychol. Learn. Mem. Cogn.* **33**, 680–703. https://doi.org/10.1037/0278-7393.33.4.680 (2007).
43. Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R. & Hütter, M. Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *J. Pers. Soc. Psychol.* **113**, 343. https://doi.org/10.1037/pspa0000086 (2017).
44. Mieth, L., Buchner, A. & Bell, R. Moral labels increase cooperation and costly punishment in a Prisoner's Dilemma game with punishment option. *Sci. Rep.* **11**, 1–13. https://doi.org/10.1038/s41598-021-89675-6 (2021).
45. Mieth, L., Buchner, A. & Bell, R. Cognitive load decreases cooperation and moral punishment in a Prisoner's Dilemma game with punishment option. *Sci. Rep.* **11**, 1–12. https://doi.org/10.1038/s41598-021-04217-4 (2021).
46. Philippsen, A., Mieth, L., Buchner, A. & Bell, R. Communicating emotions, but not expressing them privately, reduces moral punishment in a Prisoner's Dilemma game. *Sci. Rep.* **13**, 14693 (2023).
47. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191. https://doi.org/10.3758/BF03193146 (2007).
48. Parks, C. D. & Stone, A. B. The desire to expel unselfish members from the group. *J. Pers. Soc. Psychol.* **99**, 303–310. https://doi.org/10.1037/a0018403 (2010).
49. Bell, R., Mieth, L. & Buchner, A. Separating conditional and unconditional cooperation in a sequential Prisoner's Dilemma game. *PLoS ONE* **12**, e0187952. https://doi.org/10.1371/journal.pone.0187952 (2017).
50. Wang, L., Zheng, J., Meng, L., Lu, Q. & Ma, Q. Ingroup favoritism or the black sheep effect: Perceived intentions modulate subjective responses to aggressive interactions. *Neurosci. Res.* **108**, 46–54. https://doi.org/10.1016/j.neures.2016.01.011 (2016).
51. Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E. & Cohen, J. D. The neural basis of economic decision-making in the ultimatum game. *Science* **300**, 1755–1758. https://doi.org/10.1126/science.1082976 (2003).
52. Barclay, P. Enhanced recognition of defectors depends on their rarity. *Cognition* **107**, 817–828. https://doi.org/10.1016/j.cognition.2007.11.013 (2008).
53. Bell, R., Buchner, A. & Musch, J. Enhanced old–new recognition and source memory for faces of cooperators and defectors in a social-dilemma game. *Cognition* **117**, 261–275. https://doi.org/10.1016/j.cognition.2010.08.020 (2010).
54. Volstorf, J., Rieskamp, J. & Stevens, J. R. The good, the bad, and the rare: Memory for partners in social interactions. *PLoS ONE* **6**, e18945. https://doi.org/10.1371/journal.pone.0018945 (2011).
55. Ma, D. S., Correll, J. & Wittenbrink, B. The Chicago face database: A free stimulus set of faces and norming data. *Behav. Res. Methods* **47**, 1122–1135. https://doi.org/10.3758/s13428-014-0532-5 (2015).
56. Riefer, D. M. & Batchelder, W. H. Multinomial modeling and the measurement of cognitive processes. *Psychol. Rev.* **95**, 318–339. https://doi.org/10.1037/0033-295X.95.3.318 (1988).
57. Kroneisen, M. & Steghaus, S. The influence of decision time on sensitivity for consequences, moral norms, and preferences for inaction: Time, moral judgments, and the CNI model. *J. Behav. Decis. Mak.* **34**, 140–153. https://doi.org/10.1002/bdm.2202 (2021).
58. Bayen, U. J., Murnane, K. & Erdfelder, E. Source discrimination, item detection, and multinomial models of source monitoring. *J. Exp. Psychol. Learn. Mem. Cogn.* **22**, 197–215. https://doi.org/10.1037/0278-7393.22.1.197 (1996).
59. Buchner, A., Erdfelder, E. & Vaterrodt-Plünnecke, B. Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *J. Exp. Psychol. Gen.* **124**, 137–160. https://doi.org/10.1037/0096-3445.124.2.137 (1995).
60. Menne, N. M., Winter, K., Bell, R. & Buchner, A. A validation of the two-high threshold eyewitness identification model by reanalyzing published data. *Sci. Rep.* **12**, 13379. https://doi.org/10.1038/s41598-022-17400-y (2022).
61. Batchelder, W. H. & Riefer, D. M. Multinomial processing models of source monitoring. *Psychol. Rev.* **97**, 548. https://doi.org/10.1037/0033-295X.97.4.548 (1990).
62. Erdfelder, E., Cüpper, L., Auer, T.-S. & Undorf, M. The four-states model of memory retrieval experiences. *Z. Psychol./J Psychol.* **215**, 61–71. https://doi.org/10.1027/0044-3409.215.1.61 (2007).
63. Son, J.-Y., Bhandari, A. & FeldmanHall, O. Crowdsourcing punishment: Individuals reference group preferences to inform their own punitive decisions. *Sci. Rep.* **9**, 1–15. https://doi.org/10.1038/s41598-019-48050-2 (2019).
64. FeldmanHall, O., Otto, A. R. & Phelps, E. A. Learning moral values: Another's desire to punish enhances one's own punitive behavior. *J. Exp. Psychol. Gen.* **147**, 1211–1224. https://doi.org/10.1037/xge0000405 (2018).
65. Suleiman, R. & Samid, Y. Punishment strategies across societies: Conventional wisdoms reconsidered. *Games* **12**, 63. https://doi.org/10.3390/g12030063 (2021).

66. Mieth, L., Bell, R. & Buchner, A. Facial likability and smiling enhance cooperation, but have no direct effect on moralistic punishment. *J. Exp. Psychol.* **63**, 263–277. https://doi.org/10.1027/1618-3169/a000338 (2016).
67. Mieth, L., Buchner, A. & Bell, R. Effects of gender on costly punishment. *J. Behav. Decis. Mak.* **30**, 899–912. https://doi.org/10.1002/bdm.2012 (2017).
68. Capraro, V., Jordan, J. J. & Tappin, B. M. Does observability amplify sensitivity to moral frames? Evaluating a reputation-based account of moral preferences. *J. Exp. Soc. Psychol.* **94**, 104103 (2021).
69. Capraro, V. & Perc, M. Mathematical foundations of moral preferences. *J. R. Soc. Interface* **18**, 20200880 (2021).
70. Peysakhovich, A. & Rand, D. G. Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Manag. Sci.* **62**, 631–647. https://doi.org/10.1287/mnsc.2015.2168 (2016).
71. Lindström, B., Jangard, S., Selbing, I. & Olsson, A. The role of a "common is moral" heuristic in the stability and change of moral norms. *J. Exp. Psychol. Gen.* **147**, 228–242. https://doi.org/10.1037/xge0000365 (2018).
72. Minson, J. A. & Monin, B. Do-gooder derogation: Disparaging morally motivated minorities to defuse anticipated reproach. *Soc. Psychol. Personal. Sci.* **3**, 200–207. https://doi.org/10.1177/1948550611415695 (2012).
73. Loughnan, S. & Piazza, J. in *Atlas of moral psychology* (eds Kurt Gray & Jesse Graham) 165–174 (2018).

Author contributions

A.P., L.M., A.B. and R.B. contributed to the study conception and design. Material preparation, data collection and analysis were performed by A.P. All authors contributed through discussion and interpretation of the results. A.P. wrote the manuscript with subsequent input and final approval from all co-authors.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024

Published Article 3

The article includes Experiments 3.1, 3.2 and 3.3.

Philippsen, A., Mieth, L., Buchner, A., & Bell, R. (2024). Time pressure and deliberation affect moral punishment. *Scientific Reports*, *14*(1), Article: 16378. https://doi.org/10.1038/s41598-024-67268-3

scientific reports

Check for updates

OPEN

Time pressure and deliberation affect moral punishment

Ana Philippsen✉, Laura Mieth, Axel Buchner & Raoul Bell

The deliberate-morality account implies that moral punishment should be decreased with time pressure and increased with deliberation while the intuitive-morality account predicts the opposite. In three experiments, moral punishment was examined in a simultaneous one-shot Prisoner's Dilemma game with a costly punishment option. The players cooperated or defected and then decided whether or not to punish their partners. In Experiment 1, the punishment decisions were made without or with time pressure. In Experiment 2, the punishment decisions were immediate or delayed by pauses in which participants deliberated their decisions. In Experiment 3, participants were asked to deliberate self-interest or fairness before deciding whether to punish their partners. Different types of punishment were distinguished using the cooperation-and-punishment model. In Experiment 1, time pressure decreased moral punishment. In Experiment 2, deliberation increased moral punishment. So far, the evidence supports the deliberate-morality account. Experiment 3 demonstrates that the effect of deliberation depends on what is deliberated. When participants deliberated self-interest rather than fairness, moral punishment was decreased. The results suggest that unguided deliberation increases moral punishment, but the effects of deliberation are modulated by the type of deliberation that takes place. These results strengthen a process-based account of punishment which offers a more nuanced understanding of the context-specific effect of deliberation on moral punishment than the deliberate-morality account.

Keywords Cooperation, Punishment, Cognitive resources, Deliberation, Multinomial processing tree model

Cooperation forms the foundation of successful societies[1]. While cooperation among kin is a common aspect of animal behavior, humans are special in their capacity to cooperate extensively among non-kin[2]. Humans even cooperate in anonymous one-shot interactions although they cannot expect direct reciprocity[3]. Given that cooperation implies accepting costs to help others, this raises the question: What factors contribute to promoting cooperation in one-shot interactions? One factor that helps to sustain cooperation is the punishment of defection coinciding with one's own cooperation, referred to as moral punishment[4,5]. Despite its potential costs, people reliably engage in moral punishment even in anonymous one-shot interactions[6]. Since moral punishment plays a critical role in promoting cooperation, it is important to understand what processes underlie this valuable behavior.

Two conflicting positions can be contrasted: The deliberate-morality account[7,8] implies that people intuitively act selfishly, therefore shying away from the potential costs of moral punishment. This natural tendency to avoid personal costs may stop them from engaging in moral punishment unless their intuitive tendency is overridden by deliberation. From these assumptions, one can derive the hypothesis that moral punishment should be *decreased* under time pressure and increased when deliberation is encouraged and sufficient time is available. In contrast, strong moral norms guide people's intuitive responses according to the intuitive-morality account[9,10] which implies that people's intuition is to morally punish others for refusing to cooperate. Only upon deliberation should they take into account the potential costs of this behavior which then causes them to suppress their natural tendency to morally punish others. From this account, one can derive the hypothesis that moral punishment should be *increased* under time pressure and decreased when deliberation is encouraged and sufficient time is available. In the present experiments, we test these conflicting accounts in a simultaneous one-shot Prisoner's Dilemma game. We manipulated whether the participants' decision to punish the partners had to be made under time pressure (Experiment 1) or after a delay in which participants were encouraged to deliberate their decisions (Experiments 2 and 3).

On a collective level, the best outcome is typically achieved when individuals cooperate with each other. However, when cooperating the individual bears costs. This clash of collective and individual interests is called a social dilemma[11]. To study human decision making in such dilemmas, researchers often rely on economic

Department of Experimental Psychology, Heinrich Heine University Düsseldorf, Universitätsstrasse 1, 40225 Düsseldorf, Germany. ✉email: Ana.Philippsen@hhu.de

games in which the complexities of the social dilemma are broken down to a simple payoff matrix. A well-established paradigm for studying cooperation is the Prisoner's Dilemma[12]. In the Prisoner's Dilemma, two players simultaneously decide to either cooperate or defect. Depending on their decisions, different outcomes arise, as is illustrated in Fig. 1: For both players, the highest collective outcome is achieved through mutual cooperation. The highest individual outcome, however, is achieved by unilaterally defecting on a cooperating partner who, in turn, receives the worst outcome of the game. For each individual player, there is thus always a financial incentive to defect regardless of what the other player does, while collectively mutual cooperation is better than mutual defection.

Free riders tend to act on the individual rather than the collective interest and exploit others' cooperation. If too many people free ride, cooperation loses its appeal and declines to alarmingly low levels[13–15]. Here, moral punishment comes into play. The punishment of defectors offers a solution to the free-rider problem as it counters the incentive to defect, thereby effectively enforcing cooperation[16–18]. People readily sacrifice their own resources in social dilemma games to punish others for defection—even in anonymous one-shot interactions where they cannot build a reputation or coerce others into cooperation[19–22]. Due to its crucial role for promoting cooperation, the punishment of unilateral defection is typically interpreted as a behavior that enforces moral norms[16,23] which is why the punishment of defection by cooperating players is referred to as *moral punishment*[4,5].

Behaviors such as cooperation, moral punishment or telling the truth are considered prosocial in that they support the collective good at the expense of personal cost. Considerable research has been devoted to the question of whether such behaviors occur deliberately or intuitively [cf. 24]. This distinction between intuitive and deliberate processes lies at the core of dual-processing theories that conceptualize human behavior as an interplay of two different processing types which depend on the cognitive resources available in the situation[25–27]. Type-I processing defines default reactions that arise rapidly and automatically in situations with limited resources. Behaviors based on Type-I-processing are therefore qualified as intuitive reactions. Type-II-processing, in contrast, can overrule intuitive Type-1 processing and guide behavior in a deliberate way, given sufficient time and cognitive resources. To stimulate behavior that relies on intuition rather than deliberation, researchers may impose time pressure or cognitive load. Alternatively, researchers may stimulate behavior that relies on deliberation by requiring decisions to be made after a delay in which participants are encouraged to deliberate their decisions.

Research on how intuition and deliberation affect moral decision-making has yielded mixed results. For instance, there are diverging findings regarding the question of whether intuition or deliberation leads people to tell the truth despite incentives to lie[28–31]. Consistent with this broader literature on moral decision-making, the question of whether moral punishment relies on intuition or deliberation has also produced inconsistent results. This question has as yet been mainly addressed by examining people's behavior in the Ultimatum Game. In the Ultimatum Game, one player, the proposer, is endowed with a certain amount of money and is asked how much of that money they want to offer to the other player, the responder, and how much they want to keep for themselves. The responder can then decide to accept the offer, leading to the shares being paid out according to the proposer's offer, or to reject the offer in which case neither player receives any money. As rejecting an offer entails sacrificing own money to ensure that the proposer does not receive an unfair share, rejection in the Ultimatum Game is often interpreted as a form of moral punishment[32–34]. Rejection rates in the Ultimatum Game were found to be increased with restricted cognitive resources[35–38] and decreased with deliberation during a time delay[39–42]. These findings favor the idea that moral punishment relies on intuition rather than deliberation and thus support the *intuitive-morality account of punishment*. In other studies, however, these results were not replicated[43–48] or the opposite pattern was found with decreased rejection rates under cognitive load[49,50] and increased rejection rates with a time delay[51]. These latter results corroborate the *deliberate-morality account of punishment* which is further supported by neuro-imaging studies indicating that the application of punishment relies on areas of cognitive control[52–54]. In sum, the pertinent findings involving the Ultimatum Game are inconsistent. The interpretation of these findings is further complicated by the fact that the Ultimatum Game does not allow to clearly distinguish the participants' inclination to punish from their inclination to cooperate as both are intertwined in one decision [accept or reject], cf.[5,55]. It is thus interesting to test the effect of time pressure and deliberation on moral punishment in a paradigm that allows to more precisely differentiate between cooperation and punishment.

One such paradigm is the Prisoner's Dilemma game with costly punishment option[5,56–58]. In the Prisoner's Dilemma game, both players decide simultaneously whether to cooperate or to defect. Following this decision, they are informed about the outcome of the Prisoner's Dilemma game. They then decide whether or not to punish their partners by investing some of their own money to deduct money from the partner's account. To clearly

A \ B	cooperates	defects
cooperates	+10 / +10	-10 / +20
defects	+20 / -10	0 / 0

Figure 1. Payoffs in the Prisoner's Dilemma game as a function of both players' decisions. Shaded cells mark the decision and payoff of Player A, white cells mark the decision and payoff of Player B.

separate the decision processes underlying cooperation and punishment a multinomial processing tree (MPT) model was used to analyze the present data. MPT models are useful tools that serve to disambiguate observable responses by decomposing them into different underlying latent processes[59,60]. Easy-to-read tutorials[61] and user-friendly software[62] have facilitated the application of these models in a variety of fields[60], including moral judgements and decision making[63–70].

The multinomial cooperation-and-punishment model (see Fig. 2) serves to separately measure cooperation, moral punishment, hypocritical punishment, antisocial punishment and a punishment bias[5,56–58]. According to the model, a participant decides to cooperate with probability C or to defect with probability $1 - C$. In a simultaneous Prisoner's Dilemma game, the partner's behavior is revealed only after the participant has made their decision to cooperate or to defect. When the participant decides whether they want to cooperate or to defect, the participant thus cannot know whether they interact with a defecting or cooperating partner. The model therefore implies that parameter C does not differ as a function of the behavior of the partner. The P. parameters refer to the conditional probabilities of different types of punishment that are specifically elicited by, and thereby contingent upon, the outcomes of the Prisoner's Dilemma game. These can be contrasted with an unspecific punishment bias b that is assumed to be unaffected by the outcome of the Prisoner's Dilemma game. To illustrate, if a cooperating participant interacts with a defecting partner, the participant may apply *moral punishment* with the conditional probability P_{Moral}. If moral punishment is not applied which occurs with the conditional probability $1 - P_{\text{Moral}}$, the participant may still punish the partner due to an unspecific punishment bias with the conditional probability b. With the conditional probability $1-b$, no punishment is applied. If a defecting participant interacts with a defecting partner, the participant may apply *hypocritical punishment* with the conditional probability $P_{\text{Hypocritical}}$. This type of punishment can be considered hypocritical as it enforces a norm of cooperation the participant themselves failed to follow. If no hypocritical punishment is applied which occurs with the conditional probability $1 - P_{\text{Hypocritical}}$, the participant may still punish the partner due to the unspecific punishment bias with the conditional probability b. With the conditional probability $1 - b$, no punishment is applied. If a defecting participant interacts with a cooperating partner (lower tree of Fig. 2), the participant may apply *antisocial punishment* with the conditional probability $P_{\text{Antisocial}}$. This type of punishment is termed antisocial because it directly opposes the cooperation norm. If no antisocial punishment is applied which occurs with the conditional probability $1 - P_{\text{Antisocial}}$, the participant may still punish the partner due to the unspecific punishment bias with the conditional probability b. With the conditional probability $1 - b$, no punishment is applied. Mutual cooperation does not provide any specific reason to punish the partner. The punishment that still occurs in this condition is thus assumed to be caused only by the punishment bias b, representing an unspecific tendency to punish the

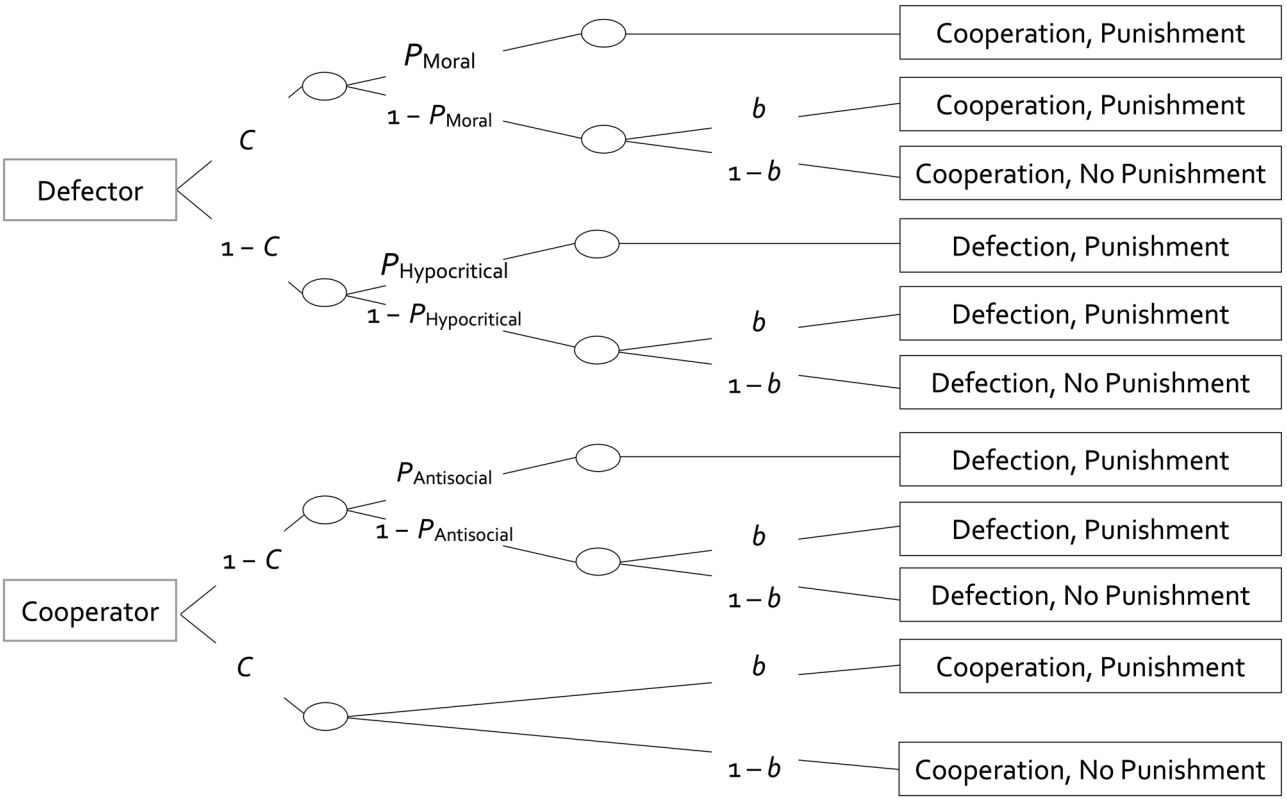

Figure 2. Graphical illustration of the cooperation-and-punishment model. The rectangles on the left represent the two types of partners in the Prisoner's Dilemma game (defector or cooperator). The rectangles on the right represent the participants' observable responses (cooperation or defection; punishment or no punishment). The letters along the branches represent the parameters for cooperation (C), moral, hypocritical and antisocial punishment (P_{Moral}, $P_{\text{Hypocritical}}$ and $P_{\text{Antisocial}}$, respectively) and the punishment bias (b).

partner irrespective of the outcome of the Prisoner's Dilemma game[5,56–58]. For example, cognitive load has been demonstrated to increase participants' inclination to indiscriminately punish partners in the Prisoner's Dilemma game[5], highlighting the necessity of accounting for bias when analyzing the punishment data, especially when manipulating the availability of cognitive resources. The concept of the punishment bias b is parallel to how response bias is taken into account in other multinomial decision-making models[71–75].

Here, a note of caution is in order: While we use the adjectives "moral," "hypocritical," and "antisocial" to refer to the punishment parameters that are clearly defined in the cooperation-and-punishment model, these labels should not be overinterpreted, particularly not against the backdrop of how these adjectives are used in everyday language. For instance, "moral" punishment may sometimes be influenced by self-interested motivations such as seeking retribution or reputation building. Furthermore, in everyday language "hypocritical" punishment may also encompass moral or antisocial motivations. Here, these adjectives merely serve as easily accessible descriptors for the parameters to simplify communication across disciplines and are not intended as exhaustive definitions of the parameters. The strength of the cooperation-and-punishment model lies in its precise structural definitions of parameters, as illustrated in Fig. 2, which remain valid regardless of the specific adjectives used as parameter labels. Therefore, the applicability of the model is not dependent on the adjectives used as verbal labels of the parameters.

The validity of the cooperation-and-punishment model has been demonstrated in various studies in which the model has been successfully used to separately measure cooperation, the different types of punishment and the punishment bias[5,56–58]. In one of these studies, Mieth et al.[5] have restricted participants' cognitive resources in the Prisoner's Dilemma game using a concurrent distractor task. This induction of cognitive load decreased moral punishment compared to a control group without distraction. The effect was specific to moral punishment. Hypocritical and antisocial punishment remained unaffected by cognitive load. The unspecific punishment bias was increased under cognitive load, suggesting that punishment was applied less purposefully when distracted by another task. These findings show that an increased availability of cognitive resources causes participants to apply moral punishment to enforce norms of cooperation, in support of the general idea that the moral use of punishment is facilitated by deliberation. This deliberate-morality interpretation implies that time pressure, like cognitive load, should specifically decrease moral punishment. Conversely, a delay during which participants are encouraged to deliberate their decisions should have the opposite effect, thereby increasing moral punishment. However, manipulations of cognitive load and time pressure have not always produced convergent results[24]. The aim of the present series of experiments was thus to dissect the effects of time pressure and deliberation on costly punishment in the Prisoner's Dilemma game. If deliberation causes punishment to be applied in a purposeful moral fashion, moral punishment should be decreased with time pressure (Experiment 1) and increased when deliberation is encouraged (Experiment 2). Furthermore, the deliberation manipulation is extended in the final experiment in which we tested whether the effect of deliberation depends on *what* is deliberated (Experiment 3): One group of participants was encouraged to deliberate self-interest while the other group was encouraged to deliberate fairness. This final experiment was performed to challenge the dichotomy that lies at the core of the dual-processes models. Specifically, the experiment served to test whether fairness-focused deliberation would favor moral punishment relative to self-interest-focused deliberation, consistent with a more nuanced process-based account according to which the effect of deliberation depends on the specific processes involved[52,76].

Experiment 1

Imposing a time constraint is a classical method to manipulate the availability of cognitive resources[10,77,78]. In line with this established approach for suppressing deliberation[24], participants had to make punishment decisions either without or with time pressure. Following the deliberate-morality account of punishment and the assumption that time pressure suppresses deliberation[7], moral punishment should be decreased in the condition with time pressure relative to the condition without time pressure. By contrast, the intuitive-morality account of punishment[9] implies moral punishment to be increased by time pressure.

Method

Participants and design

A total of 217 participants took part in the online study that was conducted via the online platform SoSci Survey[79]. Participants were recruited via a mailing list to which people could subscribe if they wished to participate in psychology experiments and with the help of social media channels directed at Heinrich Heine University students. Data had to be excluded from analyses for the following reasons: The data were not stored properly ($n = 6$), the participant was younger than 18 and thus could not legally consent to participate ($n = 1$), the participant stated to have poor eyesight ($n = 1$) or the participant withdrew their consent to the use of their data at the end of the study ($n = 3$). The final sample consisted of $N = 206$ participants (154 female, 51 male, 1 non-binary) who were between 18 and 66 (*mean age* = 25, *standard deviation* = 10) years old. Of these, 102 were assigned to the condition without time pressure and 104 participants were assigned to the condition with time pressure. The median duration of participation was 10 min. Undergraduate psychology students could receive course credit for their participation. Other students or non-students had the chance to win a shopping voucher worth 20 €; these participants knew that only one voucher was available which would be awarded through lottery after data collection had been completed. A sensitivity analysis with G*Power[80] showed that, with $\alpha = 0.05$, $N = 206$ participants and 20 behavioral choices in the Prisoner's Dilemma game, effects of time pressure on the different types of punishment as small as $w = 0.06$ could be detected with a statistical power of $1 - \beta = 0.95$.

Ethics

The present series of experiments was approved by the Ethics Committee of the Faculty of Mathematics and Natural Sciences of the Heinrich-Heine-University Düsseldorf and conducted in accordance with the requirements of the Declaration of Helsinki. All participants were informed that they would play a game involving a number of interactions with partners who would simultaneously make the same decision as themselves and that the purpose of the study was to gain insight into people's behavior in interactions. They then gave written informed consent before participating in the experiment. At the end of each experiment, participants were debriefed that the purpose of the experiment was to study how decision time affects cooperation and punishment. They were informed that they had interacted with programmed partners during the experiment and were then reminded that they could still withdraw their consent to the use of their data.

Prisoner's Dilemma

Participants played the Prisoner's Dilemma game with a costly punishment option which has been used in several previous studies to study cooperation and punishment[5,56–58,81,82]. At the start of the Prisoner's Dilemma game, participants were endowed with 400 cents (4 €). They were informed that they would receive an online-shopping voucher equivalent to the amount of money they had in their account balance at the end of the experiment (347 cents on average, *standard deviation* = 55). Each participant played 20 rounds of the Prisoner's Dilemma game with a costly punishment option with 20 different partners. Half of the partners cooperated and the other half defected. Each participant saw the partners in a different, randomly determined order.

Before each trial started, the participant was informed about their current account balance. The participant started the trial by clicking a "Continue" button. To emphasize the social nature of the game, the participant saw a facial photograph of their interaction partner. The picture was randomly drawn from 10 male and 10 female faces (between 18 and 40 years old) of the Chicago Face Database[83]. The face of the partner was shown from a frontal view with a neutral expression. The picture had a resolution of 266 × 186 pixels. The participant was asked "Do you want to cooperate or defect?" and answered by selecting either "I cooperate" or "I defect". The participant was then presented with a summary of the interaction. The participant had previously been instructed that the partner made the decision on whether to cooperate or to defect at the same time as the participant. The participant received feedback about their own decision (e.g., "You cooperate.") and the partner's decision (e.g., "Your partner defects"), and how these decisions affected the participant's account balance (e.g., "You lose 10 cents.") and the partner's account balance (e.g., "Your partner gains 20 cents."). The feedback regarding the participant's decision and outcome was displayed in black, while the feedback regarding the partner's decision and outcome was displayed in blue, corresponding to the blue frame surrounding the partner's photograph (see Fig. 3). The face of the interaction partner and the feedback remained visible until after the end of each trial.

Mutual cooperation resulted in a moderate gain (+ 10 cents) for both partners. Mutual defection resulted in neither a gain nor a loss (0 cents) for each partner. In the case of unilateral cooperation, the defecting partner made a large profit (+ 20 cents) at the expense of the cooperating partner who lost money (− 10 cents). The payoff structure thus corresponds to that of a typical Prisoner's Dilemma game in that there was a high temptation for unilateral defection, a moderate reward for mutual cooperation, no reward after mutual defection and a loss when cooperating with a defecting partner[11].

Costly punishment

Immediately after the interaction in the Prisoner's Dilemma game, the participant had the option to punish the partner. The participant could decide either not to punish their partner or to invest 1, 2, 3, 4 or 5 cents to deduct 10, 20, 30, 40 or 50 cents, respectively, from their partner's account balance, as depicted in Fig. 3. They were then automatically forwarded to the next screen which displayed the participant's investment in punishment in black, the resulting punishment for their partner in blue as well as the partner's investment in punishment in blue and the resulting punishment for the participant in black. The partners were programmed to morally punish unilateral defection of the participants with a random amount of 10, 20, 30, 40 or 50 cents. This mimics the behavior of real participants who primarily use the punishment option to punish unilateral defection[6,19–21]. Participants then initiated the next round of the game by clicking a "Continue" button.

Time-pressure manipulation

Participants were randomly assigned either to the condition without time pressure or to the condition with time pressure. In the condition without time pressure, the participant was encouraged in the instructions and before each interaction to take their time and to deliberate carefully how they wanted to respond. The participant had unlimited time both when deciding whether to cooperate and when deciding whether to punish the partner. In the condition without time pressure, the median response time was 3.6 s for the cooperation decision and 3.5 s for the punishment decision.

In the condition with time pressure, the participant was informed in the instructions and before each interaction to decide quickly whether to cooperate and whether to punish because there would be a time limit of five seconds within which either response had to be made. A countdown from 5 to 0 s was presented until each of these responses had to be made or until the countdown reached 0 s. When participants did not respond within the five seconds, a warning was displayed, asking the participant to respond more quickly. Participants then had to click on a "Continue" button to repeat the trial. In the Prisoner's Dilemma game, 92% of the 104 participants in the condition with time pressure never exceeded the time limit and another 6% exceeded the time limit only once. The summary of the Prisoner's Dilemma game interaction was presented with a countdown of five seconds and participants were automatically forwarded to the punishment option when they did not click the "Continue" button within the time limit. When making the punishment decision, 84% of the 104 participants never exceeded

You cooperate.

Your partner defects.

You lose 10 cents.

Your partner gains 20 cents.

How high should the punishment for your partner be?

- My partner is not to be punished.
- I invest 1 cent to deduct 10 cents from my partner's account balance.
- I invest 2 cents to deduct 20 cents from my partner's account balance.
- I invest 3 cents to deduct 30 cents from my partner's account balance.
- I invest 4 cents to deduct 40 cents from my partner's account balance.
- I invest 5 cents to deduct 50 cents from my partner's account balance.

Figure 3. Example trial of the Prisoner's Dilemma with punishment option. In this example trial, the participant cooperated and the partner defected. Therefore, the participant lost 10 cents and the partner gained 20 cents. The participant then decided to punish their partner by investing 3 cents to deduct 30 cents from the partner's account balance. The facial photograph of the partner was randomly selected from a set of 10 male and 10 female faces taken from the Chicago Face Database[83]. Informed consent to publish the figure in an online open-access publication has been obtained.

the time limit and another 13% exceeded the time limit only once. In the condition with time pressure, the median response time was 2.4 s for the cooperation and 2.4 s for the punishment decision.

Results

The first four trials had originally been designed as practice trials. Upon a reviewer's suggestion, these practice trials are now included in the analyses. Whether or not the practice trials are included has no effect on the statistical conclusions in any experiment reported here except that the bias parameter b differs between the two conditions in Experiment 3 when the practice trials are included.

For the model-based analyses, the α level was set to 0.05. Parameter estimates and goodness-of-fit tests were obtained using *multitree*[62]. To analyze the present data, two instances of the model depicted in Fig. 2 are needed, one for the condition without time pressure and one for the condition with time pressure. The base model fit the data, $G^2(2) = 0.18$, $p = 0.913$, indicating that the base model's parameters reflect the observed data adequately. Figure 4 displays the estimates of the cooperation parameter (left panel), the punishment parameters (middle panel) and the punishment bias (right panel).

Multinomial models make it possible to test hypotheses directly at the level of the model parameters, that is, at the level of the cognitive processes measured by the model. For example, the hypothesis that cooperation is more likely in the condition without time pressure than in the condition with time pressure can be tested by

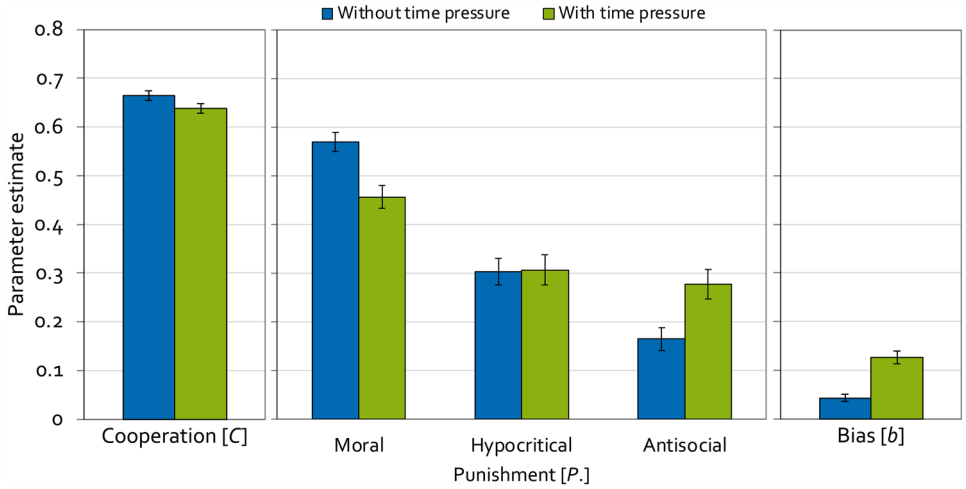

Figure 4. Estimates of the parameters of the cooperation-and-punishment model depending on whether decisions were made without or with time pressure in Experiment 1. Parameter C represents the probability of cooperation. Parameters P_{Moral}, $P_{\text{Hypocritical}}$ and $P_{\text{Antisocial}}$ represent the conditional probabilities of moral punishment, hypocritical punishment and antisocial punishment, respectively. Parameter b represents the punishment bias, that is, the probability of punishment irrespective of the outcome of the Prisoner's Dilemma game. The error bars represent the standard errors.

restricting parameter C to be equal in the two conditions. If the restricted model fits the data significantly worse than the base model, indicated by the ΔG^2 statistic that is chi-square distributed with degrees of freedom given in parentheses, then the null hypothesis that parameter C does not differ between the two conditions needs to be rejected. As a consequence, it can be concluded that cooperation differs between the two conditions[60]. In fact, the cooperation parameter could be equated across conditions without causing a significant decrease in model fit, $\Delta G^2(1) = 3.24$, $p = 0.072$, $w = 0.03$. It thus needs to be concluded that cooperation was unaffected by time pressure.

The central hypothesis test in Experiment 1 concerns the parameter representing the probability of moral punishment. In line with the deliberate-morality account of punishment, moral punishment was significantly less likely in the condition with time pressure than in the condition without time pressure, $\Delta G^2(1) = 13.61$, $p < 0.001$, $w = 0.06$. As a secondary finding, the probability of hypocritical punishment did not differ as a function of time pressure, $\Delta G^2(1) = 0.01$, $p = 0.936$, $w < 0.01$. Antisocial punishment was more likely in the condition with time pressure compared to the condition without time pressure, $\Delta G^2(1) = 8.37$, $p = 0.004$, $w = 0.05$. Furthermore, time pressure led to an increase in the punishment bias, $\Delta G^2(1) = 31.40$, $p < 0.001$, $w = 0.09$, compared to the condition without time pressure.

Discussion

Experiment 1 served to examine the effects of time pressure on moral punishment. Moral punishment was significantly decreased when punishment decisions were made with time pressure compared to when they were made without time pressure. This finding supports the deliberate-morality account of punishment according to which time pressure interferes with the moral use of punishment with the social goal of promoting cooperation[7].

The model-based analysis shows that the suppressive effect of time pressure on punishment was specific to moral punishment. Hypocritical punishment remained unaffected by time pressure. Antisocial punishment increased with time pressure. Furthermore, time pressure significantly increased participants' bias to indiscriminately punish irrespective of the outcome of the Prisoner's Dilemma interaction, in line with the prediction that time pressure causes punishment to be applied less purposefully[5].

It seems noteworthy that the effects of time pressure on punishment observed here are strikingly parallel to the effects of cognitive load on punishment reported by Mieth et al.[5]. Specifically, Mieth et al. have observed that cognitive load decreases moral punishment but has no effect on hypocritical punishment. Antisocial punishment was descriptively, but not significantly, increased in the condition with cognitive load compared to the condition without cognitive load. Furthermore, cognitive load induced an increase in the punishment bias, supporting the idea that the reduced availability of cognitive resources causes punishment to be applied less purposefully. Together with the previous findings[5], the present findings suggest that manipulations suppressing deliberation have consistent effects on punishment, regardless of whether the suppression of deliberation is caused by cognitive load or time pressure.

Experiment 2

In Experiment 2, the goal was to extend the empirical basis of the previous findings by testing whether encouraging deliberation has effects on moral punishment opposite to those of suppressing deliberation. The deliberate-morality view of punishment[7] leads to the prediction not only that moral punishment should become less likely with time pressure but also that moral punishment should become more likely with the time spent deliberating the punishment decision. To test this prediction, punishment decisions were delayed and participants were

encouraged to deliberate their punishment decision during the delay in Experiment 2. This condition with deliberation was contrasted to a condition without deliberation in which punishment was not delayed and participants received no instructions encouraging them to deliberate their punishment decision.

Method

Participants and design

A total of 646 participants took part in the online study that was conducted via the platform SoSci Survey[79]. Participants were recruited by the panel provider *mingle* (https://mingle.respondi.de). Data had to be excluded from analyses for the following reasons: The data were not stored properly ($n = 9$), the participant took part repeatedly ($n = 47$), the participant dropped out prematurely ($n = 78$) or the participant withdrew their consent to the use of their data at the end of the study ($n = 4$). The final sample consisted of $N = 508$ participants (189 female, 318 male, 1 non-binary) who were between 19 and 69 (*mean age* = 40, *standard deviation* = 12) years old. Of these, 258 participants were assigned to the condition without deliberation and 250 were assigned to the condition with deliberation. The median duration of participation was 16 min. Participants received their final account balance achieved in the game in addition to their usual compensation by the panel provider (see explanation below). A sensitivity analysis with G*Power[80] showed that, with $\alpha = 0.05$, $N = 508$ participants and 20 behavioral choices in the Prisoner's Dilemma game, effects of deliberation on punishment as small as $w = 0.04$ could be detected with a statistical power of $1 - \beta = 0.95$.

Materials and procedure

The materials and procedure were the same as those of Experiment 1 with the following exceptions. Participants of the panel provider *mingle* are used to being compensated with points that can be exchanged for online vouchers, charity donations or money. Participants were therefore informed in the experiment-specific instructions that they played for points (with 1 point corresponding to 1 Euro cent) that they would receive by *mingle* in addition to the points which they already knew they would receive as a compensation for participating from the invitation e-mail they had received by mingle prior to participating. To align their starting endowment with *mingle's* common compensation fees, participants were endowed with 80 points at the start of the game. The costly punishment option was adjusted to the lower starting endowment to avoid negative account balances. In each trial, participants could thus invest up to three points to deduct a maximum of 30 points from their partner's account balance. On average, participants achieved a final account balance of 63 points (*standard deviation* = 29).

Manipulation of deliberation

Participants were randomly assigned to the condition without deliberation or to the condition with deliberation. In the condition without deliberation, the participant decided (self-paced) whether to punish the partner right after the Prisoner's Dilemma interaction had been completed. The median response time for the cooperation decision in this condition was 2.9 s and the median response time for the punishment decision was 2.8 s.

In the condition with deliberation, the cooperation decision was made without a delay. The median response time for the cooperation decision was 3.0 s. By contrast, the punishment decision was delayed by 30 s. The participant was instructed beforehand that, following the decision in the Prisoner's Dilemma game, the participant would be given time to deliberate the punishment decision. The participant was instructed: "Please take your time to carefully deliberate on what you want to do next." Right after each cooperation decision in the Prisoner's Dilemma game, the punishment option was presented, but it was accompanied by the instruction to deliberate on whether, and if so, how to punish the partner. The punishment option was initially deactivated. Beneath the punishment option, a countdown from 30 to 0 s was presented. After 30 s, the instruction to deliberate the punishment decision was replaced by the question "How high should the punishment for your partner be?" and the punishment option was activated so that the participant could implement their punishment decision. Including the 30-s delay, the median response time for the punishment decision in the condition with deliberation was 37.4 s.

Results

To analyze the present results, two instances of the model depicted in Fig. 2 are needed, one for the condition without deliberation and one for the condition with deliberation. The base model fit the data, $G^2(2) = 1.67$, $p = 0.434$, indicating that the base model's parameters reflect the observed data adequately. Figure 5 displays the estimates of the cooperation parameter (left panel), the punishment parameters (middle panel) and the punishment bias (right panel). The probability to cooperate was significantly lower in the condition with deliberation compared to the condition without deliberation, $\Delta G^2(1) = 16.28$, $p < 0.001$, $w = 0.04$.

The central hypothesis test again concerns the parameter representing the probability of moral punishment. In line with the deliberate-morality account of punishment, moral punishment was significantly more likely in the condition with deliberation compared to the condition without deliberation, $\Delta G^2(1) = 14.64$, $p < 0.001$, $w = 0.04$. As a secondary finding, hypocritical punishment was also more likely in the condition with deliberation compared to the condition without deliberation, $\Delta G^2(1) = 6.45$, $p = 0.011$, $w = 0.03$. Neither antisocial punishment, $\Delta G^2(1) = 0.40$, $p = 0.529$, $w = 0.01$, nor the punishment bias, $\Delta G^2(1) = 0.14$, $p = 0.711$, $w < 0.01$, were affected by the deliberation manipulation.

Discussion

The aim of Experiment 2 was to extend the results of the previous experiment which demonstrated that time pressure decreased moral punishment. From the deliberate-morality account, the prediction can be derived that a time delay during which one is encouraged to deliberate one's punishment decision should have an effect on

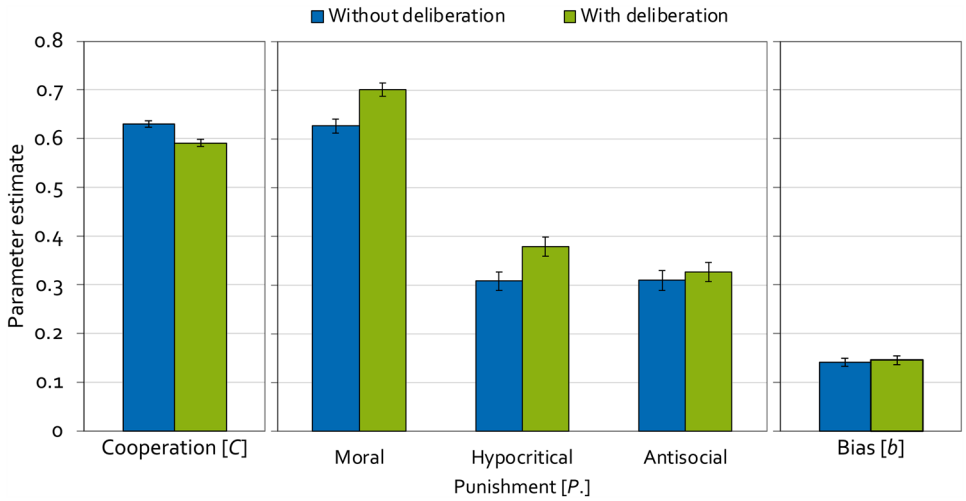

Figure 5. Estimates of the parameters of the cooperation-and-punishment model as a function of condition (without deliberation, with deliberation) in Experiment 2. Parameter C represents the probability of cooperation. Parameters P_{Moral}, $P_{Hypocritical}$ and $P_{Antisocial}$ represent the conditional probabilities of moral punishment, hypocritical punishment and antisocial punishment, respectively. Parameter b represents the punishment bias, that is, the probability of punishment irrespective of the outcome of the Prisoner's Dilemma game. The error bars represent the standard errors.

moral punishment that is opposite to the effect of time pressure. In line with this prediction, deliberation significantly increased moral punishment compared to a condition without deliberation, providing further support for the idea that deliberation favors the moral use of punishment.

Whereas the time-pressure manipulation of Experiment 1 did not affect hypocritical punishment, deliberating one's punishment decision in Experiment 2 increased hypocritical punishment relative to the condition without deliberation. Furthermore, deliberation had no effect on antisocial punishment or the punishment bias. These findings thus suggest that the lack of resources, caused by cognitive load and time pressure, affects decision-making in ways that are not simply the mirror image of deliberation.

While the decrease of cooperation in the condition with deliberation in comparison to the condition without deliberation at first glance seems to support an intuitive-morality view on cooperation[10], it is important to note that the manipulation of deliberation in Experiment 2 consisted primarily of delaying the punishment decisions. Decisions to cooperate or defect were not delayed. While the manipulation also included an instruction to "please take your time to carefully deliberate on what you want to do next", this instruction referred explicitly to the punishment and not to the cooperation decision. However, it cannot be ruled out that these instructions as well as deliberating about the punishment decisions may have caused participants to generally adopt a more deliberate processing mode, explaining why the manipulation also affected their propensity to cooperate. Note that this is only a post-hoc interpretation that requires confirmation in future studies before firm conclusions can be drawn about this issue.

Experiment 3

In the previous two experiments, we found evidence in favor of a deliberate-morality account of punishment. While this is in line with the results of some studies[49,50], it is in opposition to others that suggest punishment is intuitive rather than deliberate[35–37]. Such inconsistencies have motivated Declerck and Boone[52] to move away from the strict dichotomy of intuitive and deliberate moral behaviors. Instead, they proposed a process-based account according to which intuition and deliberation can either favor or inhibit moral behaviors, depending on contextual factors. For instance, according to this account, unguided deliberation in Experiment 2 may have increased moral punishment because participants were more likely to deliberate fairness than to deliberate self-interest. However, this does not mean that deliberation will always favor moral behaviors as the effect of deliberation may depend on *what* is deliberated which, in turn, may depend on the general context in which deliberation takes place. If participants are encouraged to deliberate moral concerns, this should have an enhancing effect on moral punishment. If participants are encouraged to deliberate self-interest, this should have a diminishing effect on moral punishment. In Experiment 3, this hypothesis was put to an empirical test. The punishment decision was delayed by a pause in which participants were explicitly encouraged to deliberate either fairness or self-interest, depending on the experimental condition they were assigned to. Instructions to deliberate fairness should increase moral punishment relative to instructions to deliberate self-interest.

Method

Participants and design

A total of 698 participants took part in the online study that was conducted via the platform SoSci Survey[79]. As in Experiment 2, participants were recruited by the panel provider *mingle*. Data had to be excluded from analyses for the following reasons: The participant took part repeatedly ($n = 45$), the participant dropped out prematurely ($n = 124$) or the participant did not consent to the use of their data at the end of the study ($n = 2$). The final sample consisted of $N = 527$ participants (206 female, 320 male, 1 non-binary) who were between 18 and 70 (*mean age* = 45, *standard deviation* = 14) years old. Of these, 255 participants were assigned to the self-interest-deliberation condition and 272 were assigned to the fairness-deliberation condition. The median duration of participation was 22 min. As in Experiment 2, participants received their final account balance in the game in addition to their usual compensation by the panel provider. A sensitivity analysis with G*Power[80] showed that, with $\alpha = 0.05$, $N = 527$ participants and 20 behavioral choices in the Prisoner's Dilemma game, effects of deliberation on punishment as small as $w = 0.04$ could be detected with a statistical power of $1 - \beta = 0.95$.

Materials, procedure and manipulation

The materials were the same as those of Experiments 1 and 2. The experimental procedure in Experiment 3 corresponded to the condition with deliberation of Experiment 2. The cooperation decisions were not delayed. The median response times for the cooperation decision were 4.5 s in the self-interest-deliberation condition and 4.7 s in the fairness-deliberation condition. By contrast, the punishment decision was delayed by a 30-s pause in which the participant was asked to deliberate the punishment decision. In Experiment 3, however, the specific content on which to deliberate was manipulated. In the self-interest-deliberation condition, the participant was instructed to deliberate "whether it is profitable to punish your partner and what a punishment means for your own account balance". The median response time for the punishment in the self-interest-deliberation condition was 38.0 s. In the fairness-deliberation condition, the participant was instructed to deliberate "how fair or unfair your partner's behavior was and whether your partner deserves a punishment for this behavior". The median response time for the punishment in the fairness-deliberation condition was 38.4 s. The same instruction was also included in the written instructions provided before the start of the Prisoner's Dilemma game. The remaining instructions and procedure were identical to those of Experiment 2 with the exception that the starting endowment was increased to 100 points (corresponding to 100 cents or 1 €). On average, participants achieved a final account balance of 85 points (*standard deviation* = 36).

Results

To analyze the present results, two instances of the model depicted in Fig. 2 are needed, one for the self-interest-deliberation condition and one for the fairness-deliberation condition. The base model fit the data, $G^2(2) = 0.36$, $p = 0.837$, indicating that the base model's parameters reflect the observed data adequately. Figure 6 displays the estimates of the cooperation parameter (left panel), the punishment parameters (middle panel) and the punishment bias (right panel). The probability to cooperate was significantly lower when participants deliberated self-interest compared to when they deliberated fairness, $\Delta G^2(1) = 35.17$, $p < 0.001$, $w = 0.06$.

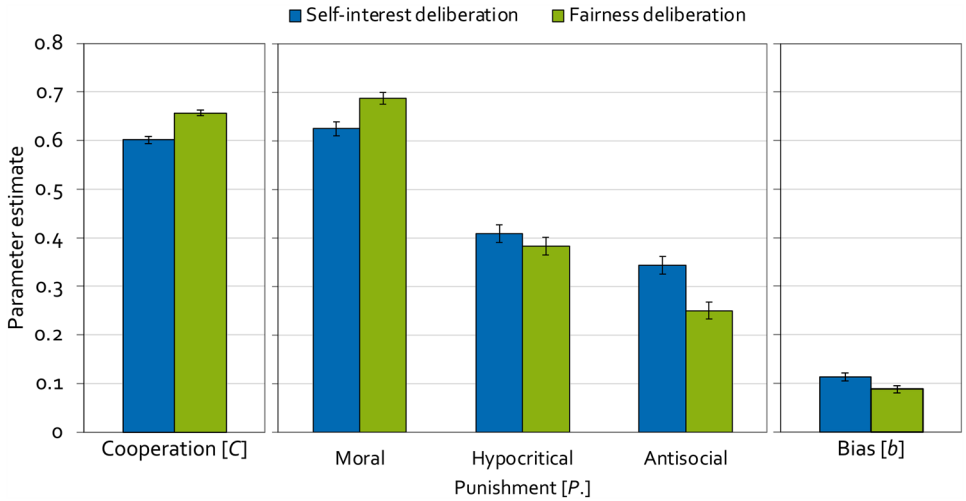

Figure 6. Estimates of the parameters of the cooperation-and-punishment model as a function of the content of deliberation (self-interest, fairness) in Experiment 3. Parameter C represents the probability of cooperation. Parameters P_{Moral}, $P_{Hypocritical}$ and $P_{Antisocial}$ represent the conditional probabilities of moral punishment, hypocritical punishment and antisocial punishment, respectively. Parameter b represents the punishment bias, that is, the probability of punishment irrespective of the outcome of the Prisoner's Dilemma game. The error bars represent the standard errors.

The central hypothesis test again concerns the parameter representing the probability of moral punishment. In line with the process-based account according to which the effect of deliberation on moral punishment depends on the content of deliberation, moral punishment was significantly less likely when participants deliberated self-interest than when they deliberated fairness, $\Delta G^2\,(1) = 11.63$, $p < 0.001$, $w = 0.03$. As a secondary finding, hypocritical punishment did not differ as a function of the content of deliberation, $\Delta G^2\,(1) = 0.96$, $p = 0.327$, $w = 0.01$. Antisocial punishment was more likely when participants deliberated self-interest compared to when they deliberated fairness, $\Delta G^2\,(1) = 13.38$, $p < 0.001$, $w = 0.04$. Further, the punishment bias was significantly higher when participants deliberated self-interest than when they deliberated fairness, $\Delta G^2\,(1) = 5.67$, $p = 0.017$, $w = 0.02$.

Discussion

The aim of Experiment 3 was to test whether moral punishment depends on the content of deliberation. The experimental procedure was identical to the condition with deliberation used in Experiment 2, but participants were instructed to deliberate either self-interest or fairness. Moral punishment was more likely when participants were instructed to deliberate fairness rather than self-interest. These results demonstrate that the effect of deliberation on moral punishment depends on contextual factors such as, in the present case, the content of deliberation. This finding supports the process-based account of punishment[52] implying that, to fully understand the effects of deliberation on moral behaviors, it is necessary to move away from the strict dichotomy that lies at the core of dual-process theories. Previous studies have already hinted at contextual factors that might modulate the effect of deliberation on punishment[37,46,51,84–86]. Here we extend these findings by demonstrating that the instructed content of deliberation affects moral punishment.

Even when participants deliberated their self-interest, they still morally punished unilateral defection with a high probability although this entailed sacrificing own costs for no apparent benefit and thereby went against their self-interest. Previous studies using the cooperation-and-punishment model have already demonstrated the robustness of participants' inclination to morally punish unilateral defection which remained at a high level even when, for example, the majority of partners defected[58] or when participants were offered other ways to communicate their anger in response to the defection[57]. Taken together, these findings suggest that, despite being sensitive to contextual factors, people have a strong and robust preference for moral punishment which aligns with the moral preference hypothesis proposed by Capraro and Perc[87].

In contrast to moral punishment, antisocial punishment was more likely when participants deliberated their self-interest compared to when they deliberated fairness. This underlines the idea that antisocial punishment is antithetical to the very fairness norms that moral punishment is assumed to promote[88,89]. Specifically, it has been suggested that antisocial punishment serves to directly oppose the normative pressure towards cooperation[58], implying that individuals are more likely to engage in antisocial punishment when they prioritize self-interest over fairness. Hypocritical punishment remained unaffected by the manipulation of the contents of deliberation while the punishment bias was increased in the self-interest-deliberation condition compared to the fairness-deliberation condition. These findings validate that the punishment-enhancing effect of deliberating on fairness versus self-interest was specific to moral punishment.

General discussion

Moral punishment of defection is essential for maintaining large-scale cooperation. People frequently accept costs to morally punish defection even if this does not yield any personal benefits for the punisher. The question prevails whether this individually irrational yet socially valuable behavior occurs intuitively, as predicted by an intuitive-morality account of punishment[10], or whether it actually requires time and deliberation to overcome selfish incentive-driven impulses, as assumed by a deliberate-morality account of punishment[7]. Evidence on that matter is mixed and mainly stems from the Ultimatum Game which does not enable to clearly differentiate between participants' inclination to cooperate and their inclination to punish[55]. In the present set of experiments, we relied on a one-shot simultaneous Prisoner's Dilemma game with a punishment option. The cooperation-and-punishment model[5,56–58] was applied to separately measure cooperation, moral punishment, hypocritical punishment and antisocial punishment as well as the punishment bias.

With respect to our initial question of whether the deliberate-morality account or the intuitive-morality account better explains how people apply punishment in social dilemmas, the conclusion is that the present results are more in line with the deliberate-morality account than with the intuitive-morality account of punishment. In Experiment 1, making punishment decisions with time pressure led to a decrease in moral punishment relative to a condition without time pressure, indicating that intuition tends to decrease moral behavior. These findings confirm and extend previous findings showing strikingly parallel effects to those of a cognitive-load manipulation[5]. Taken together, these findings suggest that the lack of cognitive resources decreases moral punishment, regardless of whether this lack of cognitive resources is caused by cognitive load or time pressure. In Experiment 2, punishment decisions were delayed by pauses in which participants were encouraged to deliberate their punishment decision. Deliberation led to an increase in moral punishment relative to a condition without deliberation. These findings support the deliberate-morality account of punishment by demonstrating that deliberation can facilitate the moral use of punishment. On the face of it, there is some plausibility to the idea that moral punishment requires deliberation. Punishment can be seen as a second-order social dilemma because participants may be torn between shying away from the costs of punishment which run against their self-interest and the moral motive of punishing unfair behaviors[90–92]. Deliberation may be needed to resolve these conflicting goals in favor of moral principles. Giving participants time to deliberate their decision may help to adhere to moral principles by suppressing their selfish impulses[51]. This is in line with neuro-imaging studies showing that punishment decisions in social dilemma games are accompanied by increased activation in brain areas associated with cognitive control[53,54]. It is also in line with research on moral behaviors more generally.

For instance, there is evidence suggesting that people's tendency to lie is increased when cognitive resources are limited, as telling the truth has been found to require self-control when lying serves immediate self-interests[30,31]. However, it is inconsistent with other findings showing evidence in the opposite direction[28,29]. Parallel to this, the present findings preserve the existing ambiguity regarding the intuitive versus deliberative nature of cooperation decisions[24], with Experiment 1 showing no effect of time pressure on cooperation and Experiment 2 showing less cooperation with deliberation in comparison to the condition without deliberation.

Based on inconsistent findings regarding the effects of deliberation on moral decision making, Declerck and Boone[52] have proposed a process-based account according to which the effect of deliberation on moral behaviors, such as cooperation and moral punishment, depends on contextual factors[8]. Several studies support this account[37,46,51,84,85]. The results of Experiment 3 extend these previous findings by demonstrating that explicit instructions on what to deliberate affect moral punishment: When participants deliberated fairness instead of self-interest, moral punishment increased.

Taken together, the findings of the present experiments thus suggest that, without manipulating the content of deliberation, unguided deliberation leads participants to use punishment in a moral way. However, the content of deliberation determines the effect of deliberation on moral punishment. Depending on the context in which the decision to punish has to be made, the content of deliberation may deviate from the moral default. Viewing punishment through the lens of the process-based account[52] might be slightly more complex than postulating a strict dichotomy between intuitive and deliberate ways of arriving at a punishment decision, but adopting such a more nuanced approach offers great potential to integrate conflicting findings on the effects of intuition and deliberation on moral cognition[52,76].

We conclude from the present findings that future research should focus on how the effects of intuition and deliberation on moral punishment are modulated by contextual factors. As an example of a potential avenue for future research, it seems possible to postulate that the effects of intuition on punishment should be context-dependent, just as the effects of deliberation on punishment observed in the present Experiment 3. That is, if heuristic cues favor moral or antisocial interpretations of situations, intuitions should favor or discourage punishment accordingly. Supporting this speculation, evidence suggests that the effect of intuition on punishment is modulated by contextual factors such as group membership[93,94].

A limitation of the present study is that, in the present series of experiments, the social context of the interaction was emphasized, for instance, by presenting facial photographs of the interaction partners to take into account that everyday interactions often contain social cues. Given sufficient time for deliberation, this may have stimulated reflection on the social aspects of the interaction. This approach contrasts with the usual approach in Experimental Economics in which, even though social information is frequently manipulated, the games themselves are typically described in neutral terms. Another difference to the usual economic approach towards studying social interactions is that the participants interacted with simulated interaction partners, a standard practice in psychological research[63,95–97]. The experimental manipulation guarantees control over the partner's behavior which is considered an extraneous factor in Experimental Psychology where the aim is to draw inferences about the cognitions underlying the individual's behavior. It seems remarkable that participants in the present study punished their partners at a rate that is typical for studies involving human interaction partners even though this implied accepting real costs[21,98]. What is more, the fact that deliberation increased the probability of moral punishment suggests that the present paradigm taps into mechanisms of social interaction. Nevertheless, it has to be counted among the limitations of the present paradigm that participants interacted with programmed partners, particularly from the perspective of Experimental Economics where the primary focus lies on studying how incentive structures affect interactions in dyads or larger groups. It is therefore an interesting avenue of future research to test whether the present conclusions generalize to settings in which participants interact in human dyads or groups. It also seems important to note that the present set of experiments focused on one-shot interactions, that is, social situations during which participants interacted with each partner only once. More specifically, participants interacted with each of 20 different partners only once. One-shot interactions may be considered particularly relevant in that people in everyday life often invest own resources to punish strangers with whom they have interacted only once, which is an interesting phenomenon that requires explanation[58]. Even though one-shot interactions are generally assumed to be less influenced by factors such as reputation building or social learning, the repeated engagement of participants in the present Prisoner's Dilemma game with different partners implies that an influence of these factors cannot be entirely ruled out. As a consequence, participants' inclination to morally punish might have changed as participants progressed from the first interaction all the way to the final interaction. To examine whether this was the case, we tested whether a model would fit the data in which the moral-punishment parameter was set to be equal across the entire sequence of 20 interactions, separately for each of the two groups in Experiments 1, 2 and 3 (generating a total of 38 degrees of freedom for this statistical test in each experiment, 19 for the equality restriction of the moral-punishment parameter across the 20 interactions for each of the two groups). This equality restriction was compatible with the data in Experiment 1, $\Delta G^2(38) = 48.86$, $p = 0.111$, in Experiment 2, $\Delta G^2(38) = 24.13$, $p = 0.960$ and in Experiment 3, $\Delta G^2(38) = 31.40$, $p = 0.767$, leading to the conclusion that the parameter representing moral punishment did not change as participants progressed from the first interaction to the final interaction. This, in turn, suggests that in the present Prisoner's Dilemma game the possible influence of factors such as reputation building or social learning on moral punishment is limited.

Another point worth discussing is that participants punished the partners they previously interacted with. As they directly suffered from their partners' defection, the processes reflected in the moral-punishment parameter might go beyond motives that can, in a strict sense, be considered truly moral. For example, participants might punish to retaliate against the partners who have caused them harm[99]. It seems interesting to replicate the present findings in a paradigm that minimizes such self-interested motivations, for instance in a third-party-punishment paradigm in which the punishing party is not directly affected by the partner's defection[100].

Conclusion

Do people punish defection intuitively or does this individually irrational, yet socially valuable, behavior rely on deliberation? Here we have found that time pressure decreased moral punishment (Experiment 1), which is parallel to what has been previously reported for a cognitive-load manipulation[5], whereas deliberation during a time delay in which participants reflect on their punishment decision increased moral punishment (Experiment 2). Furthermore, we have demonstrated that moral punishment is affected by the specific content of deliberation (Experiment 3). When participants deliberate fairness, they are more likely to rely on moral punishment than when they deliberate self-interest. The results of Experiments 1 and 2 per se are more in line with the idea that moral punishment is applied in a deliberate rather than in an intuitive way. However, those results are also compatible with a more nuanced account supported by the results of Experiment 3 in which the effect of deliberation on moral punishment was modulated by contextual factors. Moving away from the strict dichotomy of classifying moral behaviors as either strictly intuitive or strictly deliberate has great potential in deepening our understanding of the effects on deliberation on punishment and offers a promising avenue for incorporating seemingly inconsistent findings by focusing on the cognitive processes underlying the observable behavior.

Data availability

We provide the data used in our analyses via the Open Science Framework. The study was not preregistered. The data are publicly available at https://osf.io/a23gy/.

Received: 19 March 2024; Accepted: 9 July 2024
Published online: 16 July 2024

References

1. Van Lange, P. A. & Rand, D. G. Human cooperation and the crises of climate change, COVID-19, and misinformation. *Annu. Rev. Psychol.* **73**, 379–402. https://doi.org/10.1146/annurev-psych-020821-110044 (2022).
2. Clutton-Brock, T. Cooperation between non-kin in animal societies. *Nature* **462**, 51–57. https://doi.org/10.1038/nature08366 (2009).
3. Rand, D. G. & Nowak, M. A. Human cooperation. *Trends Cogn. Sci.* **17**, 413–425. https://doi.org/10.1016/j.tics.2013.06.003 (2013).
4. Kurzban, R., DeScioli, P. & O'Brien, E. Audience effects on moralistic punishment. *Evol. Hum. Behav.* **28**, 75–84. https://doi.org/10.1016/j.evolhumbehav.2006.06.001 (2007).
5. Mieth, L., Buchner, A. & Bell, R. Cognitive load decreases cooperation and moral punishment in a Prisoner's Dilemma game with punishment option. *Sci. Rep.* **11**, 1–12. https://doi.org/10.1038/s41598-021-04217-4 (2021).
6. Henrich, J. *et al.* Costly punishment across human societies. *Science* **312**, 1767–1770. https://doi.org/10.1126/science.1127333 (2006).
7. DeWall, C. N., Baumeister, R. F., Gailliot, M. T. & Maner, J. K. Depletion makes the heart grow less helpful: Helping as a function of self-regulatory energy and genetic relatedness. *Pers. Soc. Psychol. Bull.* **34**, 1653–1662. https://doi.org/10.1177/0146167208323981 (2008).
8. Isler, O., Gächter, S., Maule, A. J. & Starmer, C. Contextualised strong reciprocity explains selfless cooperation despite selfish intuitions and weak social heuristics. *Sci. Rep.* **11**, 13868. https://doi.org/10.1038/s41598-021-93412-4 (2021).
9. Zaki, J. & Mitchell, J. P. Intuitive prosociality. *Curr. Dir. Psychol. Sci.* **22**, 466–470. https://doi.org/10.1177/0963721413492764 (2013).
10. Rand, D. G., Greene, J. D. & Nowak, M. A. Spontaneous giving and calculated greed. *Nature* **489**, 427–430. https://doi.org/10.1038/nature11467 (2012).
11. Kollock, P. Social dilemmas: The anatomy of cooperation. *Annu. Rev. Sociol.* **24**, 183–214. https://doi.org/10.1146/annurev.soc.24.1.183 (1998).
12. Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396. https://doi.org/10.1126/science.7466396 (1981).
13. Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560–1563. https://doi.org/10.1126/science.1133755 (2006).
14. Fehr, E. & Fischbacher, U. Social norms and human cooperation. *Trends Cogn. Sci.* **8**, 185–190. https://doi.org/10.1016/j.tics.2004.02.007 (2004).
15. Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. The evolution of altruistic punishment. *Proc. Natl. Acad. Sci.* **100**, 3531–3535. https://doi.org/10.1073/pnas.0630443100 (2003).
16. Boyd, R. & Richerson, P. J. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195. https://doi.org/10.1016/0162-3095(92)90032-Y (1992).
17. Yamagishi, T. The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.* **51**, 110–116. https://doi.org/10.1037/0022-3514.51.1.110 (1986).
18. Axelrod, R. An evolutionary approach to norms. *Am. Polit. Sci. Rev.* **80**, 1095–1111. https://doi.org/10.2307/1960858 (1986).
19. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140. https://doi.org/10.1038/415137a (2002).
20. Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**, 980–994. https://doi.org/10.1257/aer.90.4.980 (2000).
21. Falk, A., Fehr, E. & Fischbacher, U. Driving forces behind informal sanctions. *Econometrica* **73**, 2017–2030. https://doi.org/10.1111/j.1468-0262.2005.00644.x (2005).
22. Ostrom, E., Gardner, R., Walker, J., Walker, J. M. & Walker, J. *Rules, Games, and Common-Pool Resources.* (University of Michigan Press, Michigan, 1994). https://press.umich.edu/pdf/9780472065462-fm.pdf.
23. Przepiorka, W. & Diekmann, A. Individual heterogeneity and costly punishment: a volunteer's dilemma. *Proc. R. Soc. B. Biol. Sci.* **280**, 20130247. https://doi.org/10.1098/rspb.2013.0247 (2013).
24. Capraro, V. The dual-process approach to human sociality: Meta-analytic evidence for a theory of internalized heuristics for self-preservation. *J. Pers. Soc. Psychol.* https://doi.org/10.1037/pspa0000375 (2024).
25. Evans, J. S. B. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.* **59**, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629 (2008).
26. Evans, J. S. B. & Stanovich, K. E. Dual-process theories of higher cognition: Advancing the debate. *Perspect. Psychol. Sci.* **8**, 223–241. https://doi.org/10.1177/1745691612460685 (2013).
27. Kahneman, D. *Thinking, Fast and Slow* (Farrar, 2011).

28. Capraro, V. Does the truth come naturally? Time pressure increases honesty in one-shot deception games. *Econ. Lett.* **158**, 54–57. https://doi.org/10.1016/j.econlet.2017.06.015 (2017).
29. Capraro, V., Schulz, J. & Rand, D. G. Time pressure and honesty in a deception game. *J. Behav. Exp. Econ.* **79**, 93–99. https://doi.org/10.1016/j.socec.2019.01.007 (2019).
30. Köbis, N. C., Verschuere, B., Bereby-Meyer, Y., Rand, D. & Shalvi, S. Intuitive honesty versus dishonesty: Meta-analytic evidence. *Perspect. Psychol. Sci.* **14**, 778–796. https://doi.org/10.1177/1745691619851778 (2019).
31. Shalvi, S., Eldar, O. & Bereby-Meyer, Y. Honesty requires time (and lack of justifications). *Psychol. Sci.* **23**, 1264–1270. https://doi.org/10.1177/0956797612443835 (2012).
32. Fehr, E. & Fischbacher, U. The nature of human altruism. *Nature* **425**, 785–791. https://doi.org/10.1038/nature02043 (2003).
33. Gintis, H., Bowles, S., Boyd, R. & Fehr, E. Explaining altruistic behavior in humans. *Evol. Hum. Behav.* **24**, 153–172. https://doi.org/10.1016/S1090-5138(02)00157-5 (2003).
34. Pillutla, M. M. & Murnighan, J. K. Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organ. Behav. Hum. Decis. Process.* **68**, 208–224. https://doi.org/10.1006/obhd.1996.0100 (1996).
35. Sutter, M., Kocher, M. & Strauß, S. Bargaining under time pressure in an experimental ultimatum game. *Econ. Lett.* **81**, 341–347. https://doi.org/10.1016/S0165-1765(03)00215-5 (2003).
36. Liu, Y., He, N. & Dou, K. Ego-depletion promotes altruistic punishment. *Open J. Soc. Sci.* **3**, 62–69. https://doi.org/10.4236/jss.2015.311009 (2015).
37. Halali, E., Bereby-Meyer, Y. & Meiran, N. Between self-interest and reciprocity: The social bright side of self-control failure. *J. Exp. Psychol. Gen.* **143**, 745. https://doi.org/10.1037/a0033824 (2014).
38. Cappelletti, D., Güth, W. & Ploner, M. Being of two minds: Ultimatum offers under cognitive constraints. *J. Econ. Psychol.* **32**, 940–950. https://doi.org/10.1016/j.joep.2011.08.001 (2011).
39. Neo, W. S., Yu, M., Weber, R. A. & Gonzalez, C. The effects of time delay in reciprocity games. *J. Econ. Psychol.* **34**, 20–35. https://doi.org/10.1016/j.joep.2012.11.001 (2013).
40. Smith, P. & Silberberg, A. Rational maximizing by humans (homo sapiens) in an ultimatum game. *Anim. Cogn.* **13**, 671–677. https://doi.org/10.1007/s10071-010-0310-4 (2010).
41. Wang, C. S. *et al.* Retribution and emotional regulation: The effects of time delay in angry economic interactions. *Organ. Behav. Hum. Decis. Process.* **116**, 46–54. https://doi.org/10.1016/j.obhdp.2011.05.007 (2011).
42. Grimm, V. & Mengel, F. Let me sleep on it: Delay reduces rejection rates in ultimatum games. *Econ. Lett.* **111**, 113–115. https://doi.org/10.1016/j.econlet.2011.01.025 (2011).
43. Artavia-Mora, L., Bedi, A. S. & Rieger, M. Intuitive cooperation and punishment in the field. *IZA Discussion Paper No. 9871*, https://doi.org/10.2139/ssrn.2769179 (2016).
44. Bosman, R., Sonnemans, J. & Zeelenberg, M. *Emotions, Rejections, and Cooling Off in the Ultimatum Game.* (University of Amsterdam, 2001). https://hdl.handle.net/11245/1.418488.
45. Cappelletti, D., Güth, W. & Ploner, M. Being of two minds: Ultimatum offers under cognitive constraints. *J. Econ. Psychol.*, **32**(6), 940–950. https://doi.org/10.1016/j.joep.2011.08.001 (2011).
46. Oechssler, J., Roider, A. & Schmitz, P. W. Cooling off in negotiations: Does it work?. *J. Inst. Theor. Econ.* https://doi.org/10.1628/093245615X14307212950056 (2015).
47. Achtziger, A., Alós-Ferrer, C. & Wagner, A. K. Social preferences and self-control. *J. Behav. Exp. Econ.* **74**, 161–166. https://doi.org/10.1016/j.socec.2018.04.009 (2018).
48. Olschewski, S., Rieskamp, J. & Scheibehenne, B. Taxing cognitive capacities reduces choice consistency rather than preference: A model-based test. *J. Exp. Psychol. Gen.* **147**, 462. https://doi.org/10.1037/xge0000403 (2018).
49. Achtziger, A., Alós-Ferrer, C. & Wagner, A. K. The impact of self-control depletion on social preferences in the ultimatum game. *J. Econ. Psychol.* **53**, 1–16. https://doi.org/10.1016/j.joep.2015.12.005 (2016).
50. Hochman, G., Ayal, S. & Ariely, D. Fairness requires deliberation: The primacy of economic over social considerations. *Front. Psychol.* **6**, 747. https://doi.org/10.3389/fpsyg.2015.00747 (2015).
51. Ferguson, E., Maltby, J., Bibby, P. A. & Lawrence, C. Fast to forgive, slow to retaliate: Intuitive responses in the ultimatum game depend on the degree of unfairness. *Plos One* **9**, e96344. https://doi.org/10.1371/journal.pone.0096344 (2014).
52. Declerck, C. & Boone, C. *Neuroeconomics of Prosocial Behavior: The Compassionate Egoist* (Academic Press, 2015).
53. Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V. & Fehr, E. Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science* **314**, 829–832. https://doi.org/10.1126/science.1129156 (2006).
54. Knoch, D. *et al.* Studying the neurobiology of social interaction with transcranial direct current stimulation—The example of punishing unfairness. *Cereb. Cortex* **18**, 1987–1990. https://doi.org/10.1093/cercor/bhm237 (2008).
55. Yamagishi, T. *et al.* Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proc. Natl. Acad. Sci.* **109**, 20364–20368. https://doi.org/10.1073/pnas.1212126109 (2012).
56. Mieth, L., Buchner, A. & Bell, R. Moral labels increase cooperation and costly punishment in a Prisoner's Dilemma game with punishment option. *Sci. Rep.* **11**, 1–13. https://doi.org/10.1038/s41598-021-89675-6 (2021).
57. Philippsen, A., Mieth, L., Buchner, A. & Bell, R. Communicating emotions, but not expressing them privately, reduces moral punishment in a Prisoner's Dilemma game. *Sci. Rep.* **13**, 14693. https://doi.org/10.1038/s41598-023-41886-9 (2023).
58. Philippsen, A., Mieth, L., Buchner, A. & Bell, R. People punish defection, not failures to conform to the majority. *Sci. Rep.* **14**, 1211. https://doi.org/10.1038/s41598-023-50414-8 (2024).
59. Batchelder, W. H. & Riefer, D. M. Theoretical and empirical review of multinomial process tree modeling. *Psychon. Bull. Rev.* **6**, 57–86. https://doi.org/10.3758/BF03210812 (1999).
60. Erdfelder, E. *et al.* Multinomial processing tree models: A review of the literature. *Z. Psychol.* **217**, 108–124. https://doi.org/10.1027/0044-3409.217.3.108 (2009).
61. Schmidt, O., Erdfelder, E. & Heck, D. W. How to develop, test, and extend multinomial processing tree models: A tutorial. *Psychol. Methods* https://doi.org/10.1037/met0000561 (2023).
62. Moshagen, M. multiTree: A computer program for the analysis of multinomial processing tree models. *Behav. Res. Methods* **42**, 42–54. https://doi.org/10.3758/BRM.42.1.42 (2010).
63. Bell, R., Mieth, L. & Buchner, A. Separating conditional and unconditional cooperation in a sequential Prisoner's Dilemma game. *Plos One* **12**, e0187952. https://doi.org/10.1371/journal.pone.0187952 (2017).
64. Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R. & Hütter, M. Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *J. Pers. Soc. Psychol.* **113**, 343. https://doi.org/10.1037/pspa0000086 (2017).
65. Klauer, K. C., Stahl, C. & Erdfelder, E. The abstract selection task: New data and an almost comprehensive model. *J. Exp. Psychol. Learn. Mem. Cogn.* **33**, 680–703. https://doi.org/10.1037/0278-7393.33.4.680 (2007).
66. Kroneisen, M. & Steghaus, S. The influence of decision time on sensitivity for consequences, moral norms, and preferences for inaction: Time, moral judgments, and the CNI model. *J. Behav. Decis. Mak.* **34**, 140–153. https://doi.org/10.1002/bdm.2202 (2021).
67. Moshagen, M., Hilbig, B. E. & Musch, J. Defection in the dark? A randomized-response investigation of cooperativeness in social dilemma games. *Eur. J. Soc. Psychol.* **41**, 638–644. https://doi.org/10.1002/ejsp.793 (2011).
68. Schaper, M. L., Mieth, L. & Bell, R. Adaptive memory: Source memory is positively associated with adaptive social decision making. *Cognition* **186**, 7–14. https://doi.org/10.1016/j.cognition.2019.01.014 (2019).

69. Heck, D. W., Hoffmann, A. & Moshagen, M. Detecting nonadherence without loss in efficiency: A simple extension of the crosswise model. *Behav. Res. Methods* **50**, 1895–1905. https://doi.org/10.3758/s13428-017-0957-8 (2018).
70. Hoffmann, A., Diedenhofen, B., Verschuere, B. & Musch, J. A strong validation of the crosswise model using experimentally-induced cheating behavior. *Exp. Psychol.* https://doi.org/10.1027/1618-3169/a000304 (2015).
71. Batchelder, W. H. & Riefer, D. M. Multinomial processing models of source monitoring. *Psychol. Rev.* **97**, 548. https://doi.org/10.1037/0033-295X.97.4.548 (1990).
72. Bayen, U. J., Murnane, K. & Erdfelder, E. Source discrimination, item detection, and multinomial models of source monitoring. *J. Exp. Psychol. Learn. Mem. Cogn.* **22**, 197–215. https://doi.org/10.1037/0278-7393.22.1.197 (1996).
73. Buchner, A., Erdfelder, E. & Vaterrodt-Plünnecke, B. Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *J. Exp. Psychol. Gen.* **124**, 137–160. https://doi.org/10.1037/0096-3445.124.2.137 (1995).
74. Erdfelder, E., Cüpper, L., Auer, T.-S. & Undorf, M. The four-states model of memory retrieval experiences. *Z. Psychol.* **215**, 61–71. https://doi.org/10.1027/0044-3409.215.1.61 (2007).
75. Menne, N. M., Winter, K., Bell, R. & Buchner, A. A validation of the two-high threshold eyewitness identification model by reanalyzing published data. *Sci. Rep.* **12**, 13379. https://doi.org/10.1038/s41598-022-17400-y (2022).
76. Krajbich, I., Bartling, B., Hare, T. & Fehr, E. Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nat. Commun.* **6**, 7455. https://doi.org/10.1038/ncomms8455 (2015).
77. Capraro, V. The dual-process approach to human sociality: Meta-analytic evidence for a theory of internalized heuristics for self-preservation. *J. Pers. Soc. Psychol.* https://doi.org/10.1037/pspa0000375 (in press).
78. Verkoeijen, P. P. & Bouwmeester, S. Does intuition cause cooperation?. *Plos One* **9**, e96654. https://doi.org/10.1371/journal.pone.0096654 (2014).
79. Leiner, D. J. *SoSci Survey* (Version 3.5.02) [Computer software]. Available at https://www.soscisurvey.de (2024).
80. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191. https://doi.org/10.3758/BF03193146 (2007).
81. Mieth, L., Bell, R. & Buchner, A. Facial likability and smiling enhance cooperation, but have no direct effect on moralistic punishment. *J. Exp. Psychol.* **63**, 263–277. https://doi.org/10.1027/1618-3169/a000338 (2016).
82. Mieth, L., Buchner, A. & Bell, R. Effects of gender on costly punishment. *J. Behav. Decis. Mak.* **30**, 899–912. https://doi.org/10.1002/bdm.2012 (2017).
83. Ma, D. S., Correll, J. & Wittenbrink, B. The Chicago face database: A free stimulus set of faces and norming data. *Behav. Res. Methods* **47**, 1122–1135. https://doi.org/10.3758/s13428-014-0532-5 (2015).
84. Speer, S. P., Smidts, A. & Boksem, M. A. Cognitive control and dishonesty. *Trends Cogn. Sci.* **26**, 796–808. https://doi.org/10.1016/j.tics.2022.06.005 (2022).
85. Bieleke, M., Gollwitzer, P. M., Oettingen, G. & Fischbacher, U. Social value orientation moderates the effects of intuition versus reflection on responses to unfair ultimatum offers. *J. Behav. Decis. Mak.* **30**, 569–581. https://doi.org/10.1002/bdm.1975 (2017).
86. Harris, A. *et al.* Perceived relative social status and cognitive load influence acceptance of unfair offers in the ultimatum game. *Plos One* **15**, e0227717. https://doi.org/10.1371/journal.pone.0227717 (2020).
87. Capraro, V. & Perc, M. Mathematical foundations of moral preferences. *J. R. Soc. Interface.* **18**, e20200880. https://doi.org/10.1098/rsif.2020.0880 (2021).
88. Herrmann, B., Thoni, C. & Gachter, S. Antisocial punishment across societies. *Science* **319**, 1362–1367. https://doi.org/10.1126/science.1153808 (2008).
89. Sylwester, K., Herrmann, B. & Bryson, J. J. Homo homini lupus? Explaining antisocial punishment. *J. Neurosci. Psychol. Econ.* **6**, 167–188. https://doi.org/10.1037/npe0000009 (2013).
90. Alger, I. & Weibull, J. W. Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica* **81**, 2269–2302. https://doi.org/10.3982/ECTA10637 (2013).
91. Capraro, V., Jagfeld, G., Klein, R., Mul, M. & de Pol, I. V. Increasing altruistic and cooperative behaviour with simple moral nudges. *Sci. Rep.* **9**, 1–11. https://doi.org/10.1038/s41598-019-48094-4 (2019).
92. Krupka, E. L. & Weber, R. A. Identifying social norms using coordination games: Why does dictator game sharing vary?. *J. Eur. Econ. Assoc.* **11**, 495–524. https://doi.org/10.1111/jeea.12006 (2013).
93. Yudkin, D. A., Rothmund, T., Twardawski, M., Thalla, N. & Van Bavel, J. J. Reflexive intergroup bias in third-party punishment. *J. Exp. Psychol. Gen.* **145**, 1448. https://doi.org/10.1037/xge0000190 (2016).
94. Mischkowski, D., Glöckner, A. & Lewisch, P. From spontaneous cooperation to spontaneous punishment—Distinguishing the underlying motives driving spontaneous behavior in first and second order public good games. *Organ. Behav. Hum. Decis. Process.* **149**, 59–72. https://doi.org/10.1016/j.obhdp.2018.07.001 (2018).
95. Parks, C. D. & Stone, A. B. The desire to expel unselfish members from the group. *J. Pers. Soc. Psychol.* **99**, 303–310. https://doi.org/10.1037/a0018403 (2010).
96. Wang, L., Zheng, J., Meng, L., Lu, Q. & Ma, Q. Ingroup favoritism or the black sheep effect: Perceived intentions modulate subjective responses to aggressive interactions. *Neurosci. Res.* **108**, 46–54. https://doi.org/10.1016/j.neures.2016.01.011 (2016).
97. Volstorf, J., Rieskamp, J. & Stevens, J. R. The good, the bad, and the rare: Memory for partners in social interactions. *Plos One* **6**, e18945. https://doi.org/10.1371/journal.pone.0018945 (2011).
98. Carpenter, J. P. The demand for punishment. *J. Econ. Behav. Organ.* **62**, 522–542. https://doi.org/10.1016/j.jebo.2005.05.004 (2007).
99. Nockur, L., Kesberg, R., Pfattheicher, S. & Keller, J. Why do we punish? On retribution, deterrence, and the moderating role of punishment system. *Z. Psychol.* **230**, 104–113. https://doi.org/10.1027/2151-2604/a000457 (2022).
100. Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evol. Hum. Behav.* **25**, 63–87. https://doi.org/10.1016/S1090-5138(04)00005-4 (2004).

Author contributions

A.P., L.M., A.B. and R.B. contributed to the study conception and design. Preparation of the study, data collection and analysis were performed by L.M. and A.P. All authors contributed through discussion and interpretation of the results. A.P. and L.M. wrote the manuscript with subsequent input and final approval from all co-authors.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024

Declaration of the independent contribution to the published articles included in the present dissertation

My dissertation includes three articles, with seven experiments in total, which have been published in an academic journal with an established peer-review process. In the following, I will elaborate how each author individually contributed to the article. The majority of work was always carried out by the first author of the article.

Independent contribution to Published Article 1

Publication:

Philippsen, A., Mieth, L., Buchner, A., & Bell, R. (2023). Communicating emotions, but not expressing them privately, reduces moral punishment in a Prisoner's Dilemma game. *Scientific Reports*, *13*(1), Article: 14693. https://doi.org/10.1038/s41598-023-41886-9

Study conception: Laura Mieth and I developed the experimental design of Experiment 1.1 and 1.2 with support from Axel Buchner and Raoul Bell.

Implementation: Laura Mieth and I programmed and conducted Experiment 1.1 and Experiment 1.2 with feedback from Axel Buchner and Raoul Bell.

Data analysis: I conducted the statistical analyses independently. Laura Mieth, Axel Buchner and Raoul Bell reviewed their accuracy.

Manuscript: I prepared the manuscript independently, including an extensive literature review, the design of the figures and the composition and writing of the manuscript. Laura Mieth, Axel Buchner and Raoul Bell provided feedback which I incorporated after a thorough review and subsequent consultation. I managed the peer-review process through the academic journal. During this process, I made revisions with support from Laura Mieth, Axel Buchner and Raoul Bell. I prepared the final version of the manuscript.

Independent contribution to Published Article 2

Publication:

Philippsen, A., Mieth, L., Buchner, A., & Bell, R. (2024). People punish defection, not failures to conform to the majority. *Scientific Reports*, *14*(1), Article: 1211. https://doi.org/10.1038/s41598-023-50414-8

Study conception: I developed the experimental design of Experiments 2.1 and 2.2 with support from Laura Mieth, Axel Buchner and Raoul Bell.

Implementation: I programmed Experiments 2.1 and 2.2 with feedback from Laura Mieth, Axel Buchner and Raoul Bell. I collected the data independently.

Data analysis: I conducted the statistical analyses independently. Laura Mieth, Axel Buchner and Raoul Bell reviewed their accuracy.

Manuscript: I prepared the manuscript independently, including an extensive literature review, the design of the figures and the composition and writing of the manuscript. Laura Mieth, Axel Buchner and Raoul Bell provided feedback which I incorporated after a thorough review and subsequent consultation. I managed the peer-review process through the academic journal. During this process, I made revisions with support from Laura Mieth, Axel Buchner and Raoul Bell. I prepared the final version of the manuscript.

Independent contribution to Published Article 3

Publication:

Philippsen, A., Mieth, L., Buchner, A., & Bell, R. (2024). Time pressure and deliberation affect moral punishment. *Scientific Reports*, *14*(1), Article: 16378. https://doi.org/10.1038/s41598-024-67268-3

Study conception: Laura Mieth and I developed the design of Experiments 3.1, 3.2 and 3.3 with support from Axel Buchner und Raoul Bell.

Implementation: Laura Mieth and I programmed the experiments with feedback from Axel Buchner and Raoul Bell. Laura Mieth and I supervised the data collection of the experiments.

Data analysis: I conducted the statistical analyses independently. Laura Mieth, Axel Buchner and Raoul Bell reviewed their accuracy.

Manuscript: I prepared the manuscript independently, including an extensive literature review, the design of the figures and the composition and writing of the manuscript. Laura Mieth, Axel Buchner and Raoul Bell provided feedback which I incorporated after a thorough review and subsequent consultation. I managed the peer-review process through the academic journal. During this process, I made revisions with support from Laura Mieth, Axel Buchner and Raoul Bell. I prepared the final version of the manuscript. I prepared the final version of the manuscript.

Erklärung an Eides statt

Ich versichere an Eides statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist.

Ich versichere insbesondere:

(1) Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt.

(2) Alle wörtlich oder dem Sinn nach aus anderen Texten entnommenen Stellen habe ich als solche kenntlich gemacht; dies gilt für gedruckte Texte ebenso wie für elektronische Ressourcen.

(3) Die Arbeit habe ich in der vorliegenden oder einer modifizierten Form noch nicht als Dissertation vorgelegt – sei es an der Heinrich-Heine-Universität Düsseldorf oder an einer anderen Universität.

Düsseldorf, 26. September 2024

Ana Isabel Philippsen