NeuroImage 245 (2021) 118730

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

From gratitude to injustice: Neurocomputational mechanisms of gratitude-induced injustice

Ruida Zhu[a], Zhenhua Xu[b,c,d], Song Su[a], Chunliang Feng[e], Yi Luo[f], Honghong Tang[a], Shen Zhang[b,c,d], Xiaoyan Wu[b,c,d], Xiaoqin Mai[g], Chao Liu[b,c,d,*]

[a] *Business School, Beijing Normal University, Beijing 100875, China*
[b] *State Key Laboratory of Cognitive Neuroscience and Learning and IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing 100875, China*
[c] *Centre for Collaboration and Innovation in Brain and Learning Sciences, Beijing Normal University, Beijing 100875, China*
[d] *Beijing Key Laboratory of Brain Imaging and Connectomics, Beijing Normal University, Beijing 100875, China*
[e] *Guangdong Provincial Key Laboratory of Mental Health and Cognitive Science, Centre for Studies of Psychological Application, School of Psychology, South China Normal University, Guangzhou 510631, China*
[f] *Fralin Biomedical Research Institute at VTC, Virginia Tech, Roanoke, USA*
[g] *Department of Psychology, Renmin University of China, Beijing 100872, China*

ARTICLE INFO

Keywords:
Gratitude
Protection tendency
Injustice
Mentalizing
Reward processing

ABSTRACT

Gratitude shapes individuals' behaviours and impacts the harmony of society. Many previous studies focused on its association with prosocial behaviours. A possibility that gratitude can lead to moral violation has been overlooked until recently. Nevertheless, the neurocognitive mechanisms of gratitude-induced moral violation are still unclear. On the other hand, though neural correlates of the gratitude's formation have been examined, the neural underpinnings of gratitude-induced behaviour remain unknown. For addressing these two overlapped research gaps, we developed novel tasks to investigate how participants who had received voluntary (Gratitude group) or involuntary help (Control group) punished their benefactors' unfairness with functional magnetic resonance imaging (fMRI). The Gratitude group punished their benefactors less than the Control group. The self-report and computational modelling results demonstrated a crucial role of the boosted protection tendency on behalf of benefactors in the gratitude-induced injustice. The fMRI results showed that activities in the regions associated with mentalizing (temporoparietal junction) and reward processing (ventral medial prefrontal cortex) differed between the groups and were related to the gratitude-induced injustice. They suggest that grateful individuals concern for benefactors' benefits, value chances to interact with benefactors, and refrain from action that perturbs relationship-building (i.e., exert less punishment on benefactors' unfairness), which reveal a dark side of gratitude and enrich the gratitude theory (i.e., the find-bind-remind theory). Our findings provide psychological, computational, and neural accounts of the gratitude-induced behaviour and further the understanding of the nature of gratitude.

1. Introduction

Gratitude, a typical emotional response to others' kindness, is of great importance in social, moral and religious aspects of life (McCullough, 2004). It has long been considered "the parent of other virtues" (Cicero, 1851) and "moral emotion" (McCullough and Kilpatrick, 2001). Numerous behavioural studies have demonstrated that grateful individuals are more inclined to engage in prosocial behaviours (see a review, Ma et al., 2017).

A classic theory of gratitude (i.e., the find-bind-remind theory) proposes the social function of gratitude to be interpersonal relationship enhancement with high-quality partners (Algoe, 2012). Conducting prosocial behaviours, however, is not the only way to enhance interpersonal relationship. Recently, researchers noticed that a dark side of gratitude might exist theoretically (Wood et al., 2016). A behavioural study found grateful individuals were willing to violate moral norms for protecting their benefactors (Zhu et al., 2020). The findings start to change the positive stereotype of gratitude. Nevertheless, the neurocognitive mechanisms of gratitude-induced moral violation are unclear.

Based on existing literature, the implement of the social function of gratitude (i.e., interpersonal relationship enhancement with high-quality partners) involves three steps (e.g., Algoe, 2012; Algoe et al., 2008; Yu et al., 2018). In the first step, to find a high-quality partner,

[*] Corresponding author at: State Key Laboratory of Cognitive Neuroscience and Learning and IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing 100875, China.

E-mail address: liuchao@bnu.edu.cn (C. Liu).

https://doi.org/10.1016/j.neuroimage.2021.118730.
Received 23 June 2021; Received in revised form 21 October 2021; Accepted 13 November 2021
Available online 14 November 2021.

individuals collect information on three relational features (i.e., cognitive antecedents of gratitude), including the intent of partners, the cost to partners for offering help, and the benefit individuals receive from the help (Tesser et al., 1968). In the second step, the individuals integrate the collected relational information and generate gratitude accordingly, which reminds individuals that high-quality partners are identified (Algoe, 2012; Yu et al., 2018). In the third step, gratitude guides individuals to take actions to promote a relationship in a specific situation, which eventually binds individuals and partners together in the relationship (Algoe, 2012; Bartlett et al., 2012).

Neuroimaging studies have investigated the neural correlates of evaluating the cognitive antecedents of gratitude and generating gratitude (i.e., the first and second steps). As to evaluating the cognitive antecedents of gratitude, Yu et al. (2018) revealed that the monetary cost to a partner was represented in a brain region related to mentalizing (temporoparietal junction, TPJ) and that the benefit for oneself was represented in a region related to reward processing (ventral striatum). Similarly Xiong et al. (2020) showed that the activity in a region important for mentalizing (dorsal medial prefrontal cortex) correlated with the level of cost that partners bore to provide help. These findings suggest that the cognitive antecedents of gratitude are represented by brain regions associated with reward processing and mentalizing.

As to generating gratitude, in an imagination task, Fox et al. (2015) found that the participants' feelings of gratitude were correlated with activity in a region implicated in reward processing and value representation (e.g., ventral medial prefrontal cortex, vmPFC[1]). In another imagination study, Decety et al. (2011) showed that compared to imagining help others, imagining being helped (which probably generated gratitude feelings) activated several brain regions including the vmPFC. Likewise, in interpersonal tasks (other than imagination tasks), researchers identified a positive correlation between gratitude and activation of the vmPFC when the participants received real help (Xiong et al., 2020; Yu et al., 2017). In addition, Liu et al. (2020) used an interpersonal paradigm to demonstrate that gratitude compared to joy activated the brain regions associated with reward processing (putamen) and mentalizing (precuneus). It is indicated that gratitude generation may also be implemented by reward processing and mentalizing regions, although the specific regions are not exactly the same as those related to cognitive antecedents of gratitude. Despite these findings, the neural mechanisms that directly drive grateful individuals to act for relationship enhancement (i.e., the third step) are unknown.

The two research gaps mentioned above (i.e., neurocognitive mechanisms of gratitude-induced moral violation and neural mechanisms of gratitude-driven action for relationship enhancement) are overlapped. To fill in both, we either induced participants' gratitude towards a partner (Gratitude group) or not (Control group) in one task and then measured participants' punishment for their partner's unfairness to strangers in another task with fMRI scanning (Fig. 1). We predicted that gratitude lowers due punishment for partners' unfairness (i.e., moral violation; a dark side of gratitude) and focused on how it happens (i.e., cognitive, computational, and neural mechanisms). Three hypotheses are possible for the mechanism contributing to this effect. One is that grateful individuals perceive a chance interacting with high-quality benefactors to be valuable, concern about benefactors' benefits, and refrain from action that perturbs relationship building (e.g., decrease punishment for benefactors' unfairness). This hypothesis dovetails with the find-remind-bind theory (Algoe, 2012). In this case, gratitude may modulate activity in regions associated with rewarding processing and value representation (e.g., vmPFC) (Bartra et al., 2013; Yu et al., 2017) and mentalizing (e.g., TPJ; the region plays a vital role in moral action, such as punishment) (Bellucci et al., 2020) when individuals have a chance to build relationship with their benefactors (i.e., making a decision of less punishment on benefactors).

The other hypothesis is that grateful individuals have a distorted perception of their benefactors' unfair distributions and exert less punishment. It has been demonstrated that emotions influence moral perception (judgement) (Horberg et al., 2011; Ugazio et al., 2012). For example, shame prompts individuals to concentrate on the moral self and blame themselves more strongly (Tangney and Dearing, 2003). Given that gratitude facilitates relationship-building with benefactors (Algoe, 2012), grateful individuals may perceive the unfair distributions of their benefactors to be less unfair. The distorted unfairness perception can be a determining factor for punishment decisions (e.g., Ruff et al., 2013). According to this hypothesis, gratitude is likely to influence activity in the brain regions associated with unfairness perception (e.g., insula and anterior cingulate cortex) (Zhong et al., 2016) when individuals view benefactors' distributions.

The still other hypothesis is that grateful individuals who are in a positive mood punish less. Grateful thinking keeps individuals in a positive mood within a certain time frame (Sheldon and Lyubomirsky, 2006; Watkins et al., 2003). Convergent evidence has shown that incident moods can affect punishment for unfairness (Bonini et al., 2011; Gummerum et al., 2016). For instance, incident anger increases individuals' punishment for unfair distributions (Gummerum et al., 2016). Grateful individuals may misattribute the positive mood induced by a previous event to the benefactors' distributions they see, which attenuates anger and decreases punishment. According to this hypothesis, we may observe that gratitude increases activation in the regions related to a positive mood (e.g., striatum; Suardi et al., 2016; Vytal and Hamann, 2010), regardless of when individuals view benefactors' distributions or decide a punishment on benefactors, as moods are long-lasting background states (Beedie et al., 2005).

We examined these hypotheses combining the approaches of self-report, computational models, and fMRI imaging and attempted to delineate a neurocognitive profile of gratitude-induced injustice.

2. Methods

2.1. Participants

A behavioural experiment using a similar paradigm[2] was conducted before the current fMRI experiment (Zhu et al., 2020). In the behavioural experiment, the effect of gratitude (Gratitude versus Control group) on punishment was $f = 0.36$. Based on a Gpower estimation ($f = 0.36$, $\alpha = 0.05$, 1- $\beta = 0.80$, two groups; Faul et al., 2007), we determined the minimum sample size of the current study to be 64 participants. Assuming that 5–10% of participants would be removed from the analyses due to doubt of the authenticity of the experiment or technical issues, we recruited 71 university students (Gratitude: 35 participants; Control: 36 participants). Participants had normal or corrected-to-normal vision and reported no history of neurological or psychiatric diseases. All participants provided written consent prior to the experiment and received payment for their participation. The experimental protocol was approved by the Institutional Review Board of the Department of Psychology at Renmin University of China (Beijing, China), and complied with the Declaration of Helsinki.

One participant in the Gratitude group and one participant in the Control group who doubted the authenticity of the experiment were excluded from all the analyses.[3] Two other participants in the Grat-

[1] We note that the cluster identified by Fox et al. (2015) was large, which contained the bilateral vmPFC/OFC, dmPFC, and ACC. In their article, they named the cluster "mPFC".

[2] In this behavioural experiment ($N = 62$), the participants punished the allocator at the cost of their own payoff, whereas the participants in the current fMRI experiment could punish the allocator without any cost.

[3] They reported that they doubted whether they interacted with a real person during the games, when they were asked whether they had any question about the experiment at the very end of the experiment.

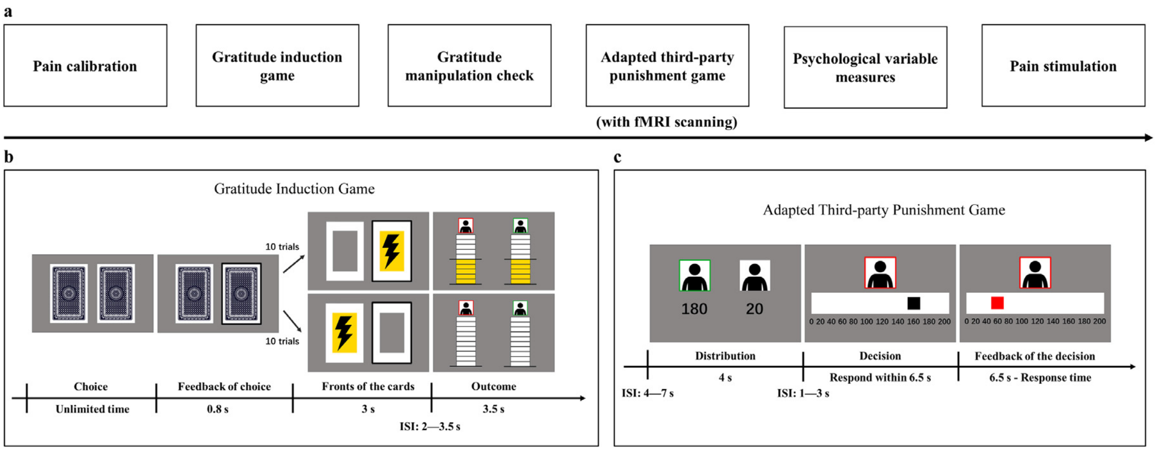

Fig. 1. Experimental procedures. **a**, Timeline for the whole experimental session. **b**, Timeline for the gratitude induction game. The participant chose one of the two covered cards. If an electric-shock card was chosen, the participant faced a risk of receiving a strong electric shock. If a blank card was chosen, no electric shock was administered. For all the participants, in 10 trials, they chose a blank card, and nothing happened; in the other 10 trials, they chose an electric-shock card but only received a moderate (rather than strong) electric shock. To manipulate gratitude, participants in the Gratitude group were informed that a partner (i.e., a confederate) voluntarily helped them by taking part of painful shock for them, whereas participants in the Control group were informed that a partner was required to help them by the rule. The bar filled with five yellow blocks indicated a moderate electric shock, whereas the empty bar indicated no electric shock. **c,** Timeline for the adapted third-party punishment game. The partner in the gratitude induction game distributed monetary tokens between themselves and a stranger. After viewing the distribution, the participant was entitled to deduct tokens from the partner without any cost. **b,c,** The partner, stranger, and participant were represented by the agents surrounded by a green, grey, and red frame, respectively. ISI, inter-stimulus interval. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

itude group who did not elicit punishment in any of the trials (i.e., all punishment decisions were zero) were excluded from the computational modelling and neuroimaging analyses, as the parameters could not be properly estimated in such a case (e.g., Stallen et al., 2018; Zhong et al., 2016). Thus, the behavioural analyses were conducted with 69 participants (Gratitude: 34 participants, 15 females and 19 males, M_{age} = 21.71 years, SD_{age} = 2.17; Control: 35 participants, 16 females and 19 males, M_{age} = 21.46 years, SD_{age} = 1.84), and the computational modelling and neuroimaging analyses were conducted with 67 participants (Gratitude: 32 participants, 15 females and 17 males, M_{age} = 21.72 years, SD_{age} = 2.23; Control: 35 participants, 16 females and 19 males, M_{age} = 21.46 years, SD_{age} = 1.84).

2.2. Experimental procedures

Each participant came to the lab individually. Upon arrival, the participant was introduced to a same-sex partner (i.e., a confederate of our lab), who would (ostensibly) interact with the participant later via an internal network. The participant was then led to an independent testing room and instructed to complete a series of tasks. Our confederate confirmed that he or she did not know and did not have a personal relationship with the participant before this experiment.

2.2.1. Pain calibration

First, the participant underwent an individual pain calibration procedure with an SXC-4A multichannel electrical stimulator. This is an important step for inducing gratitude in the following gratitude induction game. After an introduction of the calibration process and precaution, we cleaned the participant's left forearm and placed two disposable electrodes on the back of their left wrist at an interval of 5 cm. The pain stimulation was set as three repeated square waveform electrical stimulation pulses (duration of a pulse: 0.5 ms; interval between pulses: 10 ms) (e.g., Yu et al., 2014). Pain calibration began with a weak electric shock (0.5 mA) and then gradually increased or decreased in small increments (0.5 mA) with a 3:1 ratio (Crockett et al., 2014). The participant rated their subjective experience after each shock on an 11-point scale (0 = no sensation, 10 = maximum tolerable pain). The calibration continued until a rating of 8 was reported. Electric current values corresponding to self-reported pain ratings of 4 and 7 were respectively recorded as moderate and strong pain stimuli for the participant and were used in the subsequent gratitude induction game (e.g., Yu et al., 2017).

2.2.2. Gratitude induction game

We used a novel gratitude induction game to manipulate gratitude, the development of which was inspired by a previous study (Yu et al., 2017). In this game, the participant saw the backs of the two cards and chose one of them. Afterwards, the cards were turned over. If an electric-shock card was chosen, the participant faced a risk of receiving a strong electric shock. If a blank card was chosen, no electric shock was administered. The partner could see the participant's choice on another computer. Unknown to the participants, they were randomly assigned to the Gratitude group or the Control group. The participants in the Gratitude group were told that the partner could decide whether to help them by taking part of the electric shock when the electric-shock card was chosen. If the partner decided to help, the participant received a moderate electric shock instead of a strong one, and the partner received a moderate electric shock. If the partner refused to help, the participant received a strong electric shock, and the partner did not receive any shock. Different from the Gratitude group, the participants in the Control group were informed that the rules of the game required the partner to help the participant (i.e., the partner could not refuse to help). To give the participants first-hand experience, moderate and strong pain stimuli (individually calibrated; see the pain calibration section) were administered to them before the gratitude induction game. All the participants reported that the two levels of pain stimuli were clearly distinguishable.

This game contained 20 trials. For all the participants, they chose a blank card in 10 trials, and no one received any shock; in the other 10 trials, they chose an electric-shock card, and both the participant and the partner would receive a moderate electric shock. Thus, the outcome of the game was exactly the same for the two groups. The only difference was that in the Gratitude group, the partner helped voluntarily (benevolent intent), while in the Control group, the partner was required to help (ambiguous intent). The electric shocks were not administered during the gratitude induction game but were given to the participants at the end of the experiment (see the pain stimulation section). This helped to circumvent a confounding effect of pain-evoked anger on participants' punishment decisions in a later task (see below), given that pain evokes anger (Berkowitz, 1993; Trost et al., 2012) and anger enhances third-party punishment (Gummerum et al., 2016; Nelissen and Zeelenberg, 2009). Participants completed the gratitude induction game outside the scanner room without fMRI scanning.

2.2.3. Gratitude manipulation check

After the gratitude induction game, we asked the participant to complete a questionnaire. The participant rated how grateful, thankful, shameful, guilty, unhappy, disappointed, indebted, obligated (to repay), and angry they felt towards the partner. Gratitude and indebtedness were measured by two items (gratitude items: grateful and thankful; Cronbach's $\alpha = 0.970$; indebtedness items: indebted and obligated; Cronbach's $\alpha = 0.772$) (e.g., Peng et al., 2018). Each of the other emotions was measured by one corresponding item. The participant also reported their thoughts about (i) to what extent the partner was willing to help, (ii) to what extent the partner concerned for the participant's benefits, (iii) to what extent the partner concerned for the partner's own benefits, (iv) how much cost the partner undertook due to the help, (v) how much benefit the partner brought to the participant, (vi) to what extent a friendship with the partner would benefit the participant in the future and (vii) to what extent the participant was inclined to build a friendship with the partner on a nine-point scale (1 = not at all, 9 = very much).

2.2.4. Adapted third-party punishment game (with fMRI scanning)

Afterwards, participants went into the MRI machine and completed an adapted third-party punishment game (*adapted TPP*). Similar to the typical third-party punishment game (e.g., Fehr and Fischbacher, 2004), there were three players—an allocator, a receiver, and a decider—in the adapted TPP. In each trial, the allocator could freely distribute 200 tokens between the receiver and themselves (minimum unit: 10 tokens). The receiver had to accept the distribution. The distribution was then shown to the decider. The decider, owning 100 tokens, had a chance to deduct the allocator's tokens (options: deducting 0 to 200 tokens, in increments of 20) (the deducted tokens disappeared and did not go to the receiver or decider). Notably, different from the typical TPP that the decider deducted the allocator's tokens by spending their own tokens, in the adapted TPP, the decider could deduct the allocator's tokens without any cost. This change enabled us to better explain and model the decider's choices, given that the cost to the decider and the benefit of the allocator were negatively correlated with each other and could not be disentangled in a typical TPP.

In all trials, the participant and the previous partner in the gratitude induction game played as the decider and the allocator, respectively. The receiver in different trials was played by different strangers. In each trial, the participant saw one of the five types of unfair distributions, in which the partner distributed 200 tokens between themselves and a stranger (i.e., 200 (for the partner): 0 (for the stranger), 180:20, 160:40, 140:60, or 120:80; each type repeated 10 times across the game). Different distributions were presented in a pseudorandom order. To justify the different distributions made by the partner, we told a cover story to the participant. It was informed that in each trial, the partner could see a portrait of the current receiver before the distribution, but the participant could not. Though these strangers were not present in the lab at the time, they had visited the lab, spent time understanding the rule of the adapted TPP, provided their portraits, and agreed to participate as the receivers. The participant needed to indicate how many tokens they wanted to deduct from the partner by moving a black block. The initial place of the block in each trial was random, which was helpful to balance the motor action between the Gratitude and Control groups. When the participant confirmed a punishment decision by pressing a button, the black block turned red. Participants were informed that the partner did not know how many tokens were deducted from them until all trials were completed.[4] Thus, participants would not expect an influence of their punishment decisions on the partner's distributions during the game. Notably, we did not use the term "punishment" in the introduction of the game, but instead told the participants that they had chances to deduct tokens from the allocator (Fehr and Fischbacher, 2004). It was informed that at the end of the experiment the tokens were converted into real money, and the players were paid based on the amount of tokens they obtained, in addition to a fixed show-up fee. Using the adapted TPP, our study probed the effect of gratitude on participants' trade-off between the partner's benefit and inequity aversion.

2.2.5. Psychological variable measures

To investigate potential psychological contributors to gratitude-induced differences in punishment, which helps to examine our three hypotheses, participants completed a questionnaire after finishing the adapted TPP. They reported their psychological processes and feelings (one item for each variable), including (i) to what extent they wanted to reduce the allocator's losses, (ii) to what extent they concerned for the justice of their own behaviour, (iii) to what extent they concerned for the receivers' losses, (iv) to what extent they concerned whether the behaviour of the allocator was immoral, (v) to what extent they concerned whether the behaviour of the allocator was unfair, (vi) to what extent the allocator's behaviour was immoral, (vii) to what extent the allocator's behaviour was unfair, (viii) to what extent they felt angry, (ix) to what extent they felt glad, and (x) to what extent they felt excited (regardless of in a positive or negative way) on a nine-point scale (1 = not at all, 9 = very much) (see Supplementary Table 1).

2.2.6. Pain stimulation

Ten moderate electric shocks were administered on the left forearm of the participant. The strength of the shocks for each participant had been individually calibrated. The participant rated how painful the electric shocks were after the stimulation. No significant difference in pain ratings was found between the Gratitude (M = 4.76, SD = 1.69) and Control (M = 5.37, SD = 1.78) groups (F(1,67) = 2.10, P = 0.152, partial η^2 = 0.030) groups. This finding was not in line with the findings that gratitude attenuates negative experiences, such as pain (Algoe and Stanton, 2012; Roberts, 2004; Yu et al., 2017). A possible explanation is that the participants' gratitude faded over time.

2.2.7. Additional measurements and procedures

After the gratitude induction game, all the participants reported the frequency that they had chosen the electric-shock card (1 = very low, 9 = very high; Gratitude: M = 5.24, SD = 1.18; Control: M = 5.40, SD = 1.40; no significant difference was found between the two groups, F(1,67) = 0.28, P = 0.599, partial η^2 = 0.004). The Gratitude group also reported the frequency that the partner had decided to help them (1 = very low, 9 = very high; M = 8.50, SD = 1.50). These findings indirectly suggest that the participants understood the rules of the gratitude induction game.

After completing the adapted TPP, the participants rated how grateful, thankful, shameful, guilty, unhappy, disappointed, indebted, obligated (to repay), and angry they felt towards the partner once again. The results confirmed that the induced gratitude was valid during the adapted third-party punishment game (see Supplementary Table 5).

During the experiment, we did not explicitly tell the participants whether they would meet the partner after the experiment. This mimicked some real-life situations (e.g., you feel grateful towards someone who helps you, but you cannot be sure whether you will meet him or her again in the future). The situation was the same for the Gratitude and Control groups.

[4] In other words, the participants were told that their decisions would be revealed to the partner when all the trials of the adapted third-party punishment game were completed.

In the end, the participants were paid and were asked whether they had any question about the experiment.

2.3. Magnetic resonance imaging data acquisition

Whole-brain imaging data were acquired on a 3-Tesla magnetic resonance scanner (Siemens, Erlangen, Germany) with a 64-channel phase-array head coil. Field map images were collected using a vendor-provided Siemens gradient echo sequence (number of slices, 62; slice thickness, 2 mm; voxel size, 2 × 2 × 2 mm^3; field of view, 224 × 224 mm^2; repetition time, 620 ms; echo time 1, 4.92 ms; echo time 2, 7.38 ms; flip angle, 60°; and bandwidth, 565 Hz/pixel). During the adapted third-party punishment game, functional images were collected with a prototype simultaneous multislice echo-planar imaging (EPI) sequence (number of slices, 62; slice thickness, 2.0 mm; voxel size, 2 × 2 × 2 mm^3; field of view, 224 × 224 mm^2; repetition time, 2000 ms; echo time, 30 ms; flip angle, 90°; echo spacing, 0.54 ms; bandwidth, 2232 Hz/pixel; number of sessions, one; about 15 min). Structural images were collected through a three-dimensional magnetisation-prepared rapid acquisition gradient echo sequence (number of slices, 192; slice thickness, 1.0 mm; voxel size, 0.5 × 0.5 × 1 mm^3; field of view, 256 × 256 mm^2; repetition time, 2530 ms; echo time, 2.98 ms; flip angle, 7°; echo spacing, 7.1 ms; and bandwidth, 240 Hz/pixel). In addition, diffusion tensor imaging data were collected at the end of the scanning session. The data were used to test whether white matter tracts were associated with different personality traits, which was not related to the current study.

2.4. Behavioural analyses

2.4.1. Gratitude manipulation check

To test whether gratitude was successfully induced, we analysed the participants' answers to the questionnaire after they completed the gratitude induction game. We submitted the participants' gratitude ratings to an analysis of variance (ANOVA) with groups (Gratitude versus Control group) as a between-subjects factor. We examined whether in the Gratitude group the participants' gratitude ratings were higher than the other emotion ratings using repeated measures ANOVA. Additionally, we submitted each of seven variables indicating the participants' gratitude-related cognition and behavioural tendency (Supplementary Tables 3 and 4) to ANOVA with groups (Gratitude versus Control group) as a between-subjects factor.

To test whether the gratitude was maintained during the adapted third-party punishment game, we analysed the participants' emotion ratings after the adapted TPP. We submitted the participants' gratitude ratings to an ANOVA with groups (Gratitude versus Control group) as a between-subjects factor. We also examined whether in the Gratitude group the participants' gratitude ratings were higher than the other emotion ratings using repeated measures ANOVA (see Supplementary Table 5).

2.4.2. Punishment behaviour

We averaged the amount of tokens deducted from the partner by the participants across trials in the adapted TPP and submitted it to an ANOVA with groups (Gratitude versus Control group) as a between-subjects factor.

2.4.3. Psychological measures

We submitted the ten psychological measures hypothesised to drive the gratitude-induced difference in punishment (Supplementary Table 1) into a multivariate analysis of variance (MANOVA) with groups (Gratitude versus Control group) as a between-subjects factor. Confirming the effect of the MANOVA was significant, we submitted each of the ten psychological measures to an ANOVA with groups (Gratitude versus Control group) as a between-subjects factor. Bonferroni correction was used to adjust the *P* values of the ten statistical tests.[5]

2.4.4. Mediation analyses

We investigated the relationships amongst gratitude, punishment, and psychological measures at the individual level. We aggregated the data from the two groups and tested for potential psychological mediators of gratitude-induced differences in punishment with a two-stage process (Losin et al., 2020). In the first stage, we searched for significant correlations between gratitude ratings and each of the ten psychological measures. We calculated the percentage bend correlation coefficients, which are robust to outliers, using a MATLAB toolbox for robust correlation analyses (http://sourceforge.net/projects/robustcorrtool/; Pernet et al., 2013). The type I error inflation of the ten statistical tests was controlled by using Bonferroni correction. Psychological measures that revealed significant correlations with gratitude gratings were considered as candidate mediators (Losin et al., 2020).

In the second stage, we examined whether any of the candidate mediators mediated the observed gratitude-related difference in punishment using the PROCESS macro based on the SPSS software (http://www.processmacro.org/index.html). The continuous variables were normalised (*z*-scored). We tested the mediation effect, with the participants' gratitude ratings after the gratitude induction game as the predictor variable (*X*), average punishment at the individual level as the outcome variable (*Y*), and ratings on the candidate psychological mediators as the mediator variable (*M*) (one candidate mediator per mediation analysis). When a candidate psychological variable was entered into the model as the mediator variable, the other 9 psychological measures were involved in the model as covariates. This was performed to test the unique effect of the candidate psychological measure, given that some of the psychological measures were not completely exclusive to each other. A bootstrap procedure (randomly sampling 5000 observations with replacement) was used to obtain 95% confidence intervals (CI) of path coefficients for significance testing. An effect was regarded as significant, if the CI of the path coefficient did not cover zero (Hayes, 2013). The normalisation and bootstrap procedures and the threshold of the effect remained the same across all the mediation analyses in this study.

2.5. Computational modelling

Computational models are designed to estimate hidden variables that are not directly observable from behaviour itself (Farrell and Lewandowsky, 2010). With this method, we are allowed to probe individuals' psychological processes independent of a self-report approach and have a possibility to collect convergent evidence. The Fehr-Schmidt inequity aversion model (Fehr and Schmidt, 1999) is often adopted to reflect people's trade-off between their own payoff and the inequity between the allocator and the receiver in a typical third-party punishment game (e.g., Gao et al., 2018; Zhong et al., 2016). Inspired by it, we developed three new models (Models 1, 2, 3) to capture the participants' trade-off between the allocator's payoff (i.e., protecting the allocator from monetary harm) and the inequity between the allocator and the receiver in our adapted third-party punishment game.

Specifically, Model 1 assumes people are averse to inequity both when the allocator receives more than the receiver (i.e., advantageous inequity) and when the allocator receives less than the receiver (i.e., disadvantageous inequity). The advantageous and disadvantageous inequity are equivalent for them. The utility (U) of a punishment decision was described as follows (Model 1):

$$U(D) = \alpha * (200 - D) - (1 - \alpha) * \left| T_a - D - T_r \right|$$

[5] Although the Bonferroni correction helps to select reliable results, sometimes it may be too conservative. Considering the ten psychological variables were measured for an exploratory purpose, we believed it was needed to control the inflation of type I error using a strict correction method (i.e., Bonferroni correction) (e.g., Losin et al., 2020).

Table 1
Model comparison.

Model	Brief model description	Bayesian information criteria
1	Concerns for the allocator's benefits and absolute inequity	10,053
2	Concerns for the allocator's benefits and advantageous inequity	9995*
3	Concerns for the allocator's benefits, advantageous inequity, and disadvantageous inequity	10,255

Note: The model with the lowest Bayesian information criteria (BIC) value was selected as the preferred model. *The lowest BIC value.

where T_a and T_r represent the tokens the allocator (i.e., the partner) distributed to themselves and to the receiver, respectively. D represents the number of tokens deducted from the allocator by participants. The maximum punishment that the allocator could get is a 200-tokens deduction. In others word, participants could help the allocator to avoid a 200-tokens loss at the maximum. 200 − D is the extent participants protect the payoff of the allocator, whereas $|T_a - D - T_r|$ is the absolute inequity between the allocator and the receiver. α is the parameter to be estimated.

Model 2 assumes people are only averse to advantageous inequity:

$$U(D) = \alpha * (200 - D) - (1 - \alpha) * \max\left(T_a - D - T_r, 0\right)$$

where $\max(T_a - D - T_r, 0)$ is the advantageous inequity between the allocator and the receiver. α is the parameter to be estimated.

Model 3 assumes people are averse to both advantageous and disadvantage inequity and they are separately modelled by two different parameters:

$$U(D) = (200 - D) - \alpha * \max\left(T_a - D - T_r, 0\right) - \beta * \max\left(T_r - T_a + D, 0\right)$$

where $\max(T_a - D - T_r, 0)$ and $\max(T_r - T_a + D, 0)$ are the advantageous and disadvantageous inequity between the allocator and the receiver, respectively. α and β are the parameters to be estimated.

For all the three models, the trial-by-trial utility was transformed into choice probabilities (P) by the softmax function:

$$P(D) = \frac{e^{\lambda \cdot U(D)}}{\sum_{j \in J} e^{\lambda \cdot U(D_j)}}$$

where D_j denotes a possible number of tokens that could be deducted by the participant (from 0 to 200 tokens, in increments of 20), J is the full set of j, and λ is the inverse temperature parameter that represents the level of stochasticity in the decisions ($\lambda \in [0, 10]$) (Luo et al., 2018; Wu et al., 2020).

For parameter estimation, we constructed a coarse grid in the parameter space to select the best start location, which could decrease the possibility of the model converging on a local minimum. Maximum likelihood estimation was conducted at the individual level by maximizing the log likelihood function over each participant i and trial t (e.g., Gao et al., 2018):

$$\sum_i \sum_t \log\left(P_{i,t}(D)\right)$$

Based on the Bayesian information criteria (BIC), Model 2 with the lowest BIC value outperformed the other two alternative models (Table 1). Thus, we selected it as the preferred/winning model. In Model 2, with the increase in α, participants put more weight on the payoff of the allocator compared with the advantageous inequity between the allocator and the receiver.

We tested the preferred model (Model 2) with the methods of simulation and parameter recovery (see more details in the Supplementary Modelling Note). The model was able to replicate the pattern of the group difference in punishment with simulation (Supplementary Fig. 1), predict participants' punishment decisions with significantly higher accuracy than the chance level (34%, 95% CI: [29%, 40%]; chance level: 9%]; mean squared correlation coefficient between the real and simulated punishment at the trial level: $r^2 = 0.32$), and perform well in parameter recovery (mean squared correlation coefficient between the real and recovered parameter α: $r^2 = 0.81$) (Supplementary Fig. 2).

We examined the group difference in the parameter α using a Mann-Whitney U test, as we noticed several outliers of the parameter α (see Fig. 3a). Nonparametric tests (e.g., Mann-Whitney U test) help to reduce the effect of outliers on statistical results. For testing the robustness of our finding, we also used an ANOVA to examine the group difference in the parameter α. Given the parameter α was conceptually associated with self-reported protection tendency (i.e., reducing the benefactor's losses), we tested their correlations using percentage bend correlation analyses. We also tested whether the parameter α mediated the effect of gratitude on punishment in the same manner that the self-reported protection tendency did. We used a mediation model of the PROCESS macro to test the mediation effect, with participants' gratitude ratings after the gratitude induction game as the predictor variable (X), average punishment at the individual level as the outcome variable (Y), and parameter α as the mediator variable (M).

2.6. fMRI data analyses

2.6.1. Preprocessing

Imaging data were pre-processed using Statistical Parametric Mapping 12 (SPM 12, http://www.fil.ion.ucl.ac.uk/spm). Functional images were corrected for geometric distortions (using the field map images), corrected for slice acquisition delay, realigned for head-motion correction, co-registered to each participant's grey matter image segmented from the corresponding high-resolution T1-weighted image, normalised into Montreal Neurological Institute (MNI) space (voxel size, $2 \times 2 \times 2$ mm^3), and smoothed by an isotropic three-dimensional 6-mm full width at half maximum (FWHM) Gaussian kernel.

2.6.2. Univariate activation analyses

We used SPM 12 to conduct the univariate activation analyses. At the first level, we build a general linear model (GLM) with 4 boxcar regressors: (1) the period of viewing the partner's distribution (4 s), (2) the period of making punishment decision (response time), (3) the period of decision feedback (6.5 s - response time), and (4) the period of missing trials (if existed). We also used the unfairness of the partner's distribution (i.e., the difference between the tokens the partner distributed to themselves and the tokens to the receiver) as a parametric modulator for the period of viewing partner's distribution and used the amount of the participants' punishment as a parametric modulator for the period of making punishment decision.[6] These main regressors were convolved with a canonical hemodynamic response function. Six estimated head movement parameters defined as nuisance regressors were involved in the first-level model. The period of jitter (not modelled in the GLM) was considered as an implicit baseline. We entered four contrasts into the first-level analysis. The first contrast investigated brain activity correlated with unfairness during the period of viewing distributions (effect of the parametric modulator: unfairness). The second contrast investigated brain activity correlated with punishment during the period of

[6] We also built another two GLMs that respectively focusing on the neural correlates of the unfairness of the partner's distribution and the amount of the participants' punishment. The details were reported in the Supplementary Information. The results of the GLM in the main text and the results of the GLMs in the Supplementary Information had no statistical difference.

making decisions (effect of the parametric modulator: punishment). The third and fourth contrasts respectively investigated average brain activity during the period of viewing distributions (compared to the implicit baseline) (effect of the boxcar regressor: viewing distribution) and average brain activity during the period of making decisions (compared to the implicit baseline) (effect of the boxcar regressor: making punishment decision) (Losin et al., 2020).

We constructed second-level models as two-sample *T*-tests using contrast images from the first-level analyses to test the group differences (Gratitude versus Control group). As previous studies have identified brain regions associated with unfairness perception and punishment decision, we could take advantage of small-volume correction analyses to examine the gratitude's effect. Regions-of-interest masks (ROIs) were defined as spheres with 8-mm-radius centred at Montreal Neurological Institute (MNI) coordinates from previous studies. Specifically, to test the group difference in the brain activity correlated to unfairness, we defined 3 ROIs using the coordinates of the anterior cingulate cortex ([33, 0, 45]) and bilateral insula ([-27, 18, -3], [-30, 21, 6]) from an fMRI study on the third-party punishment game (Zhong et al., 2016), where activities in these regions were correlated with unfairness during the period of viewing distributions. To test the group difference in the brain activity correlated to punishment, we defined 2 ROIs using coordinates of the left ventral lateral prefrontal cortex ([-46, 30, -8]) and left TPJ ([-52, -52, 22]) from a recent meta-analysis (Bellucci et al., 2020), which showed that activities in the two regions were associated with punishment during the period of making decisions in the third-party punishment game. Besides small-volume correction analyses, whole-brain correction analyses were conducted for all types of contrast images from the first-level analyses.

Additionally, we constructed the second-level models as one-sample *T*-tests using contrast images from the first-level analyses and involving the participants' gratitude ratings after the gratitude induction game as covariates to test whether the individual feelings of gratitude were predictive of any brain activation.

The statistical threshold was defined as voxel-level $P < 0.001$ (uncorrected) combined with cluster-level $P < 0.05$ (FWE-corrected) (whole brain or within predefined ROIs using small-volume correction).

2.6.3. Neurosynth meta-analytical decoding

Activity of the vmPFC plays different roles in different situations (Hiser and Koenigs, 2018). To examine the extent to which the identified difference in the brain activity (i.e., vmPFC) between the groups was associated with different psychological components, we adopted the Neurosynth Image Decoder (https://www.neurosynth.org/decode/; Yarkoni et al., 2011) to decode the unthresholded *T* map of the Gratitude versus Control group contrast during the period of making decisions. This approach enabled us to quantitatively evaluate the level of similarity between the uploaded image and each of the chosen meta-analytical images from the Neurosynth database. The level of similarity was indexed by the Pearson correlation between the two vectorised images across all voxels. We chose the psychological components to be compared with the uploaded image based on a recent review on the roles of the vmPFC (Hiser and Koenigs, 2018), including "reward", "value", "self-referential", "ToM (theory of mind)", "emotional responses", "emotional faces", and "emotion regulation".

2.6.4. Mediation analyses (fMRI data)

The univariate activation analyses showed significant group differences in the beta estimates of the TPJ and vmPFC during the period of making decisions. We were interested in whether any of these neural indexes mediated the relationship between gratitude and punishment. We used a mediation model of the PROCESS macro to test the mediation effect. The mean estimates in the TPJ and the vmPFC were extracted from each participant's first-level contrast image and used as the mediator variable (*M*) (one mediator per mediation analysis). Participants' gratitude ratings after the gratitude induction game and average punishment at the individual level were respectively defined as the predictor variable (*X*) and the outcome variable (*Y*).

2.6.5. Percentage bend correlation analyses (fMRI data)

We did not find a mediation effect of the parameter estimates of the vmPFC in the relationship between gratitude and punishment. As supplementary analyses, we examined the direct correlations between the parameter estimates of the vmPFC and gratitude and punishment. We calculated the percentage bend correlation coefficients for them.

After confirming the involvement of the activity of the TPJ and vmPFC in the gratitude-induced injustice, we used percentage bend correlation analyses to examine whether these neural findings were correlated with the protection tendency and parameter α, which were key psychological processes associated with the gratitude-induced injustice based on the results of the self-report and computational modelling.

3. Results

3.1. Participants in the gratitude group exhibited higher levels of gratitude

Participants in the Gratitude group had more gratitude towards the partner than participants in the Control group after the gratitude induction game ($F(1,67) = 81.60$, $P < 0.001$, partial $\eta^2 = 0.549$; Fig. 2a), suggesting that our manipulation of gratitude was successful (see Supplementary Tables 2, 3, 4 and 5 for more supporting evidence).

3.2. Participants in the gratitude group decreased punishment for the partner

In the adapted TPP, participants in the Gratitude group deducted the partner fewer tokens than those in the Control group did ($F(1,67) = 22.67$, $P < 0.001$, partial $\eta^2 = 0.253$; Fig. 2b). This suggests that grateful participants were willing to violate the norm of justice to protect their benefactors from monetary harm.

3.3. Participants in the gratitude group reported a stronger tendency to protect the partner

The MANOVA showed a significant group difference in psychological measures hypothesised to explain the gratitude-induced difference in punishment (Pillai's trace: 0.33, $F(10,58) = 2.87$, $P = 0.006$, partial $\eta^2 = 0.331$). According to separate ANOVAs, amongst the 10 psychological measures (Supplementary Table 1), only one differed between the Gratitude and Control groups after Bonferroni correction. Compared to participants in the Control group, participants in the Gratitude group reported a significantly stronger tendency to reduce the partner's losses (protection tendency: $F(1,67) = 23.00$, $P_{\text{Bonferroni corrected}} < 0.001$, partial $\eta^2 = 0.256$; Fig. 2c), which implies grateful participants valued the benefits of their benefactors. No significant difference was found in other psychological measures (Supplementary Table 6).

3.4. Protection tendency mediated the effect of gratitude on participants' punishment for the partner

The relationships amongst gratitude, punishment, and psychological measures were investigated using a two-stage process (see the Method section) (Losin et al., 2020). In the first stage, we found that gratitude ratings were significantly correlated with self-reported protection tendency ($r = 0.52$, $P_{\text{Bonferroni corrected}} < 0.001$, $N = 69$) but not with any other psychological measure (Supplementary Table 7). Hence, self-reported protection tendency was regarded as a candidate mediator that was likely to explain the decreased punishment amongst grateful participants.

Self-reported protection tendency significantly mediated the relationship between gratitude and punishment (indirect effect (path ab):

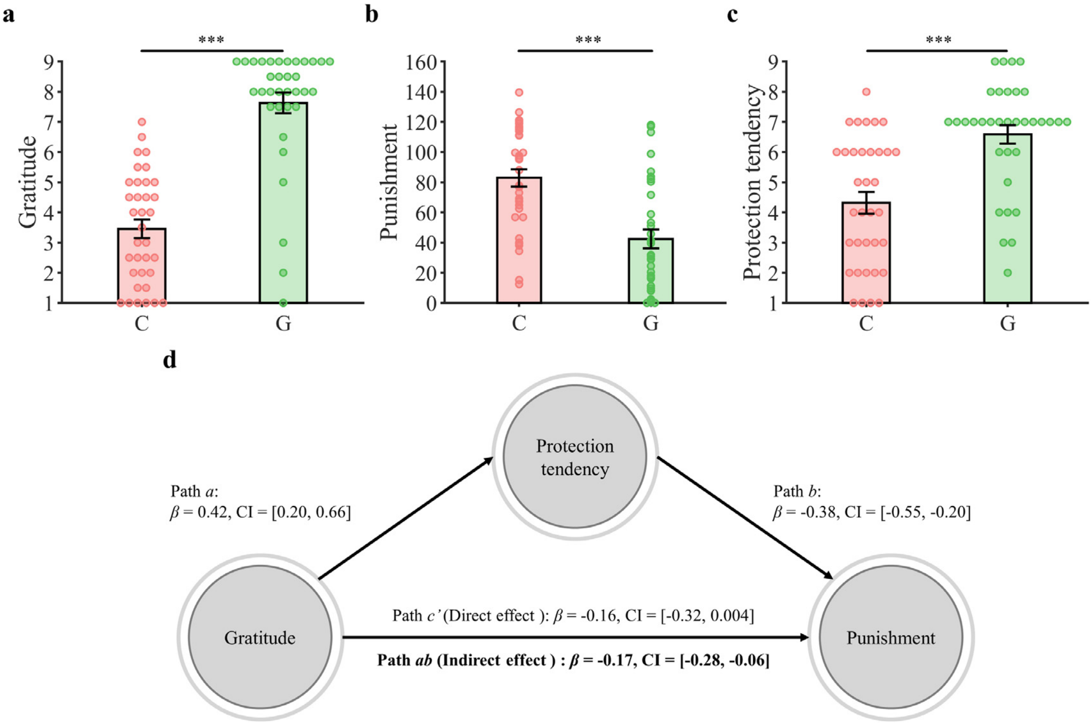

Fig. 2. Behavioural analysis results. **a–c**, Participants in the Gratitude group had stronger gratitude feelings (***P < 0.001) (**a**), performed less punishment (***P < 0.001) (**b**), and showed a stronger protection tendency (***$P_{\text{corrected}}$ < 0.001) (**c**) than participants in the Control group. C, Control group; G, Gratitude group. Data are shown as the mean ± standard error with overlaid dot plots. **d**, Self-reported protection tendency mediated the effect of feelings of gratitude on punishment. β, path coefficient; CI, 95% confidence interval of β; bold font, significant indirect effect.

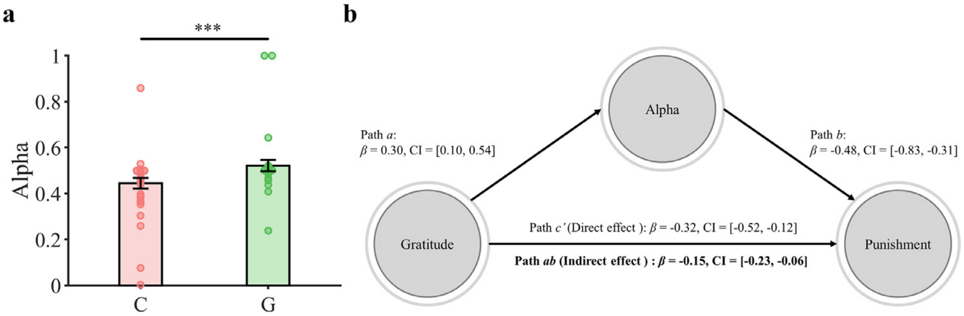

Fig. 3. Computational modelling results. **a**, Participants in the Gratitude group had higher estimates of parameter α than the Control group (***P < 0.001; Mann-Whitney U test). C, Control group; G, Gratitude group. Data are shown as the mean ± standard error with overlaid dot plots. **b,** Parameter α mediated the effect of feelings of gratitude on punishment. β, path coefficient; CI, 95% confidence interval of β; bold font, significant indirect effect.

β = -0.17, 95% confidence interval (CI) = [-0.28, -0.06]; Fig. 2d), even when the effects of the other 9 psychological measures were controlled. Additionally, for completeness, we also tested whether any of the other 9 psychological measures mediated the effect of gratitude on punishment. None of them had a significant effect (Supplementary Table 8). Thus, the findings demonstrate a crucial role of protection tendency in the gratitude-induced change in punishment. The results of self-report suggest that gratitude-induced injustice was mainly driven by the increased concerns for benefactors' benefits rather a distortion of unfairness perception or a change in mood.

3.5. Participants in the gratitude group had larger estimates of α

We found a significant group difference in the estimates of α (Mann-Whitney U test: Z = 3.48, P < 0.001, r = 0.425; ANOVA: F(1,65) = 5.16, P = 0.026, partial η^2 = 0.074) (Fig. 3a). The parameter α was positively correlated with the self-reported protection tendency (r = 0.53, P < 0.001, N = 67). Consistent with the results of self-report, the findings here suggest that more weight was placed on the benefits of benefactors (relative to the inequity aversion) in the Gratitude group than the Control group.

3.6. Parameter α mediated the effect of gratitude on punishment in a manner similar to self-reported protection tendency

As the parameter α was associated with the protection tendency, we tested whether the parameter α could also play a role as a mediator. Similar to the protection tendency, the parameter α significantly mediated the relationship between gratitude and punishment (indirect effect (path ab): β = -0.15, CI = [-0.23, -0.06]; Fig. 3b). Together, the self-report and computational modelling approaches demonstrate the key role of the increased concerns for benefactors' benefits in the gratitude-induced injustice.

3.7. Similar neural response to unfairness and similar neural activity during the period of viewing distributions were engaged in the gratitude and control groups

During the period of viewing distributions, no brain regions' response to unfairness differed significantly between the groups and no brain regions differed significantly in activation between the groups based on the predetermined threshold. No significant result was found even when we adopted a more lenient threshold (voxel-level P < 0.005

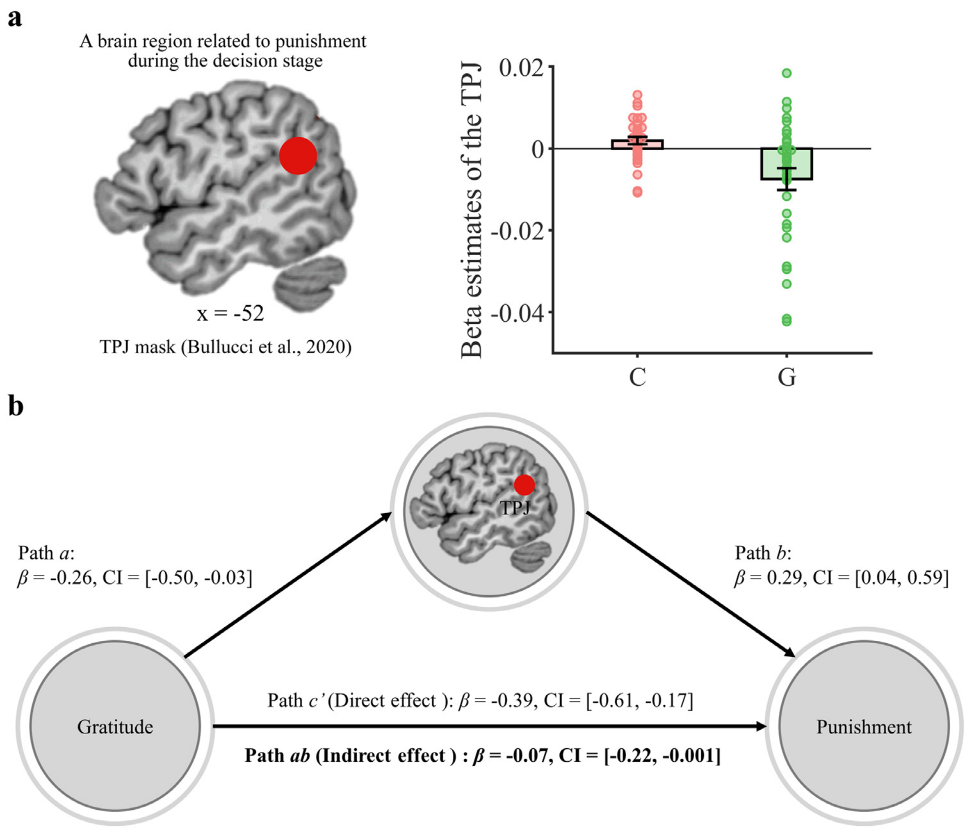

Fig. 4. fMRI results of TPJ. **a**, During the period of making decisions, TPJ activity in the Gratitude group relative to the Control group negatively correlated with the punishment (P_{FWE} = 0.021, small volume corrected). Mean parameter estimates of the TPJ were extracted from the cluster that survived the correction. The estimates are shown as the mean ± standard error with overlaid dot plots. G, Gratitude group; C, Control group. **b**, Parameter estimates of the TPJ mediated the effect of feelings of gratitude on punishment. β, path coefficient; CI, 95% confidence interval of β; bold font, significant indirect effect.

(uncorrected) combined with cluster-level $P < 0.05$ (FWE-corrected); whole brain or within predefined ROIs using small-volume correction). Thus, no neural evidence supports the hypothesis that gratitude distorts individuals' perception of unfairness.

3.8. Distinct relationships between the TPJ activity and punishment were found between the gratitude and control groups

A small-volume correction analysis revealed that the left TPJ activity (peak MNI coordinates: [-54, -48, 22]; cluster size: 6 voxels; P_{FWE} = 0.021, small volume corrected; Fig. 4a) in the Gratitude group relative to the Control group negatively correlated with punishment during the period of making decisions. As the predefined ROI (i.e., left TPJ) is implicated in punishment decisions (Bellucci et al., 2020), the results indicates that gratitude modulated the neural activity related to punishment. In a whole-brain analysis, we found no clusters exceeding the predetermined statistical threshold.

3.9. The vmPFC showed greater activation during the period of making decisions in the gratitude group

A whole-brain analysis revealed that the vmPFC (peak MNI coordinates: [-12, 42, -8]; cluster size: 156 voxels; P_{FWE} = 0.021, whole brain corrected) exhibited greater activation in the Gratitude than Control group during the period of making decisions (Fig. 5a). Additionally, after pooling the two groups, participants' feelings of gratitude after the gratitude induction game were predictive of the activation of the vmPFC (peak MNI coordinates: [-8, 50, -12]; cluster size: 160 voxels; P_{FWE} = 0.019, whole brain corrected) during the period of making decisions (Supplementary Fig. 3). The identified vmPFC cluster largely overlapped with the vmPFC cluster that showed a group difference in activation (Fig. 5b).

Given that previous studies have found multifaceted roles of the vmPFC (Hiser and Koenigs, 2018), we examined the extent to which the group differences in activation during the period of making decisions were associated with different psychological components, using the Neurosynth Image Decoder (Yarkoni et al., 2011). This analysis is conducive to inferring the psychological components associated with a contrast based on meta-analytical images instead of a possibly biased selection of literature (Yarkoni et al., 2011). The resulting *T* map of the contrast was more strongly associated with the meta-analytical patterns of "reward" (r = 0.14) and "value" (r = 0.12) than those of "self-referential" (r = 0.06), "ToM (theory of mind)" (r = 0.02), "emotional responses" (r = 0.03), "emotional faces" (r = 0.04), and "emotion regulation" (r = 0.02) (Fig. 5c). The results imply that gratitude modulated the neural activity associated with reward processing and value representation during the period of making decisions.

3.10. The activities of the TPJ and vmPFC were involved in the gratitude-induced injustice and related to the key psychological processes associated with the gratitude-induced injustice

We investigated whether the parameter estimates of the TPJ and vmPFC mediated the relationship between gratitude and punishment. The results revealed that the parameter estimates of the TPJ (i.e., correlation between the TPJ activity and punishment within each participant during the period of making decisions) significantly mediated the effect of feelings of gratitude on individuals' overall punishment (indirect effect (path ab): β = -0.07, CI = [-0.22, -0.001]; Fig. 4b). The parameter estimates of the vmPFC (i.e., the vmPFC activation during the period of making decisions) did not have a significant mediation effect (indirect effect (path ab): β = -0.09, CI = [-0.23, 0.03]). As a supplement, we examined direct correlations between the parameter estimates of the vmPFC and gratitude and punishment. The parameter estimates of vmPFC were positively correlated with gratitude (r = 0.57, P < 0.001, N = 67; Fig. 6d) and negatively correlated with punishment (r = -0.40, P < 0.001, N = 67; Fig. 6e). The findings indicate that the activities of the TPJ and vmPFC were involved in the gratitude-induced injustice.

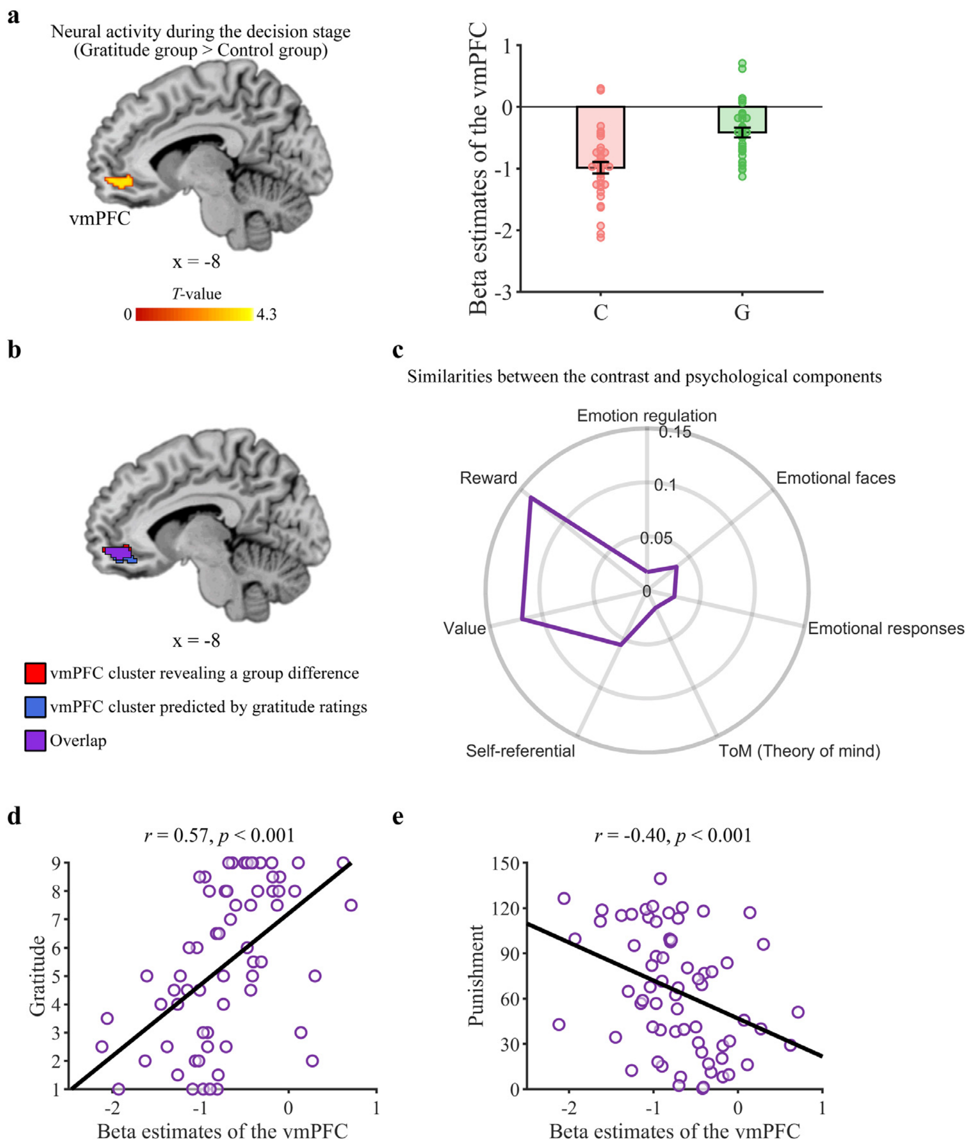

Fig. 5. fMRI results of vmPFC. **a**, During the period of making decisions, activation of vmPFC was greater in the Gratitude than Control group (P_{FWE} = 0.021, whole brain corrected). Mean parameter estimates of the vmPFC were extracted from the cluster survived the correction. The estimates are shown as the mean ± standard error with overlaid dot plots. G, Gratitude group; C, Control group. **b**, The vmPFC cluster in which activation was predicted by gratitude ratings overlapped with the vmPFC cluster that revealed a significant group difference in activation during the period of making decisions. **c**, The similarities (*r* values) between the *T* map of the Gratitude versus Control group contrast during the period of making decisions and the meta-analytical maps of terms of psychological components derived from the Neurosynth are shown. **d**, Parameter estimates of the vmPFC were positively correlated with gratitude. **e**, Parameter estimates of the vmPFC were negatively correlated with punishment.

The correlations between the activities of the TPJ and vmPFC and the key psychological variables associated with the gratitude-induced injustice were calculated. Although the parameter estimates of the TPJ were not significantly correlated with the protection tendency (r = -0.23, P = 0.058, N = 67) or parameter α (r = -0.22, P = 0.080, N = 67), trends appeared. The parameter estimates of the vmPFC were significantly correlated with the protection tendency (r = 0.47, P < 0.001, N = 67) and parameter α (r = 0.36, P = 0.003, N = 67).

4. Discussion

In spite of a number of studies focusing on gratitude's positive impact on society (e.g., Bartlett and DeSteno, 2006; DeSteno et al., 2019; Ma et al., 2017; Yu et al., 2017, 2018), a potential dark side of gratitude is recently noticed (Ng et al., 2017; Zhu et al., 2020). Confirming this dark side, the current study revealed that the Gratitude group exerted lower punishment on their benefactors' unfairness to strangers than the Control group. Importantly, here we offered accounts of how gratitude leads individuals to benefit benefactors at the cost of violating the justice norm. The results of self-report showed that the boosted protection tendency on behalf of benefactors is an important underlying psychological process of the gratitude-induced injustice. No result of the self-report advocated the hypotheses that distortions of unfairness perception or changes in mood play roles. As participants in both the Gratitude and Control groups benefited from their benefactors' costly helping, the difference in punishment between the groups could not be explained by the need for equity restoration.

Independent of the self-report approach, computational model analysis revealed that parameter α (i.e., concerns for benefactors' benefits relative to inequity aversion) was implicated in the gratitude-induced injustice. Thus, both the results from computational modelling and self-report (especially the results of mediation analyses; see Figs. 2d and 3b) suggest that gratitude promotes individuals to concern for benefactors' benefits, and in turn decreases punishment for benefactors' unfairness to strangers. Our findings provide supportive evidence for the find-bind-remind theory that gratitude promotes behaviour that is conductive to relationship-building (Algoe, 2012; Algoe et al., 2008) and extend the theory by identifying the psychological and computational mechanisms underlying such behaviour.

The implementation of the social function of gratitude involves three steps (Algoe et al., 2008; Yu et al., 2018). Compared to the neural findings of the first two steps—cognitive antecedent evaluation and gratitude generation (Decety and Porges, 2011; Fox et al., 2015; Xiong et al., 2020; Yu et al., 2017, 2018; Zahn et al., 2009)—the neural mechanisms directly driving gratitude-induced behaviour are far less known. The current study helps to close this gap in knowledge by revealing the group difference in brain activation during punishment decisions (can be considered as the neural correlates of the dark side of gratitude). Our fMRI results showed that gratitude modulated the correlation between the TPJ activity and punishment during the period of making decisions. Numerous studies on punishment have found the TPJ to be involved in a mentalizing process advocating punishment decisions (e.g., Buckholtz et al., 2008; Ginther et al., 2016; Hu et al., 2016; Krueger et al., 2014). In those studies, the TPJ is supposed to understand the suffering of the victim, including inferring the wrongdoer's spiteful/selfish intent and determining blameworthiness of the wrongdoer (Bellucci et al., 2020). In our study, we found that the TPJ activity in the Gratitude group (relative to the Control group) was related to decisions to refrain from punishment (i.e., a negative correlation between the TPJ activity and punishment; see Fig. 4a). As the location of the TPJ cluster we found is within a TPJ region which has been demonstrated to be associated with making punishment (see a meta-analysis, Bellucci et al., 2020), consistent with a typical explanation of the TPJ's role in punishment, we believe that the activity of the TPJ cluster is associated with mentalizing. Nevertheless, gratitude may change the way of mentalizing. Previous studies showed that social emotion promotes individuals to concentrate their attention on targets that mostly related to the social emotion (e.g., guilt draws individuals' attention to victims' benefits for repairing the damaged relationship with the victims and results in the neglect of the need of strangers) (de Hooge et al., 2011). Analogously, in our study the Gratitude group might withdraw their attention from understanding the pain of the victim and shift it to concerning for the benefactor's (wrongdoer's) benefits for building relationship with the benefactor. In this case, the TPJ is likely to function for speculating the benefactor's preference instead of caring about the victim.

Our fMRI results revealed that gratitude increased the activation in the vmPFC during the period of making decisions. At the group level, the Gratitude group compared with the Control group had stronger activation in the vmPFC. At the individual level, participants with stronger feelings of gratitude had greater activation in the vmPFC, the location of which overlapped with the vmPFC cluster of the group contrast. Many studies have found that the vmPFC is involved in gratitude generation (when individuals receive help from benefactors) (Fox et al., 2015; Xiong et al., 2020; Yu et al., 2017). The vmPFC activity is supposed to be associated with processing the reward of social interaction or experiencing a positive emotion (Fox et al., 2015; Yu et al., 2017). Different from these studies, we identified the activation of vmPFC during the stage of individuals making decisions regarding benefactors rather than receiving help from benefactors. The results from Neurosynth Image Decoder indicated that our findings were more likely to be associated with reward processing and value computation than other psychological components (e.g., emotional response/experience) (see Fig. 5c). In our opinion, the vmPFC activity we identified may reflect that grateful individuals are more inclined to consider the right of punishment as a valuable chance to benefit benefactors for building relationship and conceive it as a social reward.

Besides, we found that the parameter estimates of the TPJ mediated the effect of gratitude on punishment and that the parameter estimates of the vmPFC were correlated with both gratitude and punishment. We considered these findings as (preliminary) evidence that the TPJ and vmPFC activities are associated with the gratitude-induced injustice. Moreover, the activities of the vmPFC had correlations with the key psychological processes (i.e., the protection tendency and parameter α) related to the gratitude-induced injustice. These findings confirmed the important roles of the TPJ and vmPFC activities in the gratitude-induced injustice.

No brain activity associated with unfairness differed significantly and no brain activation differed significantly during the period of viewing distributions between the groups. It is in line with the results of self-report that no significant difference was found in the unfairness judgement or immoral judgement between the groups (Supplementary Table 6). Thus, the hypothesis that the gratitude-induced injustice is due to a distortion of unfairness perception is not supported. We also found that no regions' activity differed significantly during both the period of viewing distributions and making decisions between the groups. If positive mood is a cause of the gratitude-induced injustice, we may find higher activation in regions associated with positive mood in both the period of viewing distributions and making decisions, given that mood is a long-lasting background state (Beedie et al., 2005). However, it is not the case. In the same line, we found no significant difference in gladness, anger, or arousal between the groups based on the participants' self-report (Supplementary Table 6). Therefore, no evidence supports that positive mood plays a role in the gratitude-induced injustice.

Our study are related to two previous research (Vayness et al., 2020; Yu et al., 2018). However, they have fundamental differences. Vayness et al. (2020) found that incident gratitude promotes participants to punish more severely on strangers' unfairness at the cost of the participants' own benefits, which implies a prosocial feature of gratitude like many other gratitude studies (see a review, Ma et al., 2017). Our findings show that integral gratitude[7] inhibits participants from punishing their benefactors even when the benefactors violate the norm of fairness, which warms people a dark side of gratitude and emphasizes the social function of gratitude to be personal relationship-building (Algoe, 2012; Zhu et al., 2020). Yu et al. (2018) showed that participants' neural response of the anterior cingulate cortex (ACC) to benefactors' help was predictive of participants' consequent gratitude-related compensation for the benefactors. It is noted that Yu et al. (2018) focused on the neural activities of generating gratitude (when participants received help from the benefactors). Differently, we investigated the neural activities of taking actions on the benefactors. Gratitude generation and gratitude-related action towards the benefactors are not only separated by time, but also may be corresponding to distinct psychological and neural systems. Indeed, we identified different brain regions (i.e., the TPJ and vmPFC rather than ACC) that were implicated in the gratitude-related behaviour during the action stage. To the best of our knowledge, our study for the first time examines the neural activities that directly drive grateful individuals to act for relationship enhancement.

The limitations of our study present opportunities to research in the future. Our insignificant results cannot support the argument that the distortion of unfairness perception and the positive mood are not involved in the gratitude-induced injustice at all. Thus, based on our findings, we cautiously propose that the increased concerns for benefactors' benefits are of importance in the gratitude-induced injustice. Future studies are needed to examine our findings with a larger sample size. On the other hand, the psychological mechanisms underlying high-level social behaviours, such as gratitude-induced injustice, are complicated. It is plausible that our study has not exhausted all the possibilities. We call for future studies to further our explorations, such as investigating whether grateful individuals conduct less punishment on their benefactors in the adapted TPP for reciprocating the benefactors' help in the gratitude induction game (i.e., a one-time repayment gesture; Algoe et al., 2008) and whether grateful individuals themselves regard conducting less punishment as unrighteous. Last but not the least, we focus on the neural activities associated with the gratitude-induced injustice. Studying the neuroanatomical correlates of this phenomenon is an interesting future direction (e.g., Liu et al., 2018).

[7] The influence of emotion is integral or incident depends on whether it is directly related to the current situation (e.g., Zhu et al., 2019). An example of integral gratitude is that you feel grateful due to receiving benevolent help from Person A and later the gratitude modulates your behaviour on Person A; An example of incident gratitude is that you feel grateful because of receiving benevolent help from Person A and the gratitude towards A modulates your behaviour on Person B (a stranger unrelated to the induction of your gratitude).

Our study makes several novel contributions. Firstly, we develop a new task (i.e., the gratitude induction game) which can reliably induce gratitude and ensure its maintenance for a relatively long time (see Supplementary Tables 2–5) and create an adapted version of the TPP which can disentangle the benefit of the allocator from the cost to the decider. Secondly, we delineate the process of a newly identified phenomenon, gratitude-induced injustice, at behavioural and computational levels. Understanding the key motivation of grateful individuals being benefitting benefactors helps to dispel a potential positive stereotype about gratitude (i.e., "gratitude is the parent of other virtues"), which echoes a recent trend of comprehending emotions based on their adaptive functions (Sznycer, 2019; Sznycer and Lukaszewski, 2019). Thirdly, we provide empirical evidence for the neural mechanism directly underlying the gratitude-induced behaviour. The brain imaging results fill in the missing piece of the neural substrates of the find-remind-bind theory, which suggests the involvement of the regions associated with mentalizing (i.e., TPJ) and reward processing (i.e., vmPFC) in social decision-making across grateful individuals. Finally, we also deepen the understanding of injustice by revealing the possibility that individuals are willing to violate moral norms for their benefactors.

In conclusion, we show that gratitude leads individuals to decrease due punishment for their benefactors, during which the boosted protection tendency on behalf of benefactors is a key motivation. Furthermore, we demonstrate that the brain activities associated with mentalizing and reward processing support the gratitude-induced injustice. Our findings provide mechanistic and neural accounts of gratitude-induced behaviour and shed light on the nature of gratitude.

Declaration of Competing Interest

The authors declare no conflicts of interest.

Credit authorship contribution statement

Ruida Zhu: Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Zhenhua Xu:** Investigation, Visualization, Writing – review & editing. **Song Su:** Writing – review & editing, Supervision. **Chunliang Feng:** Visualization, Writing – review & editing. **Yi Luo:** Visualization, Writing – review & editing. **Honghong Tang:** Visualization, Writing – review & editing. **Shen Zhang:** Methodology. **Xiaoyan Wu:** Visualization, Writing – review & editing. **Xiaoqin Mai:** Writing – review & editing, Supervision. **Chao Liu:** Conceptualization, Methodology, Visualization, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition.

Acknowledgments

We thanked Jiahua Xu for providing useful advice. This work was supported by the National Natural Science Foundation of China (NSFC) (32130045, 31871094), the Major Project of National Social Science Foundation (19ZDA363), the Beijing Municipal Science and Technology Commission (Z151100003915122), the National Program for Support of Top-notch Young Professionals, and the International Postdoctoral Exchange Fellowship Program (No. 2020002).

Data and code availability statements

Data could be made available by contacting the corresponding author (Chao Liu, liuchao@bnu.edu.cn). Data availability is subject to the approval from the institutional review board with a data sharing agreement. Behavioural analyses were completed with the Statistical Product and Service Solutions (SPSS) software, unless stated otherwise. The code used for percentage bend correlation analyses was retrieved from the Robust Correlation Toolbox (http://sourceforge.net/projects/robustcorrtool/). The code used for computational modelling was retrieved from the Optimisation Toolbox of the Matrix Laboratory (MATLAB) (https://fr.mathworks.com/help/optim/ug/fmincon.html). The code used for fMRI analyses was retrieved from the Statistical Parametric Mapping (SPM) platform (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118730.

References

Algoe, S.B., 2012. Find, remind, and bind: the functions of gratitude in everyday relationships. Soc. Personal. Psychol. Compass 6, 455–469.
Algoe, S.B., Haidt, J., Gable, S.L., 2008. Beyond reciprocity: gratitude and relationships in everyday life. Emotion 8, 425–429.
Algoe, S.B., Stanton, A.L., 2012. Gratitude when it is needed most: social functions of gratitude in women with metastatic breast cancer. Emotion 12, 163–168.
Bartlett, L., DeSteno, D., 2006. Gratitude and prosocial behavior. Psychol. Sci. 17, 319–325.
Bartlett, M.Y., Condon, P., Cruz, J., Baumann, J., Desteno, D., 2012. Gratitude: prompting behaviours that build relationships. Cogn. Emot. 26, 2–13.
Bartra, O., McGuire, J.T., Kable, J.W., 2013. The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. Neuroimage 76, 412–427.
Beedie, C.J., Terry, P.C., Lane, A.M., 2005. Distinctions between emotion and mood. Cogn. Emot. 19, 847–878.
Bellucci, G., Camilleri, J.A., Iyengar, V., Eickhoff, S.B., Krueger, F., 2020. The emerging neuroscience of social punishment: meta-analytic evidence. Neurosci. Biobehav. Rev. 113, 426–439.
Berkowitz, L., 1993. Pain and aggression: some findings and implications. Motiv. Emot. 17, 277–293.
Bonini, N., Hadjichristidis, C., Mazzocco, K., Demattè, M.L., Zampini, M., Sbarbati, A., Magon, S., 2011. Pecunia Olet: the role of incidental disgust in the ultimatum game. Emotion 11, 965–969.
Buckholtz, J.W., Asplund, C.L., Dux, P.E., Zald, D.H., Gore, J.C., Jones, O.D., Marois, R., 2008. The neural correlates of third-party punishment. Neuron 60, 930–940.
Cicero, M.T., 1851. The Orations of Marcus Tullius Cicero (Vol. III). George Bell & Sons, London, UK (C. D.Younge, Trans.).
Crockett, M.J., Kurth-Nelson, Z., Siegel, J.Z., Dayan, P., Dolan, R.J., 2014. Harm to others outweighs harm to self in moral decision making. Proc. Natl. Acad. Sci. U.S.A. 111, 17320–17325.
de Hooge, I.E., Nelissen, R.M.A., Breugelmans, S.M., Zeelenberg, M., 2011. What is moral about guilt? Acting "prosocially" at the disadvantage of others. J. Personal. Soc. Psychol. 100, 462–473.
Decety, J., Porges, E.C., 2011. Imagining being the agent of actions that carry different moral consequences: an fMRI study. Neuropsychologia 49, 2994–3001.
DeSteno, D., Duong, F., Lim, D., Kates, S., 2019. The grateful do not cheat: gratitude as a fount of virtue. Psychol. Sci. 30, 979–988.
Farrell, S., Lewandowsky, S., 2010. Computational models as aids to better reasoning in psychology. Curr. Dir. Psychol. Sci. 19, 329–335.
Faul, F., ErdFelder, E., Lang, A.G., Buchner, A, 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39, 175–191.
Fehr, E., Fischbacher, U., 2004. Third-party punishment and social norms. Evol. Hum. Behav. 25, 63–87.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Q. J. Econ. 114, 817–868.
Fox, G.R., Kaplan, J., Damasio, H., Damasio, A., 2015. Neural correlates of gratitude. Front. Psychol. 6, 1–11.
Gao, X., Yu, H., Sáez, I., Blue, P.R., Zhu, L., Hsu, M., Zhou, X., 2018. Distinguishing neural correlates of context-dependent advantageous-and disadvantageous-inequity aversion. Proc. Natl. Acad. Sci. 115, E7680–E7689.
Ginther, M.R., Bonnie, R.J., Hoffman, M.B., Shen, F.X., Simons, K.W., Jones, O.D., Marois, R., 2016. Parsing the behavioral and brain mechanisms of third-party punishment. J. Neurosci. 36, 9420–9434.
Gummerum, M., Van Dillen, L.F., Van Dijk, E., López-Pérez, B., 2016. 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.
Hayes, A., 2013. Introduction to mediation, moderation, and conditional process analysis. J. Educ. Meas. 51, 335–337.
Hiser, J., Koenigs, M., 2018. The multifaceted role of the ventromedial prefrontal cortex in emotion, decision making, social cognition, and psychopathology. Biol. Psychiatry 83, 638–647.
Horberg, E.J., Oveis, C., Keltner, D., 2011. Emotions as moral amplifiers: an appraisal tendency approach to the influences of distinct emotions upon moral judgment. Emot. Rev. 3, 237–244.
Hu, Y., Scheele, D., Becker, B., Voos, G., David, B., Hurlemann, R., Weber, B., 2016. The effect of oxytocin on third-party altruistic decisions in unfair situations: an fMRI study. Sci. Rep. 6, 1–10.

Krueger, F., Hoffman, M., Walter, H., Grafman, J., 2014. An fMRI investigation of the effects of belief in free will on third-party punishment. Soc. Cogn. Affect. Neurosci. 9, 1143–1149.

Liu, G., Cui, Z., Yu, H., Rotshtein, P., Zhao, F., Wang, H., Sui, J., 2020. Neural responses to intention and benefit appraisal are critical in distinguishing gratitude and joy. Sci. Rep. 10, 1–12.

Liu, G., Zeng, G., Wang, F., Rotshtein, P., Peng, K., Sui, J., 2018. Praising others differently: neuroanatomical correlates to individual differences in trait gratitude and elevation. Soc. Cogn. Affect. Neurosci. 13, 1225–1234.

Losin, E.A.R., Woo, C.W., Medina, N.A., Andrews-Hanna, J.R., Eisenbarth, H., Wager, T.D., 2020. Neural and sociocultural mediators of ethnic differences in pain. Nat. Hum. Behav. 1–14.

Luo, Y., Hétu, S., Lohrenz, T., Hula, A., Dayan, P., Ramey, S.L., Ramey, C., 2018. Early childhood investment impacts social decision-making four decades later. Nat. Commun. 9, 1–10.

Ma, L.K., Tunney, R.J., Ferguson, E., 2017. Does gratitude enhance prosociality?: a meta–analytic review. Psychol. Bull. 143, 601–635.

McCullough, M., Kilpatrick, S., 2001. Is gratitude a moral affect? Psychol. Bull. 127, 249–266.

McCullough, R.A.E.M.E., 2004. The Psychology of Gratitude. Oxford University Press.

Nelissen, R.M.A., Zeelenberg, M., 2009. Moral emotions as determinants of third-party punishment: anger, guilt, and the functions of altruistic sanctions. Judgm. Decis. Mak. 4, 543–553.

Ng, J.W.X., Tong, E.M.W., Sim, D.L.Y., Teo, S.W.Y., Loy, X., Giesbrecht, T., 2017. Gratitude facilitates private conformity: a test of the social alignment hypothesis. Emotion 17, 379–387.

Peng, C., Nelissen, R.M.A., Zeelenberg, M., 2018. Reconsidering the roles of gratitude and indebtedness in social exchange. Cogn. Emot. 32, 760–772.

Pernet, C.R., Wilcox, R., Rousselet, G.A., 2013. Robust correlation analyses: false positive and power validation using a new open source matlab toolbox. Front. Psychol. 3, 1–18.

Roberts, R.C., 2004. The blessings of gratitude: a conceptual analysis. In: The Psychology of Gratitude, pp. 58–78.

Ruff, C.C., Ugazio, G., Fehr, E., 2013. Changing social norm compliance with noninvasive brain stimulation. Science 342, 482–484.

Sheldon, K.M., Lyubomirsky, S., 2006. How to increase and sustain positive emotion: the effects of expressing gratitude and visualizing best possible selves. J. Posit. Psychol. 1, 73–82.

Stallen, M., Rossi, F., Heijne, A., Smidts, A., De Dreu, C.K.W., Sanfey, A.G., 2018. Neurobiological mechanisms of responding to injustice. J. Neurosci. 38, 2944–2954.

Suardi, A., Sotgiu, I., Costa, T., Cauda, F., Rusconi, M., 2016. The neural correlates of happiness: a review of PET and fMRI studies using autobiographical recall methods. Cogn. Affect. Behav. Neurosci. 16, 383–392.

Sznycer, D., 2019. Forms and functions of the self-conscious emotions. Trends Cogn. Sci. 23, 143–157.

Sznycer, D., Lukaszewski, A.W., 2019. The emotion–valuation constellation: multiple emotions are governed by a common grammar of social valuation. Evol. Hum. Behav. 40, 395–404.

Tangney, J.P., Dearing, R.L., 2003. Shame and Guilt. Guilford Press, New York.

Tesser, A., Gatewood, R., Driver, M., 1968. Some determinants of gratitude. J. Personal. Soc. Psychol. 9, 233–236.

Trost, Z., Vangronsveld, K., Linton, S.J., Quartana, P.J., Sullivan, M.J.L., 2012. Cognitive dimensions of anger in chronic pain. Pain 153, 515–517.

Ugazio, G., Lamm, C., Singer, T., 2012. The role of emotions for moral judgments depends on the type of emotion and moral scenario. Emotion 12, 579–590.

Vayness, J., Duong, F., DeSteno, D., 2020. Gratitude increases third-party punishment. Cogn. Emot. 34, 1020–1027.

Vytal, K., Hamann, S., 2010. Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. J. Cogn. Neurosci. 22, 2864–2885.

Watkins, P.C., Woodward, K., Stone, T., Kolts, R.L., 2003. Gratitude and happiness: development of a measure of gratitude, and relationships with subjective well-being. Soc. Behav. Personal. 31, 431–452.

Wood, A.M., Emmons, R.A., Algoe, S.B., Froh, J.J., Lambert, N.M., Watkins, P., 2016. A dark side of gratitude? distinguishing between beneficial gratitude and its harmful impostors for the positive clinical psychology of gratitude and well-being. In: The Wiley Handbook of Positive Clinical Psychology, pp. 137–151.

Wu, X., Feng, C., Zhang, S., Liang, Z., Dong, W., Qin, S., Liu, C. (2020). The selfish nature in interpersonal exchange among adolescents: imposing the kind and submitting to the exploitative. BioRxiv, 2020.10.26.354704.

Xiong, W., Gao, X., He, Z., Yu, H., Liu, H., Zhou, X., 2020. Affective evaluation of others' altruistic decisions under risk and ambiguity. Neuroimage 218, 116996.

Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., Wager, T.D., 2011. Large-scale automated synthesis of human functional neuroimaging data. Nat. Methods 8, 665–670.

Yu, H., Cai, Q., Shen, B., Gao, X., Zhou, X., 2017. Neural substrates and social consequences of interpersonal gratitude: intention matters. Emotion 17, 589–601.

Yu, H., Gao, X., Zhou, Y., Zhou, X., 2018. Decomposing gratitude: representation and integration of cognitive antecedents of gratitude in the brain. J. Neurosci. 38, 4886–4898.

Yu, H., Hu, J., Hu, L., Zhou, X., 2014. The voice of conscience: neural bases of interpersonal guilt and compensation. Soc. Cogn. Affect. Neurosci. 9, 1150–1158.

Zahn, R., Moll, J., Paiva, M., Garrido, G., Krueger, F., Huey, E.D., Grafman, J., 2009. The neural basis of human social values: evidence from functional MRI. Cereb. Cortex 19, 276–283.

Zhong, S., Chark, R., Hsu, M., Chew, S.H., 2016. Computational substrates of social norm enforcement by unaffected third parties. Neuroimage 129, 95–104.

Zhu, R., Xu, Z., Tang, H., Liu, J., Wang, H., An, Y., Liu, C., 2019. The effect of shame on anger at others: awareness of the emotion-causing events matters. Cogn. Emot. 33, 696–708.

Zhu, R., Xu, Z., Tang, H., Wang, H., Zhang, S., Zhang, Z., Liu, C., 2020. The dark side of gratitude: gratitude could lead to moral violation. J. Exp. Soc. Psychol. 91, 104048.