Third-party punishers who express emotions are trusted more

Kupfer, Tom R.; Tybur, Joshua M.

published in
Proceedings of the Royal Society B: Biological Sciences
2023

DOI (link to publisher)
10.1098/rspb.2023.0916

document version
Publisher's PDF, also known as Version of record

PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

Cite this article: Kupfer TR, Tybur JM. 2023 Third-party punishers who express emotions are trusted more. *Proc. R. Soc. B* **290**: 20230916.
https://doi.org/10.1098/rspb.2023.0916

Received: 16 August 2022
Accepted: 26 July 2023

Subject Category:
Behaviour

Subject Areas:
behaviour, evolution

Keywords:
third party punishment, trust, cooperation, emotion expression, anger, disgust

Author for correspondence:
Tom R. Kupfer
e-mail: thomas.kupfer@ntu.ac.uk

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6777772.

THE ROYAL SOCIETY PUBLISHING

Third-party punishers who express emotions are trusted more

Tom R. Kupfer[1] and Joshua M. Tybur[2]

[1]Psychology, Nottingham Trent University, Nottingham, UK
[2]Experimental and Applied Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

TRK, 0000-0003-1147-6082; JMT, 0000-0002-0462-6508

Third party punishment (TPP) is thought to be crucial to the evolution and maintenance of human cooperation. However, this type of punishment is often not rewarded, perhaps because punishers' underlying motives are unclear. We propose that the expression of moral emotions could solve this problem by advertising such motives. In each of three experiments ($n = 1711$), a third-party punishment game was followed by a trust game. Third parties expressed anger or disgust instead of, or in addition to, financial punishment. Results showed that third parties who expressed these emotions were trusted more than those who didn't express (Experiment 1), and more than those who financially punished (Experiment 2). Moreover, third parties who expressed while financially punishing were trusted more than those who punished without expressing (Experiment 3). Findings suggest that emotion expression might play a role in the evolution and maintenance of cooperation by facilitating TPP.

1. Introduction

Across cultures and time, societies have relied upon cooperation to flourish. Cooperation has an Achilles' heel, though: free riding. How is cooperation maintained when individuals have incentives to behave selfishly? Cross-disciplinary research has converged on an answer to this question: cooperation is made possible when observers punish non-cooperators—that is, through third party punishment (TPP) [1,2]. However, TPP is costly in itself because targets can retaliate against punishers, and because observers of TPP may avoid interacting with the punisher [3,4]. Due to these costs, the evolutionary origins of TPP are mysterious, as are the incentives for engaging in TPP in contemporary society [5]. One prominent account suggests that individuals who engage in TPP benefit via indirect reciprocity—that is, they gain a reputation of being cooperative and trustworthy, and hence are preferred as cooperation partners in future interactions [6,7]. However, findings are equivocal; some evidence suggests that punishers gain reputation benefits, such as increased trustworthiness [8,9], whereas other evidence suggests that punishers are not rewarded or trusted more than non-punishers [10,11]. Further, evidence that some punishers conceal their behaviour from observers suggests that people are aware that punishing may not give them a good reputation [12].

Recent theorizing suggests that TPP does not consistently improve a punisher's reputation because observers can make multiple distinct inferences regarding the motives underlying a punisher's decision to impose a financial penalty [13]. Although TPP signals a willingness to incur costs to inflict harm [9], the punisher could have selfish motives for paying these costs, such as spite [14], vengefulness [15] or deterrence [16]. If observers are likely to make negative inferences about the motives behind that TPP, then reputational enhancement cannot adequately account for the evolution off TPP in the face of the costs of punishing. We argue that the expression of moral emotions may provide a solution to the costliness of punishment by conveying information about the motives of the expresser.

(a) The role of emotion in cooperation and punishment

Cooperation researchers have long emphasized the importance of emotions in driving cooperative behaviour, including TPP [17,18]. But research has predominantly emphasized the *intrapersonal* role of emotion in motivating TPP. For example, multiple studies have demonstrated that anger and moral outrage motivate punishment of noncooperators (e.g. [2,19]). However, considerable theory and research in social psychology shows that emotions also have *interpersonal* functions, including the communication of socially important information about an expresser's motives and behavioural intentions [20–23]. Little work has investigated the role of emotion expression in TPP, despite cooperation researchers hypothesizing that expressed emotions could themselves function as a form of punishment [24].

Some existing research supports the possibility that expressed emotions can promote cooperation. For example, one study reported that participants allowed to express emotion via written notes after receiving an unfair offer in an ultimatum game were less likely to reject the offer, suggesting that expressing emotion satisfies motives to punish [25]. Another study reported that, in public goods games, selfish participants who received 'disapproval points' from others—rather than financial punishment—increased their contribution in subsequent rounds [26]. A similar study showed that individuals given the option to assign disapproval points financially punished low contributors less than those not given that option [27]. In another experiment, receivers' reported anger increased when they knew their anger would be conveyed to proposers in an ultimatum game, and proposers responded by offering more money to receivers who expressed more anger [28]. These findings point to the efficacy of emotion expression as a form of punishment.

However, no research, to our knowledge, has examined the benefits of expressing emotions during TTP. We propose that expressing disgust or anger (the primary emotions of moral outrage) enable condemners of non-cooperative behaviour to gain greater reputational benefits than can be gained from employing financial TPP alone, which gives observers little information about the punisher's motives and dispositions.

Whereas some research suggests that anger and disgust are interchangeable [29], or co-occur in blended forms [30], other accounts suggest that they have differing interpersonal effects [31,32]. Whereas anger relates to approach and attack motivational tendencies [33], disgust does not; instead, it relates to indirect aggression [34]. Individuals who express disgust towards a violation are perceived as less motivated by self-interest and more by moral concerns than those who express anger [35]. These findings suggest that disgust and anger expressing third parties might be perceived differently, with disgust-expressing third parties being perceived as less aggressive and more trustworthy than anger expressers.

(b) Research overview

The current research sought to establish whether the expression of moral emotions (disgust and anger) enhances the reputation of third-party punishers, beyond costly (financial) punishment alone. We employed a standard TPP paradigm: the third-party punishment game (TPPG) followed by the trust game (TG) [24]. TPPGs were modified to allow third-parties to respond to selfish dictators by showing emotion expressions, instead of, or in addition to, financial punishment. In the subsequent TG, trustor participants could make money by sending the third party money, but only if the third party subsequently returned more than one-third of the amount sent. Trustors were therefore incentivized to accurately assess the trustworthiness of third parties. Similarly, because third parties knew that a TG followed the TPPG, third parties' reputations were at stake, so they were incentivized to respond in a way that best conveyed trustworthiness.

In Experiment 1, we tested whether expressing anger or disgust in response to a selfish dictator increased trust in third parties relative to third parties who expressed no emotion (represented by a neutral facial expression). In Experiment 2, we examined trust in third parties who expressed anger or disgust in comparison to trust in third parties who punished financially. Finally, in Experiment 3, we tested whether expressing moral emotion while financially punishing improved third-party punishers' reputation relative to financially punishing without emotional expression. In all experiments, trustors and third parties also reported perceptions of trustworthiness and aggressiveness of third parties who made each response. Before making their decisions, participants answered comprehension check questions after reading the TPPG instructions and TG instructions and those who twice answered one or more incorrectly were excluded from analysis. All experiments were conducted with participants recruited online via Prolific and directed to online Qualtrics surveys. Methods, materials, hypotheses and power analyses were pre-registered for all experiments. Preregistration documents, materials, data and syntax are available on the Open Science Framework (https://osf.io/q89b2/).

2. Experiment 1

(a) Methods

(i) Participants

We derived the sample size by conducting a power analysis on the smallest effect size we anticipated – the difference in trust between disgust and anger expressors. For binomial tests of the frequency of anger and disgust expressions chosen by third parties, 199 participants would give 80% power to detect a 10% difference (i.e. 60% versus 50%) in frequency. We increased this number by 10% to account for participants who chose a neutral expression, and we increased that value by 25% to account for exclusions based on failed attention and comprehension checks, resulting in a total of 275 participants per role (dictator, third party and trustor) and 825 in total. All hypotheses and predictions concerned data from participants allocated to third party and trustor roles. Thirty-eight of the 276 participants allocated to the third-party role failed one or more comprehension check questions, leaving 238 participants ($M_{age} = 35.76$, $s.d._{age} = 12.18$; 114 female). Sixty-five of the 277 participants allocated to the trustor role failed one or more comprehension check questions, leaving 212 participants ($M_{age} = 34.02$, $s.d._{age} = 11.27$; 113 female).

(ii) Procedure

In the TPPG, a dictator was endowed with 30 cents and chose to either share that money with a receiver or to keep it for themselves (the 'selfish' decision). Another group of participants was assigned to the third party role, and chose

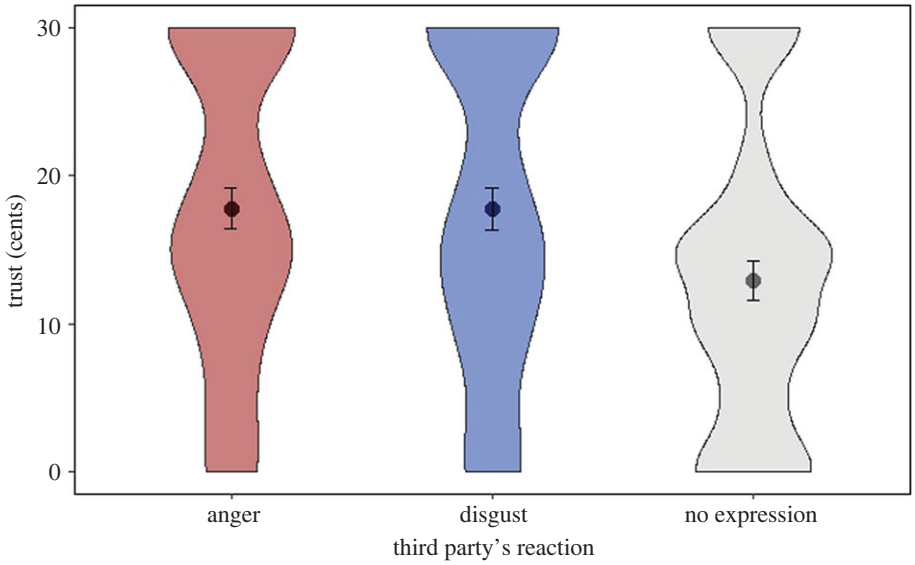

Figure 1. Average amount (in cents) allocated by trustors to third parties who expressed anger, disgust or no expression, in response to a selfish dictator (Experiment 1). Shaded areas of violin plots represent smoothed density of raw data. Points and error bars represent means and 95% confidence intervals, respectively.

to react to a selfish dictator by expressing either anger, disgust or no expression (represented by a neutral facial expression). To represent facial expressions, photographs with the highest validity (i.e. most frequently recognized as the intended expression) were selected from the Radboud Faces Database [36]. Resulting stimuli showed an adult male with an anger, disgust or neutral expression accompanied by verbal labels 'I am angry', 'I am disgusted' or 'no emotion expression', respectively.

A separate group of participants was assigned to the trustor role in a TG. After learning about the third party's response in the TPPG, the trustor was endowed with 30 cents and decided how much to send to the third party, with options varying between 0 and 30 cents in 5-cent increments. The amount sent was tripled, and the third party decided what percentage to return to the trustor. The trustors made three decisions: one for a third party who expressed anger, one for a third party who expressed disgust, and one for a third party who expressed nothing in response to the dictator's selfish decision. They did not learn the behaviour of the third party observer they were paired with until after the experiment was over and payouts were made. Evidence suggests that this 'strategy method', in which participants respond to each of the third party's possible reactions, produces similar behaviour to methods in which participants make only one decision [37].

After making their decisions, trustors reported their perceptions of third parties who had chosen each expression option. Items measuring trustworthiness (trustworthy', 'likeable' and 'cooperative) and aggressiveness (competitive', 'dominant' and 'aggressive) were rated on 7-point scales (ranging from 0, 'not at all', to 6, 'extremely'). Using the same items, third parties rated how they expected trustors to perceive them if they had made each expression decision.

(b) Results

In all experiments, data were analysed using ANOVA. When data violated sphericity, *F*-values and degrees of freedom based on Greenhouse–Geisser corrections are reported, rounded to two decimal places.

(c) Trust

Trustors allocated money differently across expression conditions, $F_{1.47,\ 310.16} = 42.57$, $p < 0.001$, $\eta_p^2 = 0.17$, with more money allocated to disgust expressers ($M = 17.74$, *s.d.* = 10.64), $p < 0.001$, $d = 0.47$, and anger expressers ($M = 17.76$, *s.d.* = 10.21), $p < 0.001$, $d = 0.48$ than to non-expressers ($M = 12.93$, *s.d.* = 9.87) (figure 1). There was no difference in the amount entrusted to disgust and anger expressers, $p > 0.99$, $d = 0.002$.

Trustors' ratings were consistent with these behavioural trust findings. Third parties who expressed disgust ($M = 3.24$, *s.d.* = 1.54) were seen as more trustworthy[1] than anger expressers ($M = 2.97$, *s.d.* = 1.37), $p = 0.001$, $d = 0.19$, and both disgust ($p < 0.001$, $d = 0.51$) and anger ($p < 0.001$, $d = 0.36$) expressers were rated more trustworthy than non-expressing third-parties ($M = 2.46$, *s.d.* = 1.50). Anger expressors ($M = 3.55$, *s.d.* = 1.41) were rated as more aggressive than disgust expressors ($M = 2.89$, *s.d.* = 1.40), $p = 0.001$, $d = 0.47$, and non-expressors ($M = 1.80$, $SD = 1.36$), $p < 0.001$, $d = 1.26$. Disgust expressors were seen as more aggressive than non-expressors, $p < 0.001$, $d = 0.79$. Full details of these analyses and results of third parties' expression decisions, expected trustworthiness and aggressiveness ratings, and trustee decisions, are reported in the electronic supplementary material.

3. Experiment 2

Results from Experiment 1 suggest that expressing moral emotions provides reputational benefits to third parties who observe selfish behaviour. The account we described predicts that third parties who express moral emotions should be trusted more than costly (financial) punishers, because the former communicates motives more effectively than the latter. By contrast, costly signalling accounts [9] predict that financial punishment should be trusted more because it is a more costly signal than emotion expression. Therefore, whereas Experiment 1 showed that emotion-expressing third parties gain reputation benefits over third parties who don't express, Experiment 2 aimed to test whether expressing moral emotions enhances the reputation of third-parties more than costly (i.e. financial) punishment does.

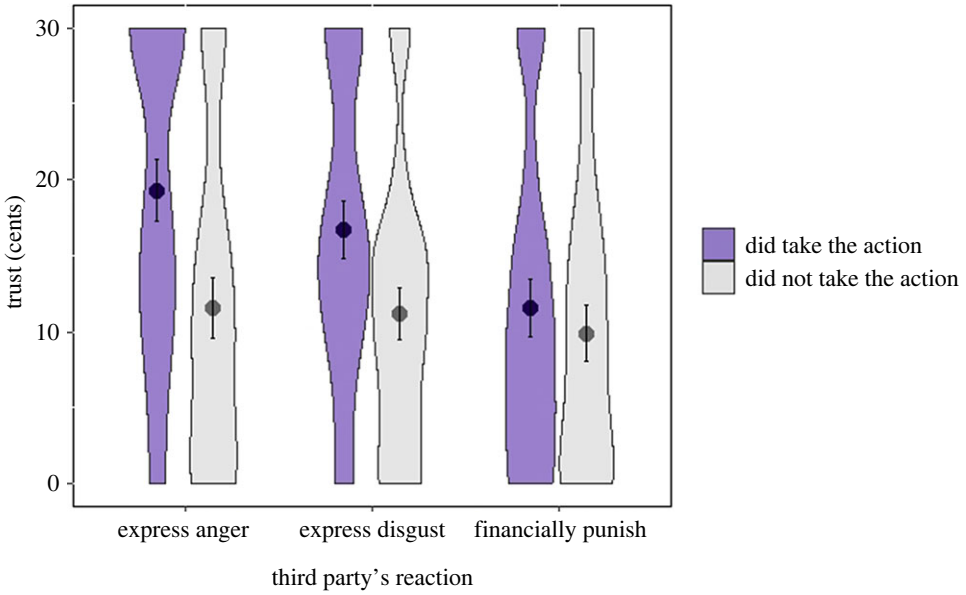

Figure 2. Average amount (in cents) allocated by trustors to third parties who did or did not express anger, express disgust or financially punish, in response to a selfish dictator (Experiment 2). Shaded areas of violin plots represent smoothed density of raw data. Points and error bars represent means and 95% confidence intervals, respectively.

(a) Methods

(i) Participants

We considered the most important effect to be the difference in the amount entrusted to third parties who expressed versus punished financially. In Experiment 1, the effect size for the comparison of expression conditions was $\eta_p^2 = 0.17$. Given that effects comparing expression to financial punishment could be smaller, we sought 95% power to detect an effect size of $\eta_p^2 = 0.05$. To compare disgust expressers, anger expressers, and financial punishers, G*Power 3.1.9.2 recommended a sample size of 297 per role. Estimating that around 25% of participants would fail comprehension check items, we aimed to recruit 371 participants per role (dictator, third party and trustor) giving a total sample size of 1114. All hypotheses and predictions concerned data from participants allocated to the third party and trustor roles. Fifty-eight out of 372 third parties failed one or more comprehension check questions, leaving 314 participants ($M_{age} = 35.56$, $s.d._{age} = 12.38$; 221 female). Fifty-five of 373 trustors failed one or more comprehension check questions, leaving 318 participants ($M_{age} = 33.00$, $s.d._{age} = 11.74$; 178 female).

(ii) Procedure

Similar to Experiment 1, Experiment 2 involved a modified TPPG followed by a TG. However, in the TPPG, after observing a dictator make a selfish decision, participants assigned to the third-party role were randomly assigned to an anger, disgust or financial punishment between-subjects condition. Third parties in the disgust condition chose between expressing disgust or a neutral expression, and in the anger condition between expressing anger or a neutral expression. In the financial punishment condition, which followed standard TPPG procedures [2], third parties were endowed with 20 cents and decided whether to pay 5 cents to cause the dictator to lose 15 cents, or to pay nothing, not causing the dictator to lose anything. To make the financial punishment condition parallel with the expression conditions, neutral facial expressions accompanied both punishment choices. Next, participants assigned to the trustor role in a TG decided how much of their 30-cent endowment to send to a third party after learning about the third party's response in the TPPG. Again, a strategy method was used so that trustors made decisions for both possible third-party responses. After completing both stages, trustors rated the trustworthiness and aggressiveness of third parties using the same items as in Experiment 1. Third parties rated the same items on how they expected trustors would perceive them if they had made each decision.

(b) Results

We detected an interaction between emotional expression and whether third parties took action, $F_{2,\ 315} = 7.55$, $p < 0.001$, $\eta_p^2 = 0.05$. As depicted in figure 2, more money was entrusted to third parties who expressed anger ($M = 19.27$, $SD = 10.51$), $p < 0.001$, $d = 0.76$, or disgust ($M = 16.71$, $s.d. = 9.79$), $p < 0.001$, $d = 0.53$, than to third parties who financially punished ($M = 11.54$, $s.d. = 9.89$). The amount entrusted to anger- and disgust-expressing third parties did not differ, $p = 0.07$, $d = 0.25$ (figure 4).

Moreover, more money was entrusted to third parties who expressed anger than to those who didn't ($M = 11.55$, $s.d. = 10.27$), $p < 0.001$, $d = 0.74$, and more money was entrusted to third parties who expressed disgust than to those who didn't ($M = 11.20$, $s.d. = 8.94$), $p < 0.001$, $d = 0.59$. However, third parties who financially punished were not entrusted with more than those who didn't financially punish, ($M = 9.91$, $s.d. = 9.49$), $p = 0.14$, $d = 0.17$.

Trustors' ratings were consistent with these behavioural trust findings. Third parties who expressed anger ($M = 3.33$, $s.d. = 1.41$) were rated more trustworthy than those who chose not to ($M = 2.34$, $s.d. = 1.42$), $p < 0.001$, $d = 0.70$, and third parties who expressed disgust ($M = 3.27$, $s.d. = 1.45$) were rated more trustworthy than those who chose not to ($M = 2.38$, $s.d. = 1.26$), $p < 0.001$, $d = 0.66$. However, third parties who financially punished ($M = 2.87$, $s.d. = 1.48$) were not rated more trustworthy than those who chose not to financially

punish ($M = 3.27$, $s.d. = 1.64$), $p = 0.07$, $d = 0.26$. Third parties who acted (by expressing or punishing) were rated more aggressive than those who chose not to act, $ps < 0.001$. (Additional details of these analyses, along with third parties' expression decisions and trustworthiness and aggressiveness ratings are reported in the electronic supplementary material.)

4. Experiment 3

Results from Experiment 2 suggest that emotion expression enhances trust more than financial punishment does. But punishment can have tangible costs, such as paying to impose a fine or risking retaliation during a confrontation, and third-party financial punishments model these costs [2]. The costs of third-party punishment might be offset by the signalling benefits of expressing moral emotions revealed in Experiments 1 and 2. Experiment 3 therefore aimed to test whether expressing moral emotions concurrently with financial punishment increases trust compared to financial punishment alone.

(a) Methods

(i) Participants

We considered the most important effect in Experiment 3 to be the amount entrusted between different third-party reaction conditions. In Experiment 2, the effect of condition (anger; disgust; financial punishment) on amount entrusted was $\eta_p^2 = 0.06$. We therefore sought to power Experiment 3 to detect a small effect size ($\eta_p^2 = 0.05$) for any between-subjects difference between third party reaction conditions. For 95% power to detect an effect of this size, G*Power 3.1.9.2 recommended a sample size of 314. Estimating that around 25% of participants would fail comprehension check items, we aimed to recruit 394 participants per role giving a total sample size of 1182. All hypotheses and predictions concerned data from participants allocated to third party and trustor roles. Eighty-four of 417 participants allocated to be third parties failed one or more comprehension check questions, leaving 333 participants ($M_{age} = 35.55$, $s.d._{age} = 12.51$; 154 female). Ninety-nine of 395 participants allocated to be trustors failed one or more comprehension check questions, leaving 296 participants ($M_{age} = 34.38$, $s.d._{age} = 12.13$; 166 female).

(ii) Procedure

After observing a dictator make a selfish decision, a 2 (express or don't) × 2 (financially punish or don't) design rendered four possible options for third parties: express only; financially punish only; express and financially punish; or don't express and don't financially punish.

As an extra between-subjects factor, participants were randomly assigned to conditions in which the available expression was either anger or disgust. Reactions were represented again by faces expressing anger/disgust or neutral expressions and accompanied by verbal labels, along with the relevant financial punishment information represented by the words 'Pay 5 cents to make Player 1 lose 15 cents' or 'Pay nothing and don't cause Player 1 to lose money'. Trustors decided how much to entrust to third parties who made each of the four reaction decisions. Trustors were also randomly assigned to one of two between-subject conditions in which the expression was either anger or disgust. Trustors then rated the trustworthiness and aggressiveness of third parties who had responded in each of the four ways (only express; only punish; express and punish; don't express and don't punish) using the same items as in Experiments 1 and 2. After making their TG decision, third parties rated how trustworthy and aggressive they expected the trustor to have seen them had they made each reaction.

(b) Results

(i) Trust

We first examined the amount trusted to third parties. One interaction emerged: that between whether third parties expressed and whether they punished, $F_{1,\ 294} = 4.23$, $p = 0.04$, $\eta_p^2 = 0.01$, on the amount entrusted to third parties. To assess whether expressing increased trust in punishers, we tested the simple effect of decision to express within decisions to punish. Third parties who punished and also expressed anger or disgust ($M = 16.81$, $s.d. = 9.57$) were entrusted with more than third parties who punished and didn't express ($M = 13.50$, $s.d. = 9.26$), $p < 0.001$, $d = 0.35$ (figure 3). Third parties who didn't punish but expressed anger or disgust ($M = 11.64$, $s.d. = 9.21$) were entrusted with more than third parties who didn't punish and didn't express ($M = 9.48$, $s.d. = 9.47$), $p < 0.001$, $d = 0.23$. Testing the simple effect of decision to punish within the decision to express showed that punishing increased trust in expressing third parties, $p < 0.001$, $d = 0.52$, and also increased trust in non-expressing third parties, $p < 0.001$, $d = 0.39$.

Trustors' ratings were broadly consistent with these behavioural trust findings, albeit with a complex pattern of interactions (see electronic supplementary material for a full description). Most importantly, however, expressers ($M = 3.08$, $s.d. = 0.90$) were rated more trustworthy than non-expressers ($M = 2.86$, $s.d. = 0.88$), $F_{1,\ 295} = 14.79$, $p < 0.001$, $\eta_p^2 = 0.05$. The effect of third parties' expression decisions on trustworthiness and aggressiveness ratings are also reported in the electronic supplementary material.

5. General discussion

Third parties who expressed moral emotions (anger or disgust) towards a dictator who acted selfishly were entrusted with more money than third parties who didn't express (Experiments 1 and 2) and with more money than third parties who financially punished (Experiment 2). Moreover, third parties who expressed anger or disgust *while* engaging in financial punishment were trusted more than those who financially punished without expressing (Experiment 3). Perceptions of third parties were broadly consistent with these behavioural measures of trust revealed in TGs: third parties who expressed anger or disgust were rated as more trustworthy than non-expressing third parties in Experiment 1. In Experiment 2, anger and disgust expressers were rated more trustworthy than third parties who chose not to express, whereas third parties who chose to punish were not rated more trustworthy than those who chose not to punish. In Experiment 3, third parties who expressed anger or disgust while punishing were rated more trustworthy than those who punished without expressing. Overall, these findings support the hypothesis that the expression of moral emotions (anger and disgust) enhances the reputation of third party punishers, beyond costly (financial) punishment.

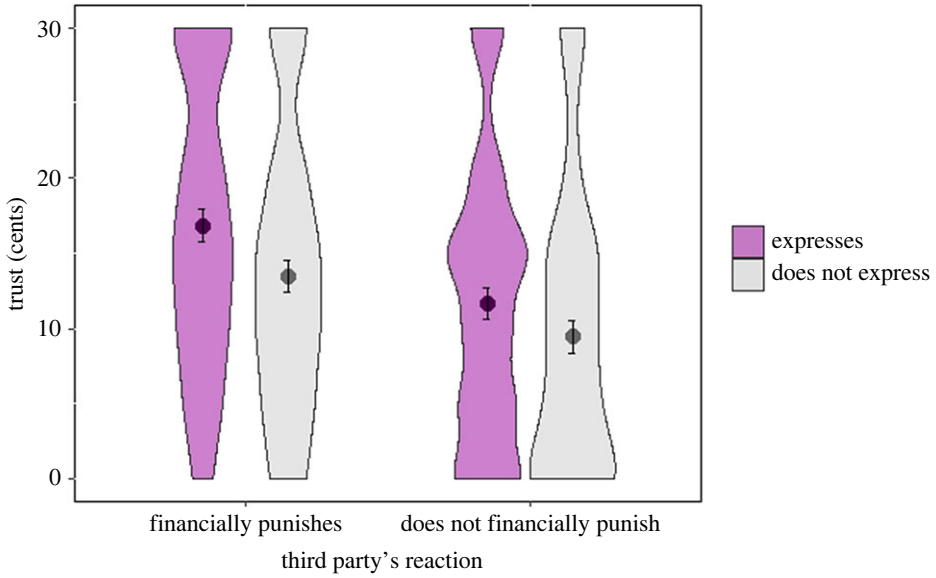

Figure 3. Average amount (in cents) allocated by trustors to third parties who expressed (anger or disgust, collapsed) or not while financially punishing or not, in response to a selfish dictator (Experiment 3). Shaded areas of violin plots represent smoothed density of raw data. Points and error bars represent means and 95% confidence intervals, respectively.

Whereas disgust and anger expressers were consistently trusted more than non-expressers, we did not find greater trust in financial punishers relative to non-punishers, contrary to some previous findings [9]. Our results are more consistent with findings showing that punishers are not trusted more than non-punishers [10,11], which, according to some accounts, is because the motives behind costly punishment are ambiguous [13]. An exception was that in Experiment 3, in which third parties chose whether to engage in financial punishment or not *and* whether to express or not, both financial punishment and expression increased trust. This pattern of findings is consistent with the idea that emotion expressions reduce uncertainty about the intentions behind financial punishment, and thereby increase the likelihood that punishers will be trusted. Multiple theoretical accounts of emotion expression have theorized that emotion expression evolved in part to efficiently convey social information about motives and intentions [17,20–22], a function that may be particularly important in the context of third-party punishment. By enhancing reputation, emotion expression may enable people to offset the costs of punishing third parties.

However, although trustors' behaviour and self-reported perceptions indicated that they expected anger and disgust expressing third parties to be more trustworthy and co-operative, other inferences could have been made regarding third parties' motives. Third parties may punish due to principles such as retribution or to achieve consequences such as deterrence [38] and people perceive others to punish for both reasons [39]. Emotion expressions communicate information about motivational dispositions and behavioural tendencies [35]. One possibility is that third parties who express moral emotions while punishing are perceived to do so for more praiseworthy reasons. Further research is needed to identify which perceptions explain the reputation-enhancing effects of emotion expression on third party punishment.

Expressing emotions is not the only means of increasing the likelihood that TPP leads to reputational payoffs. If punishment is carried out by institutions or coalitions, rather than individuals, TPP can be seen as more legitimate [40], but this type of punishment is often unavailable during informal or small group situations. Third parties can also communicate condemnation of non-cooperative individuals through gossip [41], or communicate their own virtue by compensating the victim instead of punishing a non-cooperator [7,42]. However, because these alternatives are not directed at the non-cooperator, they may not promote cooperative behaviour with the same efficiency as direct punishment [1,2]. Future research could examine how well expression of emotion by third parties increases trust in comparison to other mechanisms such as gossip and compensation.

Numerous studies have shown that punishment of non-cooperators is an effective way to promote cooperation [1]. However, future research is needed to examine whether punishment with concurrent expression also promotes cooperation. A further possibility is that emotion expression could by itself efficiently promote cooperation, without the need for more costly forms of punishment. The notion that emotion expression could serve as a low-cost form of third-party punishment has been proposed [24], but remains to be tested.

We also assessed whether anger and disgust expressing third parties would be perceived and treated differently from each other. We did not detect a difference in the amount sent by trustors to anger versus disgust expressors in the TG in any of the experiments. In Experiment 1, but not in Experiment 2, third parties who expressed disgust were perceived as more trustworthy and less aggressive than anger expressers. This finding is consistent with prior research showing that disgust expressers are perceived as less self-interested and more morally motivated than anger expressers [35] and as less aggressive than anger expressers [34]. Future research is needed to understand the situations in which the differing but overlapping information communicated by anger and disgust can lead to different behavioural consequences. Additionally, our research used expression images from only one adult male individual. Although this approach yielded findings in line with predictions, future research could examine how effects vary when different stimuli are used, such as expressions produced by individuals of different gender, race or age. Likewise, all participants in our studies were from Western, English-speaking

countries. Findings could vary if the research was conducted with participants from other cultures, especially those with different punishment [43] or emotion expression [44] norms.

Abundant research has shown that TPP is motivated by emotions, specifically moral outrage which consists of anger and disgust [2,19]. However, to our knowledge, the experiments reported here are the first to investigate the reputational effects of third parties showing these emotions. We conclude that expressing disgust or anger enables condemners of non-cooperative behaviour to gain greater reputational benefits than can be gained from employing financial TPP alone. By reducing the ambiguity of a punisher's motives and intentions, the expression of moral emotions may contribute to the stability of human cooperation.

Ethics. This work received ethical approval from the Vrije Universiteit ethics board.

Data accessibility. The data are provided in electronic supplementary material [45].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. T.R.K.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, writing—original draft; J.M.T.: conceptualization, funding acquisition, investigation, methodology, writing—review and editing.

Both authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This work was supported by European Research Council grants to Tom R. Kupfer (MSCA-IF-2017-800096-EmoPun) and Joshua M. Tybur (StG-2015-680002-HBIS).

Endnote

[1]Factor analysis revealed trustworthiness and aggressiveness items formed two factors comprising the expected items in all three experiments. See electronic supplementary material for details.

References

1. Balliet D, Mulder LB, Van Lange PA. 2011 Reward, punishment, and cooperation: a meta-analysis. *Psychol. Bull.* **137**, 594. (doi:10.1037/a0023489)
2. Fehr E, Gächter S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140. (doi:10.1038/415137a)
3. Balafoutas L, Grechenig K, Nikiforakis N. 2014 Third-party punishment and counter-punishment in one-shot interactions. *Econ. Lett.* **122**, 308–310. (doi:10.1016/j.econlet.2013.11.028)
4. Eriksson K, Andersson PA, Strimling P. 2016 Moderators of the disapproval of peer punishment. *Group Process. Intergroup Relations* **19**, 152–168. (doi:10.1177/1368430215583519)
5. Dreber A, Rand DG, Fudenberg D, Nowak MA. 2008 Winners don't punish. *Nature* **452**, 348–351. (doi:10.1038/nature06723)
6. Ohtsuki H, Iwasa Y, Nowak MA. 2009 Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature* **457**, 79–82. (doi:10.1038/nature07601)
7. Raihani NJ, Bshary R. 2015 Third-party punishers are rewarded, but third-party helpers even more so. *Evolution* **69**, 993–1003. (doi:10.1111/evo.12637)
8. Barclay P. 2006 Reputational benefits for altruistic punishment. *Evol. Hum. Behav.* **27**, 325–344. (doi:10.1016/j.evolhumbehav.2006.01.003)
9. Jordan JJ, Hoffman M, Bloom P, Rand DG. 2016 Third-party punishment as a costly signal of trustworthiness. *Nature* **530**, 473–476. (doi:10.1038/nature16981)
10. Horita Y. 2010 Punishers may be chosen as providers but not as recipients. *Lett. Evol. Behav. Sci.* **1**, 6–9. (doi:10.5178/lebs.2010.2)
11. Kiyonari T, Barclay P. 2008 Cooperation in social dilemmas: free riding may be thwarted by second-order reward rather than by punishment. *J. Pers. Soc. Psychol.* **95**, 826. (doi:10.1037/a0011381)
12. Rockenbach B, Milinski M. 2011 To qualify as a social partner, humans hide severe punishment, although their observed cooperativeness is decisive. *Proc. Natl Acad. Sci. USA* **108**, 18 307–18 312. (doi:10.1073/pnas.1108996108)
13. Raihani NJ, Bshary R. 2015 The reputation of punishers. *Trends Ecol. Evol.* **30**, 98–103. (doi:10.1016/j.tree.2014.12.003)
14. Jensen K. 2010 Punishment and spite, the dark side of cooperation. *Phil. Trans. R. Soc. B* **365**, 2635–2650. (doi:10.1098/rstb.2010.0146)
15. Przepiorka W, Liebe U. 2016 Generosity is a sign of trustworthiness—the punishment of selfishness is not. *Evol. Hum. Behav.* **37**, 255–262. (doi:10.1016/j.evolhumbehav.2015.12.003)
16. Krasnow MM, Delton AW, Cosmides L, Tooby J. 2016 Looking under the hood of third-party punishment reveals design for personal benefit. *Psychol. Sci.* **27**, 405–418. (doi:10.1177/0956797615624469)
17. Frank RH. 1988 *Passions within reason: the strategic role of the emotions*. New York, NY: Norton.
18. Loewenstein G. 2000 Emotions in economic theory and economic behavior. *Am. Econ. Rev.* **90**, 426–432. (doi:10.1257/aer.90.2.426)
19. Jordan JJ, Rand DG. 2020 Signaling when no one is watching: a reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *J. Pers. Soc. Psychol.* **118**, 57. (doi:10.1037/pspi0000186)
20. Fessler DM, Haley KJ. 2003 The strategy of affect. In *The genetic and cultural evolution of cooperation* (ed. P Hammerstein), pp. 7–36. Cambridge, MA: MIT Press.
21. Fischer AH, Manstead AS. 2008 Social functions of emotion. *Handb. Emot.* **3**, 456–468.
22. Keltner D, Haidt J. 1999 Social functions of emotions at four levels of analysis. *Cogn. Emot.* **13**, 505–521. (doi:10.1080/026999399379168)
23. Van Kleef GA. 2009 How emotions regulate social life: the emotions as social information (EASI) model. *Curr. Direct. Psychol. Sci.* **18**, 184–188. (doi:10.1111/j.1467-8721.2009.01633.x)
24. Fehr E, Fischbacher U. 2004 Third-party punishment and social norms. *Evol. Hum. Behav.* **25**, 63–87. (doi:10.1016/S1090-5138(04)00005-4)
25. Xiao E, Houser D. 2005 Emotion expression in human punishment behavior. *Proc. Natl Acad. Sci. USA* **102**, 7398–7401. (doi:10.1073/pnas.0502399102)
26. Dickinson DL, Masclet D. 2015 Emotion venting and punishment in public good experiments. *J. Pub. Econ.* **122**, 55–67. (doi:10.1016/j.jpubeco.2014.10.008)
27. Masclet D, Noussair C, Tucker S, Villeval MC. 2003 Monetary and nonmonetary punishment in the voluntary contributions mechanism. *Am. Econ. Rev.* **93**, 366–380. (doi:10.1257/000282803321455359)
28. Andrade EB, Ho TH. 2009 Gaming emotions in social interactions. *J. Consumer Res.* **36**, 539–552. (doi:10.1086/599221)
29. Nabi RL. 2002 The theoretical versus the lay meaning of disgust: implications for emotion research. *Cogn. Emot.* **16**, 695–703. (doi:10.1080/02699930143000437)
30. Salerno JM, Peter-Hagene LC. 2013 The interactive effect of anger and disgust on moral outrage and judgments. *Psychol. Sci.* **24**, 2069–2078. (doi:10.1177/0956797613486988)
31. Giner-Sorolla R, Kupfer T, Sabo J. 2018 What makes moral disgust special? An integrative functional review. *Adv. Exp. Soc. Psychol.* **57**, 223–289. (doi:10.1016/bs.aesp.2017.10.001)
32. Tybur JM, Lieberman D, Kurzban R, DeScioli P. 2013 Disgust: evolved function and structure. *Psychol. Rev.* **120**, 65. (doi:10.1037/a0030778)
33. Carver CS, Harmon-Jones E. 2009 Anger is an approach-related affect: evidence and implications. *Psychol. Bull.* **135**, 183. (doi:10.1037/a0013965)
34. Molho C, Tybur JM, Güler E, Balliet D, Hofmann W. 2017 Disgust and anger relate to different aggressive responses to moral violations. *Psychol. Sci.* **28**, 609–619. (doi:10.1177/0956797617692000)
35. Kupfer TR, Giner-Sorolla R. 2017 Communicating moral motives: the social signaling function of

disgust. *Soc. Psychol. Person. Sci.* **8**, 632–640. (doi:10.1177/1948550616679236)

36. Langner O, Dotsch R, Bijlstra G, Wigboldus DH, Hawk ST, Van Knippenberg AD. 2010 Presentation and validation of the Radboud Faces Database. *Cogn. Emot.* **24**, 1377–1388. (doi:10.1080/02699930903485076)
37. Jordan J, McAuliffe K, Rand D. 2016 The effects of endowment size and strategy method on third party punishment. *Exp. Econ.* **19**, 741–763. (doi:10.1007/s10683-015-9466-8)
38. Crockett MJ, Özdemir Y, Fehr E. 2014 The value of vengeance and the demand for deterrence. *J. Exp. Psychol.: Gen.* **143**, 2279. (doi:10.1037/xge0000018)
39. Marshall J, Gollwitzer A, Bloom P. 2022 Why do children and adults think other people punish? *Dev. Psychol.* **58**, 1783–1792. (doi:10.1037/dev0001378)
40. Pfattheicher S, Boehm R, Kesberg R. 2018 The advantage of democratic peer punishment in sustaining cooperation within groups. *J. Behav. Decision Making* **31**, 562–571. (doi:10.1002/bdm.2050)
41. Wu J, Balliet D, Van Lange PA. 2016 Gossip versus punishment: the efficiency of reputation to promote and maintain cooperation. *Sci. Rep.* **6**, 1–8. (doi:10.1038/s41598-016-0001-8)
42. Batistoni T, Barclay P, Raihani NJ. 2022 Third-party punishers do not compete to be chosen as partners in an experimental game. *Proc. R. Soc. B* **289**, 20211773. (doi:10.1098/rspb.2021.1773)
43. Eriksson K *et al.* 2021 Perceptions of the appropriate response to norm violation in 57 societies. *Nat. Commun.* **12**, 1481. (doi:10.1038/s41467-021-21602-9)
44. Maitner AT *et al.* 2022 Perceptions of emotional functionality: similarities and differences among dignity, face, and honor cultures. *J. Cross-Cult. Psychol.* **53**, 263–288. (doi:10.1177/00220221211065108)
45. Kupfer TR, Tybur JM. 2023 Third-party punishers who express emotions are trusted more. Figshare. (doi:10.6084/m9.figshare.c.6777772)