



The mutuality of social emotions: How the victim's reactive attitude influences the transgressor's emotional responses



Xiaoxue Gao^{a,b,1,*}, Hongbo Yu^{c,1,*}, Lu Peng^b, Xiaoliang Gong^d, Yang Xiang^d, Changjun Jiang^d, Xiaolin Zhou^{a,b,e,f,g,*}

^a Shanghai Key Laboratory of Mental Health and Psychological Crisis Intervention, School of Psychology and Cognitive Science, East China Normal University, Shanghai 200062, China

^b School of Psychological and Cognitive Sciences, Peking University, Beijing 100871, China

^c Department of Psychological and Brain Sciences, University of California Santa Barbara, Santa Barbara, CA, 93106-9660, USA

^d Key Laboratory of Embedded System and Service Computing (Ministry of Education), Tongji University, Shanghai 201804, China

^e School of Business and Management, Shanghai International Studies University, Shanghai 200083, China

^f Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing 100871, China

^g PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, China

ARTICLE INFO

Keywords:

Guilt
Anger
Expectation violation
Ventral striatum
fMRI

ABSTRACT

Would a transgressor be guiltier or less after receiving the victim's forgiving or blaming attitude? Everyday intuitions and empirical evidence are mixed in this regard, leaving how interpersonal attitudes shape the transgressor's reactive social emotions an open question. We combined a social interactive game with multivariate pattern analysis of fMRI data to address this question. Participants played an interactive game in an fMRI scanner where their incorrect responses could cause either high or low pain stimulation to an anonymous co-player. Following incorrect responses, participants were presented with the co-player's (i.e., the victim's) attitude towards the harm (Blame, Forgive, or Neutral). Behaviorally, the victim's attitude and the severity of harm interactively modulated the transgressor's social emotions, with expectation violation serving as a mediator. While unexpected forgiveness following severe harm amplified the participants' guilt, unexpected blame following minor harm reduced the participants' guilt and increased their anger. This role of expectation violation was supported by multivariate pattern analysis of fMRI, revealing a shared neural representation in ventral striatum in the processing of victim's attitude-induced guilt and anger. Moreover, we identified a neural re-appraisal process of guilt in the transgressor, with the involvement of area related to self-conscious processing (i.e., perigenual anterior cingulate cortex) before knowing the victim's attitude transiting to the involvement of other-regarding related area (i.e., temporoparietal junction) after knowing the victim's attitude. These findings uncover the neurocognitive bases underlying the transgressor's social emotional responses, and highlight the importance of the mutuality of social emotions.

1. Introduction

Imagine that you felt guilty for accidentally damaging your friend's bicycle that was the present from her beloved grandmother (De Hooge et al., 2011). If you were expecting that your friend would be angry and blame you, but in fact she forgave you, would you feel more guilty or less? These everyday anecdotes demonstrate the mutuality of social emotions – the nature and strength of social emotions in dynamic interactions are shaped by how one party expects and reacts to the attitudes of the other party (Helm, 2017; Strawson, 1974). This mu-

tuality is fundamental to human adaptations to volatile social environments (Gassin and Elizabeth, 1998; Gross, 2002; Louie and De Martino, 2013). However, empirical studies on the transgressor's social emotions have mainly focused on how self-conscious appraisals (e.g., responsibility for the other's harm) give rise to guilt (Baumeister et al., 1994; Tangney et al., 2007; Vaish and Hepach, 2019). The neurocognitive bases underpinning how transgressors process interpersonal feedbacks (e.g., the victim's attitudes of forgiveness and blame) and adjust their emotional and behavioral responses remain unclear.

Social psychological studies examining the influences of the victim's attitude on the transgressor's reactions are inconclusive (Exline et al.,

* Corresponding authors.

E-mail addresses: gxx114455@gmail.com (X. Gao), hongbo.yu@psych.ucsb.edu (H. Yu), xz104@pku.edu.cn (X. Zhou).

¹ These authors contributed equally to this work.

2003; Gassin and Elizabeth, 1998). Several studies suggested that forgiveness reduces guilt (McNulty, 2010, 2011) and blame increases it (Kubany and Watson, 2003; Parkinson and Illingworth, 2009), whereas others observed the opposite effects, namely forgiveness enhances guilt (Wallace et al., 2008), and blame reduces guilt and even induces anger in the transgressor (Jennings et al., 2016; Lemay Jr et al., 2012; Zechmeister and Romero, 2002). One potential explanation for these inconsistencies is that these studies overlooked the expectation violation derived from the interaction between the victim's attitude and the severity of harm. Specifically, individuals can form expectations about others' attitudes and behaviors according to social norms and experiences (Ci, 2006; Olsson et al., 2018; Olsson et al., 2020). Others' actual attitudes or behaviors may deviate from these expectations, forming expectation violations (also referred to as prediction errors in the literature of reward learning and decision-making) that could be crucial sources of social emotions (Chang and Jolly, 2017; Chang and Smith, 2015; Miceli and Castelfranchi, 2014). For example, theories on equity and justice have suggested that while unexpected over-benefit contributes to guilt, unexpected under-benefit or over-punishment contributes to anger (Adams, 1965; Baumeister et al., 1994; Blair, 2012; Donnerstein and Hatfield, 1982; Homans, 1974; Walster et al., 1978). Extending these results to the transgression context, transgressors commonly expect to be blamed for high harm (Young and Saxe, 2009) and to be forgiven for low harm (Malle, 2021). Therefore, we hypothesize that when the victim's attitude is more favorable than what the transgressor expects (e.g., forgiving a high harm), this positive expectation violation (over-benefit) may exacerbate guilt. In contrast, when the victim's attitude is more hostile than expected (e.g., blaming a low harm), this negative expectation violation (over-punishment) may induce anger and reduce guilt.

Neurally, studies applying functional magnetic resonance imaging (fMRI) have investigated the neurocognitive bases of guilt in the context of interpersonal transgression without social feedbacks (i.e., *non-reactive guilt*). Results of univariate analysis (Basile et al., 2011; Chang et al., 2011; Koban et al., 2013; Wagner et al., 2011; Yu et al., 2014) and multivariate pattern analysis (MVPA) (Yu et al., 2020a; Yu et al., 2020b) of fMRI data consistently showed that the process of non-reactive guilt involves activities in anterior/middle cingulate cortex (dACC/aMCC) and bilateral anterior insula (aINS), regions that are implicated in distress and anxiety processing. However, after receiving the victim's reactive attitude, the transgressor may exhibit a reappraisal process that contributes to the reactive experience of guilt in response to the victim's attitude (i.e., *reactive guilt*). Although non-reactive and reactive guilt are indistinguishable in self-report, question remains as to whether the neural bases of these two types of guilt are similar or dissociable. Analogously, previous studies investigated the neural bases of the victim's anger in response to transgressions, revealing the involvements of ACC and aINS (Blair, 2012; Chang and Smith, 2015; Denson et al., 2009; Klimecki et al., 2018), as well as amygdala, a region implicated in negative emotion processing (Denson et al., 2009; Klimecki et al., 2018). Yet, the neural bases underlying the transgressor's reactive anger in response to the victim's attitude remain unclear.

In the current study, we aim to fill these gaps by addressing 1) how do the victim's attitude and the extent of harm interact to influence the transgressor's social emotional responses and what is the role of expectation violation during this process? 2) what are the neural bases underlying the transgressor's reactive guilt and anger in response to the victim's attitude? To this end, we combined fMRI with a multi-round interactive game with social feedbacks, developed on the basis of previous studies on guilt (Gao et al., 2018; Koban et al., 2013; Yu et al., 2014) (Fig. 1). In each round, the participant was paired with an anonymous co-player (confederate) and played a dot estimation task. The co-player (the victim) would receive either high or low pain stimulation with 50% probability when the participant responded incorrectly (*Harm to the co-player: High vs. Low*). After seeing the extent of harm to the co-player (i.e., Outcome phase), the participant was presented with the co-player's

attitude towards the harm caused by the participant (*Attitude of the co-player: Blame vs. Forgive vs. Neutral; Attitude phase*). Before fMRI scanning, the participant made predictions of the co-players' attitudes when they received a high or a low pain stimulation. During scanning, unbeknown to the co-player, the participant made monetary allocations between him/herself and the co-player paired in the current trial; this allocation could be taken as an index for guilt-induced compensation or anger-induced aggressive behaviors. After scanning, the participant rated his/her feelings of guilt, anger, gratitude, sadness and embarrassment in response to the co-player's attitude in each condition (the participant's reactive social emotions). MVPA was applied to explore the neural commonalities and differences between the participant's reactive guilt and anger in response to social feedbacks from the victim, as well as between the participant's non-reactive guilt before receiving the victim's attitude and reactive guilt after receiving the victim's attitude.

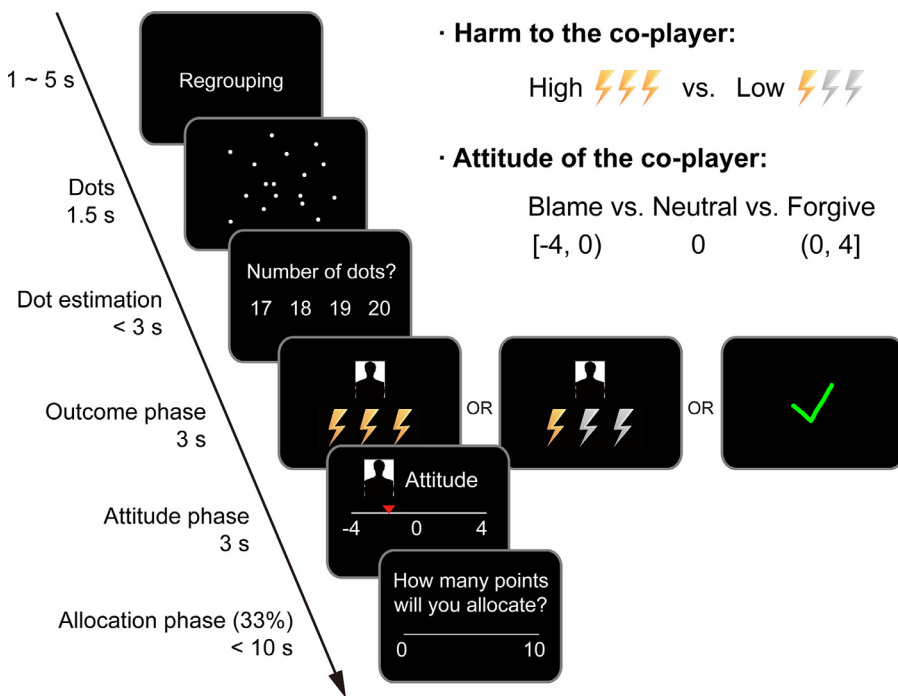
2. Materials and methods

2.1. Participants

A total of 32 graduate and undergraduate students from universities in Shanghai, China were recruited for the fMRI experiment. Five participants were excluded due to excessive head motion (> 3 mm of translation or 3 degrees of rotation) in the fMRI scanner, leaving 27 participants (13 female, 21.70 ± 1.71 (SD) years) for further analysis. Additional 30 graduate and undergraduate students from universities in Beijing, China were recruited for a behavioral replication (20 females, 22.23 ± 2.43 years). All participants were right-handed with normal or corrected-to-normal vision and with no self-reported history of neurological and psychological problems. The experiment was carried out in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the School of Psychological and Cognitive Sciences, Peking University. Informed written consent was obtained from each participant before the experiment.

2.2. Pain titration

Upon arrival, each participant met three co-players (i.e., confederates). All the three confederates were university students of the same sex as the participant to avoid the possibility that the perceived age and sex of the co-players influence the participant's responses. The three male co-players were the same for all the male participants, and the three female co-players were the same for all the female participants. The participant was told that he/she was assigned to the role of player A, and the three co-players were assigned to the role of player B according to their enrollment orders. They would later play an interactive game together through intranet in separate rooms. No other information about the co-players was communicated to the participant. Then the three co-players were led to another testing room. Pain titration was conducted following the procedure of previous studies (Gao et al., 2018; Xiong et al., 2020; Yu et al., 2017; Yu et al., 2018; Yu et al., 2014). An intra-epidermal needle electrode was attached to the back of the left hand of the participant for cutaneous electrical stimulation (Inui et al., 2002). The first pain stimulation was set as 8 repeated pulses, each of which was 0.2 mA and lasted 0.5 ms with a 10 ms interval in between. We gradually increased the intensity of each single pulse until the participant reported 8 on a 10-level pain scale (1 = not painful, 10 = intolerable, linearly increased). Participants reported that they could only experience the whole train of pulses as a single stimulation rather than as separate shocks. They were told that the two levels of pain stimulation that each player B would receive in the interactive game would be the ones that each player B rated as "4" and "8" (i.e., low and high pain stimulation) in pain intensity. All participants reported that the two levels of pain could be clearly distinguished.



periment, five rounds paired with each co-player (15 rounds in total) would be randomly selected and realized to determine the participant's and each co-player's monetary bonus and the final amount of pain stimulation each co-player would receive. Note, before and after Outcome phase and Attitude phase, a fixation cross was presented for a variable interval ranging from 1 to 6 s for the purpose of fMRI signal deconvolution.

2.3. The interactive game

The interactive game was developed based on previous studies on non-reactive guilt (Gao et al., 2018; Koban et al., 2013; Yu et al., 2014). After the pain titration, the participant was instructed on the general rules of the interactive dot-estimation task. In each round of the task (Fig. 1), the participant was paired with one of the three co-players and performed a dot-estimation task. The participant was explicitly informed that the co-player in each round was selected randomly from the three co-players by a computer program; the co-player in the current round could or could not be the same co-player in the previous trial. To avoid the possibility that the participant learned from the co-player's attitudes, the participant was instructed that the interactive task was anonymous and he/she would not know the identity of the co-player in each round throughout the task. If the estimation was correct, the current trial terminated and the next round began. Otherwise, the co-player in the current round would receive a high or low intensity pain stimulation, randomly determined by the computer program (i.e., **Harm to the co-player: High harm vs. Low harm**), with the level of the pain stimulation for the co-player being presented on the screen (i.e., Outcome phase, 3 s). The participant was informed that both the co-player and him/herself would see the outcome of dot-estimation (correct vs. incorrect) and the intensity of harm to the co-player; but the co-player would not see the picture of the dots and response options. The co-player would then indicate his/her attitude toward the harm in estimated-incorrect trials on a scale from -4 to 4 (i.e., **Attitude of the co-player**). Three types of attitudes could be expressed through this scale: (1) positive value for forgiveness, with a higher positive value indicating the co-player's higher willingness to forgive the co-player (**Forgive condition**), (2) negative value for blame, with a lower negative value indicating the co-player's higher willingness to blame the co-player (**Blame condition**), and (3) zero indicated the co-player's neutral attitude (neither forgiveness nor blame) toward the participant's performance (**Neutral condition**). The participant would be presented with the co-player's attitude on the scale from -4 to 4 (i.e., Attitude phase, 3 s). After the presentation of the co-player's attitude, in one-third of the trials, the participant was asked to divide 10 points

(1 point = 2 Yuan; 20 Yuan \approx 3.1 USD) between him/herself and the co-player paired in this trial (i.e., Allocation phase, < 10 s); in the remaining trials, the current trial terminated after the presentation of the co-player's attitude. The participant was informed that his/her paired co-player was unaware of this allocation procedure, eliminating the possibility that the co-player's attitude expression was out of monetary concerns. The money allocated to the co-player was treated as an index for guilt-induced compensation or anger-induced aggressive behaviors. After the experiment, five rounds paired with each co-player (15 rounds in total) would be randomly selected and realized to determine the participant's and each co-player's monetary bonus and the final amount of pain stimulation each co-player would receive. Since the participant was paired with only one co-player at each round, which might lead to the concern of unequal workload as the other two unpaired co-players would not be in a task state in the current round, we told the participant that the unpaired co-players would carry out a task that were irrelevant to the interactive game during the un-pairing time.

Unbeknown to the participant, the correctness of dot estimation and each co-player's attitude in each trial were predetermined by a computer program. The estimation-correct trials were treated as fillers in data analyses. There were 6 possible conditions of Harm to the co-player and Attitude of the co-player for the estimation-incorrect trials, forming a 2 (Harm to the co-player: High vs. Low harm) \times 3 (Attitude of the co-player: Blame vs. Forgive vs. Neutral) within-participant design. To ensure that the three types of the co-player's attitude could be distinguished clearly, in the current experiment, trials with attitude ratings of "3" and "4" were regarded as Forgive trials, trials with attitude ratings of "-3" and "-4" were regarded as Blame trials, and trials with attitude ratings of "0" were regarded as Neutral trials. Trials with attitude ratings of "-2," "-1," "1," and "2" were regarded as filler trials. Therefore, the average values of the co-player's actual attitude feedbacks in Blame, Neutral and Forgive conditions were -3.5, 0, and 3.5 respectively.

The experiment consisted of 114 trials (34 estimation-correct trials and 80 estimation-incorrect trials). Among the estimation-incorrect trials, there were 12 trials for Blame (ratings of "-3" and "-4"), Forgiveness (ratings of "3" and "4") and Neutral (rating of "0") attitudes under High

harm and Low harm conditions respectively. The other 8 trials with attitude ratings of "-2," "-1," "1," and "2" were filler trials. See Table S1 in *Supplementary Materials* for the distribution of estimation-incorrect trials in different Harm-Attitude conditions. The task was divided into 3 runs with equal number of trials for each condition in each run. Each run consisted of 38 trials in total and lasted for about 14.5 minutes. Trials within a run were pseudo-randomly mixed to ensure that no more than two consecutive trials were from the same condition. During the scanning, before and after Outcome phase and Attitude phase, a fixation cross was presented for a variable interval ranging from 1 to 6 s for the purpose of fMRI signal deconvolution.

2.4. Subjective ratings for the interactive task

Before the interactive task, the participant was asked to predict the co-player's attitudes on the scale from -4 ('higher willingness to blame') to 4 ('higher willingness to forgive') under High harm and Low harm conditions respectively (i.e., the participant's predicted attitude of the co-player). After the interactive task and before the payment for participation, the participant recalled and rated his/her emotional responses to the co-player after the co-player's attitude was revealed (i.e., Attitude phase) on a scale from 1 ('not at all') to 7 ('very strong') under each of the six conditions respectively, including guilt, anger, gratitude, sadness, and embarrassment (i.e., the participant's reactive social emotions). This way of post-experiment ratings has been proven effective in previous studies on social emotions, such as guilt (Chang et al., 2011; Gao et al., 2018; Li et al., 2020; Yu et al., 2014; Zhu et al., 2018) and gratitude (Liu et al., 2020; Yu et al., 2017; Yu et al., 2018; Zhu et al., 2018). We did not ask participants to report their emotions during Outcome phase because knowing that they would need to rate their emotions during Attitude phase could influence how they evaluate their emotions for Outcome phase. No participant doubted the believability of the experimental setup when he/she was asked to comment on the procedures after the experiment.

2.5. Behavioral replication

To validate the robustness of our behavioral results of the fMRI experiment, additional 30 graduate and undergraduate students from universities in Beijing, China were recruited for another behavioral experiment (20 females, 22.23 ± 2.43 years). The procedure of this behavioral experiment was the same as the fMRI experiment, except that there was no time jittering before or after Outcome phase or Attitude phase.

2.6. Behavioral analyses

First, we fed participants' ratings of guilt and anger and the amounts of monetary allocation into 2 (Harm to the co-player: High vs. Low harm) \times 3 (Attitude of the co-player: Blame vs. Forgive vs. Neutral) repeated-measures ANOVAs to examine whether Harm to the co-player and Attitude of the co-player had modulated participants' guilt, anger and subsequent behaviors.

Second, to test the hypothesis regarding the relationships between expectation violation and ratings of guilt and anger, we calculated expectation violation about the co-player's attitude in each condition according to the participant's predicted attitude of the co-player before the task. Specifically, for the High harm situation, expectation violations in Blame, Forgive and Neutral conditions were computed respectively as the difference between the average value of the co-player's actual attitude feedback in each corresponding condition during the task and the participant's predicted attitude of the co-player in High harm situation before the task. Given that the co-players' attitude feedbacks were predetermined, with 12 trials for each attitude, namely, Blame (ratings of "-3" and "-4"), Forgiveness (ratings of "3" and "4"), and Neutral (rating of "0"), in High harm and Low harm conditions respectively, the average values of the co-player's actual attitude feedbacks were -3.5 for Blame, 0 for Neutral, and 3.5 for Forgive. Similarly, for the Low harm

situation, expectation violations in Blame, Forgive and Neutral conditions were computed respectively as the difference between the average value of the co-player's actual attitude feedback in each corresponding condition during the task (i.e., -3.5 for Blame, 0 for Neutral, and 3.5 for Forgive) minus the participant's predicted attitude of the co-player in Low harm situation before the task. The values of expectation violation in the six conditions were then fed into three separate linear mixed models (LMMs) as the predictor (fixed effect) for ratings of guilt, ratings of anger and amounts of allocation respectively. By-participant random slopes for each fixed effect were included in each LMM. LMM estimations were conducted using "lme4" package in R (Bates et al., 2014). All the variables were centered and normalized in each LMM to obtain standardized coefficients.

Third, to investigate whether guilt and anger served as the two main motivations underlying monetary allocation, LMMs were conducted with the amount of allocation as the dependent variable. By-participant random slope for each fixed effect was included in each LMM. Seven models were included and compared (Table S3 in *Supplementary Materials*). Model 1 included both guilt and anger ratings as fixed effects. To test the necessity of guilt and anger in allocation, Model 2 with guilt rating as the single predictor and Model 3 with anger rating as the single predictor were included. To test whether adding ratings of other emotions could explain more variance in the allocation, in Models 4, 5 and 6, in addition to guilt and anger ratings, ratings of gratitude, sadness and embarrassment were included as additional fixed effects in each of the three models respectively. To exclude the possibility that the amount of allocation could be better explained by other emotions rather than guilt and anger, Model 7 included only ratings of gratitude, sadness and embarrassment as fixed effects. Model goodness of fit was assessed using the Bayesian information criterion (BIC; Lewandowsky and Farrell, 2010), which takes into account both model fitness and complexity. Parameters were estimated based on the best model (lowest BIC).

To test whether guilt and anger mediated the effect of expectation violation on monetary allocation, we conducted a multiple mediation model analyses using structural equation modeling through the 'lavaan' package in R software (Rosseel, 2012). In this model, ratings of guilt and anger were included as two mediators simultaneously. Mediating effects of guilt and anger were compared in this model. All the variables of the multivariate mediation model were centered and normalized within participant before the analyses. Standardized coefficient for each path of the model was labeled on corresponding figures. For each path of the model, the values of c indicated the effect of the corresponding independent variable on dependent variable before controlling for the effect of mediator; the values of c' indicated the effect of the corresponding independent variable on dependent variable after controlling for the effect of mediator. A significant c' indicated a partial mediation while a non-significant c' indicated a complete mediation (Preacher and Hayes, 2008).

In the interactive task, we deliberately attempted to minimize the participant's learning on the victim's attitudes or characters by making the co-players anonymous throughout the task. Using LMM, we further tested whether the participant's response pattern changed over trials, with the amount of monetary allocation as the dependent variable, and Harm to the co-player, Attitude of the co-player, trial ID and their interactions as predictors. By-participant random slope for each fixed effect was included in this LMM. Results showed that neither the main effect of trial ID nor its interactions with other experimental factors significantly contributed to the participant's amounts of monetary allocation (Table S8).

2.7. Neuroimaging data acquisition and preprocessing

Images were acquired on a 3.0 T MR scanner (GE MR750) with an eight-channel head coil at Tongji University, Shanghai. T2-weighted functional images were acquired in 40 axial slices parallel to the an-

terior commissural–posterior commissural line with no inter-slice gap, affording full-brain coverage. Images were acquired using an EPI pulse sequence (TR = 2000 ms; TE = 30 ms; flip angle = 90°; FOV = 192 mm × 192 mm; slice thickness = 3 mm; voxel size x = 3 mm, voxel size y = 3 mm). An ascending, interleaved slice acquisition order was used starting from the odd slices. A high-resolution, whole-brain structural scan (1 mm³ isotropic voxel MPRAGE) was acquired after functional imaging. Imaging processing was conducted following the standard pre-processing procedures in the Statistical Parametric Mapping software SPM12 (Wellcome Trust Department of Cognitive Neurology, London, UK), including 1) discarding the first 5 volumes of the functional images to allow for stabilization of magnetization; 2) correcting for within-scan acquisition time difference between slices, with the middle (i.e., the 39th) slice as the reference, i.e., slice-time correction; 3) realigning the remaining volumes to the sixth volume to correct for head-motion, and generate the six rigid-body motion parameters; 4) spatially normalizing functional images to the Montreal Neurological Institute (MNI) space using the EPInorm approach (Calhoun et al., 2017) in which functional images are aligned to an EPI template, nonlinearly warped to stereotactic space, and resampled to 3 mm × 3 mm × 3 mm isotropic voxels; 5) spatially smoothing functional images with an 8 mm FWHM Gaussian filter; and 6) temporally filtering using a high-pass filter with a cutoff frequency of 1/128 Hz. Prior work has shown that spatial smoothing does not decrease the sensitivity of MVPA (Hendriks et al., 2017; Op de Beek, 2010).

2.8. Univariate general linear model analyses

Univariate general linear model (GLM) analyses were conducted at individual level (i.e., first-level analysis) in SPM12. In the GLM, we built a design matrix with separable run-specific partitions. For each run, we modeled twelve separate regressors in estimation-incorrect trials corresponding to the six key conditions in Outcome phase and Attitude phase respectively, spanning from the presentation of the corresponding screen to the end of this event (3 s):

High_harm_Blame_Outcome, High_harm_Forgive_Outcome,
High_harm_Neutral_Outcome, Low_harm_Blame_Outcome,
Low_harm_Forgive_Outcome, Low_harm_Neutral_Outcome,
High_harm_Blame_Attitude, High_harm_Forgive_Attitude,
High_harm_Neutral_Attitude, Low_harm_Blame_Attitude,
Low_harm_Forgive_Attitude, Low_harm_Neutral_Attitude
(R1 to R12).

Regressors of no interest included: Attitude fillers (onsets of Attitude phases in which the ratings were "-2," "-1," "1," and "2", R13, 3 s), Correct outcome (onsets of Outcome phase in which the estimation was correct, R14, 3 s), Dot estimation (start from the presentation of dots to the end of dot estimation phase, R15, 4.5 s) and Allocation phase (the phase for allocation, R16, response time as duration). Six rigid-body motion parameters were also included as regressors of no interest (R17-R22) to reduce the impact of head motion on the patterns of functional activation in the current event-related design (Johnstone et al., 2006; Wilke, 2012). See descriptive statistics for head motion parameters in Table S4. Three baseline regressors modeling the average activity in each run were included at the end of the design matrix. All regressors were convolved with a canonical hemodynamics response function (HRF). The statistical maps estimation was conducted using restricted maximum likelihood (ReML), where temporal autocorrelation was estimated globally given the residuals from an initial OLS model estimation. An autoregressive AR(1) model was used during ReML parameter estimation to account for serial correlations (Friston et al., 2002; Penny et al., 2003). The ReML procedure then pre-whitened both the data and the design matrix, and estimated the model. The contrast images corresponding to the main effects of the twelve regressors of interest (R1 - R12) were extracted and used for training and test in the multivariate pattern analysis.

2.9. Multivariate pattern analysis (MVPA)

Multivariate pattern analysis was carried out in Python 3.6.8 using the NLTools package version 0.3.14 (<https://nltools.org/>). For each binary classification, we used the contrast images for all participants in the corresponding phase and linear Support Vector Machine (SVM) (Friedman et al., 2001; Wager et al., 2013) to train a whole-brain between-participant multivariate pattern classifier discriminating the corresponding binary conditions (e.g., "High harm" vs. "Low harm" in Outcome phase). SVM was conducted using linear kernel (regularization parameter C = 1), which has been suggested as a reasonable setting in multivariate pattern analysis (Varoquaux et al., 2017). With a 5-folds cross-validation method, we calculated the accuracy and significance of the SVM classifier using the forced-choice discrimination test (Chang et al., 2015; Wager et al., 2013; Woo et al., 2014). Receiver-Operating-Characteristic (ROC) curve was created based on the performance of the classification. Similar classification procedure was applied to all the following classification analyses. To be noted, first, during the cross-validation, all images from the same participant were held out together, which ensured that the training data and test data were independent from each other. Second, in the forced-choice discrimination test, pattern expression values (i.e., the dot-product of a vectorized activation image with the classifier weights) were compared for two conditions/groups tested within the same out-of-sample individual; the higher was chosen as "high harm" and the lower as "low harm" (Chang et al., 2015; Wager et al., 2013; Woo et al., 2014). Forced-choice tests are particularly suitable for the current study because they do not compare the signature responses with a threshold that is fixed across participants (Wager et al., 2013). Thus, noises or confounding factors, such as the individual differences in the emotion rating scale for one specific emotion, the differences in the emotion rating scale across emotions, or the differences in shape and amplitude of fMRI responses across studies, do not influence the performance of this kind of test (Wager et al., 2013). This method enabled us to conduct cross-emotion and cross-study classifications and predictions in the following analyses, which were similar to those used in previous studies (Chang et al., 2015; Wager et al., 2013; Woo et al., 2014; Yu et al., 2020a).

2.9.1. Whole-brain pattern classifications for reactive guilt and anger in Attitude phase

To investigate the neural underpinnings of participants' reactive guilt after receiving the victim's attitude feedback (i.e., Attitude phase), for each participant, we categorized the six conditions in Attitude phase into "High guilt" and "Low guilt" groups based on the post-experiment guilt rating of this participant. Specifically, for each participant, the contrast maps of the three conditions with the highest ratings of guilt were labeled as "High guilt", and the others were labeled as "Low guilt". Combining the contrast maps of all the participants, we applied linear support vector machine (SVM) (Friedman et al., 2001; Wager et al., 2013) to train a whole-brain between-participant multivariate pattern classifier discriminating these two groups of maps. The same set of analyses was conducted to train a whole-brain multivariate pattern classifier for reactive anger during Attitude phase by categorizing the contrast images in Attitude phase into "High anger" and "Low anger" groups based on post-experiment anger ratings. To be noted, in the current study, the categorizations of guilt groups were not correlated with those of anger groups ($\beta = 0.27 \pm 0.45$, $z = 0.61$, $p = 0.545$), which ensured that the multivariate pattern classification procedures for reactive guilt and anger were independent from each other.

To test whether the whole-brain multivariate patterns for reactive guilt generated from binary classification in Attitude phase could predict condition-wise post-experiment guilt ratings, the contrast images for each of the six conditions for each participant were used to obtain pattern expression values based on the reactive guilt pattern. To obtain single pattern expression value of each condition and each participant, we computed the dot product of the cross-validated weight map of the

reactive guilt pattern and the individual contrast images. This value reflects the distance between a given activation map and the classifier represented by a hyperplane in the feature space. The pattern expression values of all conditions and all the participants were then fed into a repeated-measures ANOVA to test whether there existed a significant 2 (Harm to the co-player: High vs. Low harm) \times 3 (Attitude of the co-player: Blame vs. Forgive vs. Neutral) interaction effect as observed in guilt rating. Additionally, the pattern expression values for the six conditions in Attitude phase were regressed against guilt ratings to assess their ability in predicting the feelings of guilt by using LMM. By-participant random slope for each fixed effect was included in this LMM and all the LMMs on pattern expression values in the following analyses. The same set of analyses were conducted to test whether the whole-brain multivariate pattern for reactive anger generated from binary classification could predict condition-wise post-experiment anger ratings.

To test whether the processing of reactive guilt and anger in Attitude phase exhibited shared or differential neural bases at whole-brain pattern level, first, we conducted cross-emotion classification by calculating the forced-choice classification accuracy for how well High and Low guilt groups were correctly classified based on pattern expression values obtained from the reactive anger pattern, and how well High and Low anger groups were correctly classified based on pattern expression values obtained from the reactive guilt pattern. Second, we tested whether the pattern expression values for the six conditions in Attitude phase generated based on reactive guilt pattern were correlated with those generated based on reactive anger pattern using LMM.

2.9.2. Whole-brain multivariate classifications: comparing non-reactive guilt processing across studies

Previous studies have suggested that a transgressor's guilt is positively correlated with the extent of harm (Berndsen and McGarty, 2010; Renetzky, 2015). This was supported by our behavioral results that participants felt guiltier in High harm condition than in Low harm condition. Therefore, we predicted that neural differences between High and Low harm conditions in Outcome phase, where the co-player's attitude was not yet available, should reflect the neural processing of non-reactive guilt and should be similar to the neural signature of non-reactive guilt established in previous studies (e.g., Yu et al. 2020a). Yu et al. (2020a) used the data of one previous univariate fMRI study on non-reactive guilt (Yu et al., 2014), in which the participant played a dot estimation task together with an anonymous co-player on each trial. If anyone responded incorrectly, the co-player would receive pain stimulation. Therefore, there were three conditions in Yu et al. (2014) where the co-player would have to receive pain stimulation, namely "Co-player_Responsibile", "Both_Responsibile", and "Self_Responsibile". Participants' self-reported guilt ratings were obtained for these three conditions. Using this dataset, Yu et al. (2020a) trained a the non-reactive guilt pattern (GRBS) discriminating High (Self_Responsibile) vs. Low (Both_Responsibile) guilt conditions by applying linear SVM (Friedman et al., 2001; Wager et al., 2013).

To identify neurocognitive processing related to non-reactive guilt in Outcome phase, we used linear SVM (Friedman et al., 2001; Wager et al., 2013) to develop a whole-brain classifier to discriminate High harm vs. Low harm conditions in Outcome phase in the current study. To validate that the classifier was indeed related to non-reactive guilt, we applied it to discriminate High (Self_Responsibile) vs. Low (Both_Responsibile) guilt conditions in a previous study on non-reactive guilt (Yu et al., 2014; Yu et al., 2020a). In a complementary manner, the non-reactive guilt pattern discriminating High (Self_Responsibile) vs. Low (Both_Responsibile) guilt conditions obtained from Yu et al. (2020a) was used to discriminate High vs. Low harm conditions in Outcome phase in the current study (i.e., cross-study classification). The rationale is that, if the whole-brain classifier we developed based on the current dataset (Outcome phase) indeed captured the neurocognitive processing of non-reactive guilt, then the accuracy of cross-study classification should be higher than chance in both directions.

We also tested this hypothesis using values of pattern expression. We computed the pattern expression values for the three conditions (Self_Responsibile, Both_Responsibile, Co-player_Responsibile) in Yu et al. (2020a) based on the pattern classifier for harm in Outcome phase, and tested whether these pattern expression values could predict the guilt ratings in the corresponding conditions in Yu et al. (2020a) using LMM. Because knowing that they would need to rate their emotions during Attitude phase could influence how participants evaluate their emotions for Outcome phase, we did not collect participants' self-reported guilt ratings regarding Outcome phase. Therefore, we could not conduct analyses concerning whether the pattern expression values in Outcome phase generated from the guilt pattern in Yu et al. (2020a) could predict the guilt ratings in Outcome phase.

2.9.3. Whole-brain multivariate classifications: comparing reactive and non-reactive guilt processing

To compare the neurocognitive processing of guilt before (i.e., Outcome phase) and after (i.e., Attitude phase) receiving the victim's attitude feedback, we first developed a whole-brain classifier of reactive guilt to discriminate the High vs. Low guilt conditions in Attitude phase. Then we examined how well this reactive guilt classifier in Attitude phase discriminated the High vs. Low guilt conditions in Outcome phase, and conversely, how well the non-reactive guilt classifier in Outcome phase discriminated the High vs. Low guilt conditions in Attitude phase (i.e., cross-phase classification). We also compared the neural processing of reactive guilt in this study and that of non-reactive guilt reported in Yu et al. (2020a). We examined how well the reactive guilt classifier in Attitude phase discriminated the High (Self_Responsibile) vs. Low (Both_Responsibile) guilt conditions in Yu et al. (2020a), and conversely, how well the non-reactive guilt classifier developed in Yu et al. (2020a) discriminated the High vs. Low guilt conditions in Attitude phase (i.e., cross-study classification).

We also tested this hypothesis using values of pattern expression. We computed the pattern expression values for the six conditions in Attitude phase based on the non-reactive guilt pattern in Yu et al. (2020a) and based on the pattern classifier for harm in Outcome phase, and tested whether the pattern expression values generated from these two patterns could predict the guilt ratings in the corresponding conditions in Attitude phase. Conversely, we computed the pattern expression values for the three conditions (Self_Responsibile, Both_Responsibile, Co-player_Responsibile) in Yu et al. (2020a) based on the pattern classifier for reactive guilt in Attitude phase, and tested whether these pattern expression values could predict the guilt ratings in the corresponding conditions in Yu et al. (2020a). Given the lack of guilt ratings in Outcome phase as stated above, we could not conduct analyses regarding whether the pattern expression values in Outcome phase generated from the pattern classifier for reactive guilt in Attitude phase could predict the guilt ratings in Outcome phase.

As a supplementary analysis, we examined whether the responses of reactive anger in Attitude phase could be distinguished by the multivariate pattern of anger developed in a previous meta-analysis (Wager et al., 2015). This meta-analytical pattern of anger was used to discriminate High vs. Low anger conditions in Attitude phase in the current study (i.e., cross-study classification).

2.9.4. Functional parcellation-based MVPA

We searched for specific brain regions that were involved in reactive guilt and reactive anger processing in Attitude phase and harm processing in Outcome phase respectively. To reduce the search space in the brain, we used an a priori 200-parcel whole-brain parcellation based on meta-analytically functional co-activation of the Neurosynth database (Chang et al., 2021; de la Vega et al., 2016; van Baar et al., 2019) (parcellation available at <http://neurovault.org/images/39711/>) and divided each contrast image for each condition and each participant into 200 parcels. The use of a parcellation scheme has several advantages over

the more conventional searchlight approach, such as less computationally demanding and higher homogeneity with functional neuroanatomy (Chang et al., 2021; Craddock et al., 2012; van Baar et al., 2019), and has been proven efficient in multivariate based analysis (Chang et al., 2021; van Baar et al., 2019). Next, to identify parcels contributing to reactive guilt processing, for each parcel, we applied SVM (Friedman et al., 2001; Wager et al., 2013) to train a multivariate pattern classifier discriminating High guilt vs. Low guilt groups in Attitude phase (Chang et al., 2015; Wager et al., 2013; Woo et al., 2014). The same set of analyses was conducted to identify parcels contributing to reactive anger processing in Attitude phase and non-reactive guilt processing in Outcome phase. Results were thresholded at $q < 0.05$, FDR (false discovery rate) corrected, two-tailed.

Post hoc permutation tests were performed for each identified parcel to illustrate how likely parcel-wise classification accuracies were achieved by chance, compared with data-driven permutation-based null distributions. For each parcel, by resampling the order of contrast images with 2500 permutations (2500-fold), we computed the classification accuracies for reactive guilt and anger in each shuffled sample and the probability of the estimated classification accuracies in permutations being greater than the observed classification accuracies (i.e., permutation p). To further identify regions that were more sensitive to reactive guilt processing than for reactive anger processing, for each permutation, we computed the difference between classification accuracies for reactive guilt and reactive anger for each parcel. Permutation p s for accuracy differences were computed for the probability of the estimated accuracy differences in permutations being greater than the observed accuracy differences. The same set of analyses was conducted to further identify regions that were more sensitive to reactive anger processing than for reactive guilt processing in Attitude phase, and regions that showed different sensitivities to non-reactive guilt processing in Outcome phase and reactive guilt processing in Attitude phase. Results were thresholded at $q < 0.05$, FDR corrected, two-tailed.

2.9.5. Shared and differential neural processing in overlapped regions for reactive guilt and anger

Although we observed that the involvements of dACC, pre-SMA, dmPFC and ventral striatum in the processing of reactive guilt and anger in Attitude phase, it is possible that the patterns for reactive guilt and anger in these overlapped regions may be different from each other. To test this possibility, we computed pattern expression values for the six conditions in Attitude phase based on guilt pattern and anger pattern respectively for each of the four regions. If there existed a shared neural representation of guilt and anger in a given brain region, then 1) the condition-wise pattern expression values obtained from the guilt pattern and from the anger pattern should be positively correlated, and 2) the guilt classifier should be able to discriminate High and Low anger conditions, and the anger classifier should be able to discriminate High and Low guilt conditions. Therefore, for each of the four regions, we tested whether the pattern expression values for the six conditions in Attitude phase generated from guilt pattern and those generated from anger pattern were correlated with each other using LMMs. In each LMM, pattern expression values for reactive guilt and anger were regarded as dependent variable and fixed effect respectively, with by-participant random slopes for the fixed effect included. Moreover, for each of the four regions, we calculated the forced-choice classification accuracies for how well High and Low guilt groups were correctly classified by the anger pattern and how well High and Low anger groups were correctly classified by the guilt pattern.

2.9.6. False-positive control

For behavioral analyses, we used an independent replication to minimize potential false positives of results. For fMRI analyses, in addition to the FDR corrections applied for functional parcellation-based MVPA, we conducted FDR corrections ($q < 0.05$, two-tailed) on other fMRI analyses to minimize false positives (i.e., Type II error). We explicitly defined six

families of analyses with separate and independent statistical hypotheses and conducted multiple comparison corrections within each family:

- 1) *Whether the reactive guilt pattern classifier identified in the current study could distinguish between High vs. Low guilt conditions and predict guilt ratings in Attitude phase ($N_{Test} = 3$)?* This family included the binary classification between High vs. Low guilt conditions in Attitude phase, the AVOVA and the LMM analyses on the pattern expression values generated from the reactive guilt pattern classifier.
- 2) *Whether the reactive anger pattern classifier identified in the current study could distinguish between High vs. Low anger conditions and predict anger ratings in Attitude phase ($N_{Test} = 3$)?* This family included the binary classification between High vs. Low anger conditions in Attitude phase, the AVOVA and the LMM analyses on the pattern expression values generated from the reactive anger pattern classifier.
- 3) *Whether there existed shared or differential representations for reactive guilt and reactive anger in Attitude phase ($N_{Test} = 21$)?* This family included cross-emotion binary classifications between reactive guilt and anger, and the LMMs testing the relationships between pattern expression values for reactive guilt and anger, both at whole-brain level and local level (i.e., VS, dACC, pre-SMA, and dmPFC).
- 4) *Whether the non-reactive guilt in Outcome phase could be generalized to and be distinguished by the previous study on non-reactive guilt (Yu et al., 2020a) ($N_{Test} = 4$)?* This family included the binary classification between High vs. Low harm conditions in Outcome phase, the cross-study binary classifications and the LMM analysis testing the relationships between pattern expression values for non-reactive guilt (harm pattern) in Outcome phase and non-reactive guilt in Yu et al. (2020a).
- 5) *Whether there existed shared or differential representations for reactive guilt and non-reactive guilt ($N_{Test} = 9$)?* This family included cross-phase binary classifications between non-reactive guilt (harm pattern) in Outcome phase and reactive guilt in Attitude phase, cross-study binary classifications between non-reactive guilt in Yu et al. (2020a) and reactive guilt in Attitude phase, and corresponding LMM analyses on pattern expression values, both at whole-brain level and local level (i.e., VS).
- 6) *Whether patterns for other related psychological processes could discriminate conditions in the current study.* This series of analyses included the whole-brain patterns for physical pain (Woo et al., 2014), social rejection (Woo et al., 2014), vicarious pain (Krishnan et al., 2016), empathic distress and empathic care (Ashar et al., 2017), and skin conductance and heart rate (Eisenbarth et al., 2016) established in previous studies. We tested whether these patterns could discriminate responses of non-reactive guilt (High vs. Low harm conditions) in Outcome phase (sub-family 1; $N_{Test} = 7$), reactive guilt (High vs. Low guilt conditions) in Attitude phase (sub-family 2; $N_{Test} = 7$), and reactive anger (High vs. Low anger conditions) in Attitude phase (sub-family 3; $N_{Test} = 7$), respectively. Given that separate and independent hypotheses were tested in these three sub-families, multiple comparison corrections were conducted within each sub-family.

2.9.7. Statistical power analysis

We conducted statistical power simulations to formally test 1) whether the current sample size of fMRI study was adequate to draw our main conclusions regarding the neural processes, and 2) whether the insignificant correlations between pattern expression values for reactive guilt and anger arose from the differential neural representations or just resulted from a limited power of analyses. To our knowledge, there is no standard way of calculating effect size and statistical power for the multivariate classification analysis directly from the fMRI data. Therefore, we focused on pattern expression values generated from the multivariate analyses instead. For each of the LMM analyses on pattern expression values, we computed the observed statistical power, as well as the simulations of statistical power assuming the number of partic-

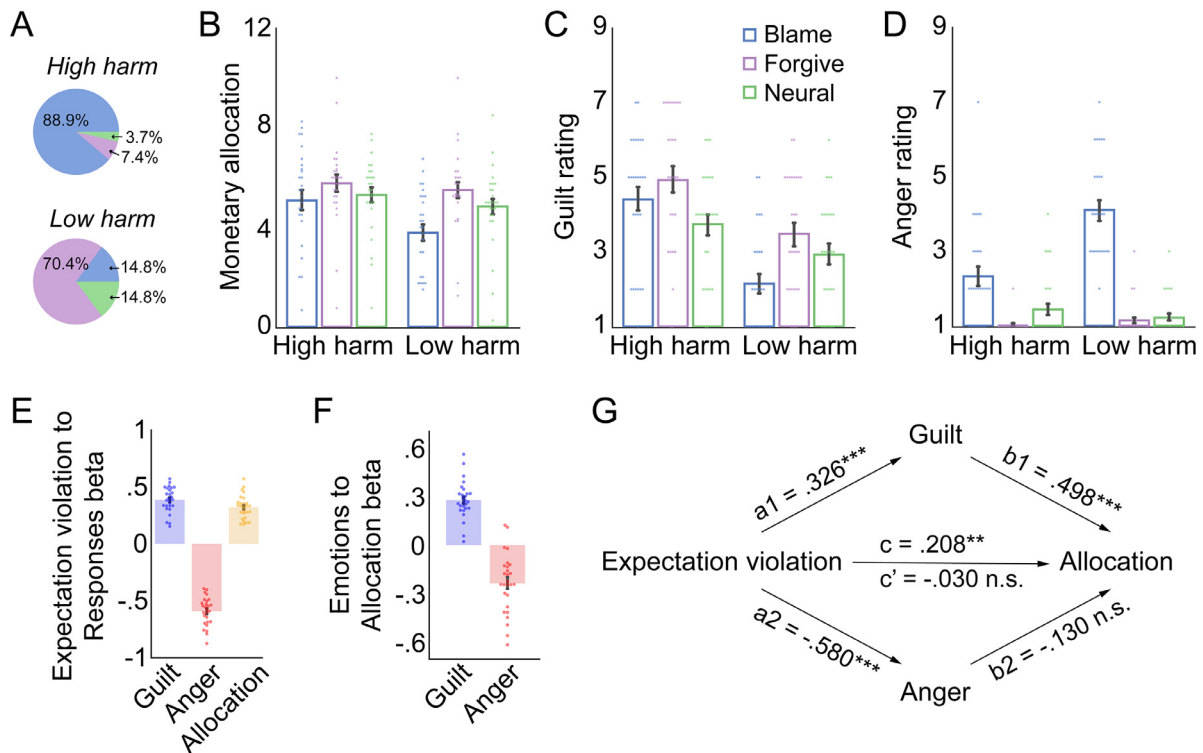


Fig. 2. Behavioral results. (A–D) Participants' pre-experiment prediction of the co-player's attitudes in High and Low harm conditions (A), amount of monetary allocation (B), post-experiment rating of guilt (C), and post-experiment rating of anger (D) in the six conditions. The scale of monetary allocation ranged from 0 to 10 points (1 point = 2 Yuan). The scales of guilt and anger ratings ranged from 1 ('not at all') to 7 ('very strong'). (E) The regression betas indicating the contributions of expectation violation to the amount of monetary allocation, and post-experiment ratings of guilt and anger. (F) The regression betas indicating the contributions of post-experiment ratings of guilt and anger to the amount of monetary allocation. (G) Post-experiment ratings of guilt and anger significantly mediated the effect of expectation violation on the amount of monetary allocation. *n.s.* $p > 0.05$, * $p < .05$, ** $p < .01$, *** $p < .001$. Error bars represent SEs.

ipients ranging from 0 to 200 using “simr” package in R (Green and MacLeod, 2016) (Fig. S3).

3. Results

3.1. Effects of Harm and Attitude of victim on guilt, anger and monetary allocation

Prior to the task, 88.9% of the participants predicted that the co-player would blame them for high harm, and 70.4% of the participants predicted that the co-player would forgive them for low harm. This indicated that harm severity manipulation successfully modified participants' expectations of the co-player's attitudes (Fig. 2A). Note that the patterns of results reported below remained the same if the data of participants whose pre-task predictions were contrary to those of the majority were excluded.

Table 1 presents descriptive statistics for the amount of allocation and ratings of guilt and anger in each condition. At behavioral level, 2 (Harm to the co-player: High vs. Low harm) \times 3 (Attitude of the co-player: Blame vs. Forgive vs. Neutral) repeated-measures analysis of variance (ANOVA) revealed significant main effects of Harm ($F_{1,26} = 23.62$, $p < 0.001$, $\eta^2_{\text{partial}} = 0.48$, statistical power = 0.99) and Attitude ($F_{2,52} = 14.98$, $p < 0.001$, $\eta^2_{\text{partial}} = 0.37$, power = 1.00) on the amount of monetary allocation (Fig. 2B). On one hand, the amount of monetary allocation was greater in High harm conditions than in Low harm conditions ($F_{1,26} = 23.62$, $p < 0.001$, $\eta^2_{\text{partial}} = 0.48$, power = 1.00). On the other hand, the amount of monetary allocation was lower in Blame conditions than in Forgive conditions (with Bonferroni correction, $t_{53} = -4.27$, $p = 0.001$, *Cohen's d* = 0.70) and Neutral conditions ($t_{53} = -3.17$, $p = 0.012$, *Cohen's d* = 0.38). The amount

of monetary allocation was higher in Forgive conditions than in Neutral conditions ($t_{53} = 3.55$, $p = 0.004$, *Cohen's d* = 0.37). Further, the 2 (Harm: High vs. Low harm) \times 3 (Attitude: Blame vs. Forgive vs. Neutral) interaction effect was significant ($F_{2,52} = 6.74$, $p = 0.003$, $\eta^2_{\text{partial}} = 0.21$, power = 0.90). In Low harm conditions, the amount of monetary allocation was lower in Blame condition than in Forgive condition ($t_{26} = -5.27$, $p < 0.001$, *Cohen's d* = 1.02) and Neutral condition ($t_{26} = -4.05$, $p = 0.001$, *Cohen's d* = 0.62); such differences were reduced in High harm conditions (Blame vs. Forgive: $t_{26} = -2.09$, $p = 0.140$, *Cohen's d* = 0.35; Blame vs. Neutral: $t_{26} = -0.82$, $p = 1.000$, *Cohen's d* = 0.11).

For self-reported emotions, we found significant main effects of Harm on ratings of guilt ($F_{1,26} = 68.66$, $p < 0.001$, $\eta^2_{\text{partial}} = 0.73$, power = 1.00; Fig. 2C) and anger ($F_{1,26} = 26.35$, $p < 0.001$, $\eta^2_{\text{partial}} = 0.50$, power = 1.00; Fig. 2D) after receiving the co-player's attitude feedbacks (i.e., participants' reactive social emotions). Specifically, High harm conditions were associated with higher level of guilt and lower level of anger, relative to Low harm conditions. Attitude also exhibited significant main effects on guilt ratings ($F_{2,52} = 7.38$, $p < 0.001$, $\eta^2_{\text{partial}} = 0.22$, power = 0.93) and anger ratings ($F_{2,52} = 70.60$, $p < 0.001$, $\eta^2_{\text{partial}} = 0.73$, power = 1.00). The feeling of guilt was higher in Forgive conditions than in Blame conditions ($t_{53} = 3.17$, $p = 0.012$, *Cohen's d* = 0.62) and Neutral conditions ($t_{53} = 3.59$, $p = 0.004$, *Cohen's d* = 0.59); no difference was observed between Blame and Neutral conditions ($t_{53} = -0.20$, $p = 1.000$, *Cohen's d* = 0.04). In contrast, the feeling of anger was greater in Blame conditions than in Forgive conditions ($t_{53} = 9.18$, $p < 0.001$, *Cohen's d* = 2.40) and Neutral conditions ($t_{53} = 8.50$, $p < 0.012$, *Cohen's d* = 1.99); no difference was observed between Forgive and Neutral conditions ($t_{53} = -2.16$, $p = 0.119$, *Cohen's d* = 0.55). These results indicated that overall, while the victim's forgive-

Table 1
Descriptive statistics for behavioral results.

Experiment	Variable	High harm			Low harm		
		Blame	Forgive	Neutral	Blame	Forgive	Neutral
fMRI	Monetary allocation	5.08 ± 0.42	5.77 ± 0.29	5.30 ± 0.30	3.78 ± 0.34	5.50 ± 0.32	4.84 ± 0.31
	Guilt rating	4.41 ± 0.29	4.93 ± 0.37	3.74 ± 0.28	2.15 ± 0.25	3.48 ± 0.30	2.93 ± 0.28
	Anger rating	2.33 ± 0.27	1.04 ± 0.04	1.44 ± 0.16	4.11 ± 0.28	1.15 ± 0.09	1.22 ± 0.10
Behavioral Replication	Monetary allocation	2.43 ± 0.29	3.71 ± 0.38	2.86 ± 0.32	1.64 ± 0.27	3.24 ± 0.37	2.56 ± 0.29
	Guilt rating	4.47 ± 0.27	4.87 ± 0.29	3.17 ± 0.23	2.29 ± 0.22	3.97 ± 0.26	2.93 ± 0.21
	Anger rating	2.86 ± 0.31	1.30 ± 0.15	1.47 ± 0.16	3.43 ± 0.34	1.13 ± 0.10	1.43 ± 0.16

Note: Each value in a cell represents Mean ± SE. During the task, participants made monetary allocation on a scale from 0 to 10 points in one-third of the trials. After the task, they rated their feelings of guilt and anger to co-players after the co-player's attitude was revealed on a scale of 1 ('not at all') to 7 ('very strong') under each of the six conditions respectively.

ness induced increased feeling of guilt in the transgressor, the victim's blame fueled feeling of anger.

Importantly, significant 2 (Harm: High vs. Low harm) * 3 (Attitude: Blame vs. Forgive vs. Neutral) interaction effects were observed on ratings of both guilt ($F_{2,52} = 10.77, p < 0.001, \eta^2_{\text{partial}} = 0.29, \text{power} = 0.97$; Fig. 2C) and anger ($F_{2,52} = 30.23, p < 0.001, \eta^2_{\text{partial}} = 0.54, \text{power} = 1.00$; Fig. 2D). On one hand, in Low harm conditions, the feeling of guilt was lower and the feeling of anger was higher in Blame condition than in Forgive condition (guilt: $t_{26} = -5.21, p < 0.001, \text{Cohen's } d = 0.91$; anger: $t_{26} = 10.79, p < 0.001, \text{Cohen's } d = 2.75$) and Neutral condition (guilt: $t_{26} = -2.95, p = 0.020, \text{Cohen's } d = 0.56$; anger: $t_{26} = 10.35, p < 0.001, \text{Cohen's } d = 2.66$); such differences were significantly reduced in High harm conditions (guilt Blame vs. Forgive: $t_{26} = -1.25, p = 0.667, \text{Cohen's } d = 0.29$; guilt Blame vs. Neutral: $t_{26} = 1.84, p = 0.231, \text{Cohen's } d = 0.41$; anger Blame vs. Forgive: $t_{26} = 4.78, p < 0.001, \text{Cohen's } d = 1.30$; anger Blame vs. Neutral: $t_{26} = 3.25, p = 0.010, \text{Cohen's } d = 0.77$). Since a large proportion of participants predicted that the co-player would forgive them in Low harm conditions, these results indicated that an unexpected negative attitude feedback (i.e., blame) in Low harm condition might induce reduced feeling of guilt and enhanced feeling of anger. On the other hand, in High harm conditions, compared with Neutral condition, the feeling of guilt was higher ($t_{26} = 4.12, p = 0.001, \text{Cohen's } d = 0.69$) and the feeling of anger was lower ($t_{26} = -2.51, p = 0.056, \text{Cohen's } d = 0.67$) in Forgive condition; such differences were significantly reduced in Low harm conditions (guilt: $t_{26} = 2.20, p = 0.110, \text{Cohen's } d = 0.36$; anger: $t_{26} = -0.57, p = 1.000, \text{Cohen's } d = 0.15$). Since a large proportion of participants predicted that the co-player would blame them in High harm conditions, these results indicated that an unexpected positive attitude feedback (i.e., forgiveness) in High harm condition might induce increased feeling of guilt and decreased feeling of anger. These results demonstrated potential relationships between expectation violation and feelings of guilt and anger.

To directly examine the relationships between expectation violation and feelings of guilt and anger, we calculated expectation violation of co-player's attitude in each condition according to participants' predicted attitudes of co-player before the task. Consistent with our hypothesis, results of linear mixed modeling (Fig. 2E) revealed that while self-reported guilt was positively correlated with the level of expectation violation ($\beta = 0.37 \pm 0.07$ (SE), $t = 5.84, p < 0.001, \text{power} = 1.00$), self-reported anger was negatively correlated with the level of expectation violation ($\beta = -0.59 \pm 0.07, t = -8.44, p < 0.001, \text{power} = 1.00$). Moreover, the amount of allocation was positively with the level of expectation violation ($\beta = 0.32 \pm 0.05, t = 6.00, p < 0.001, \text{power} = 1.00$).

Next, we conducted linear mixed modeling and model comparison to examine whether guilt and anger, rather than other possible emotions such as gratitude, sadness and embarrassment, served as the main motivations for monetary allocation. Results demonstrated that the model with both ratings of guilt and anger as predictors for the amount of allocation outperformed other models that had: (1) only one predictor of ei-

ther guilt or anger, (2) guilt, anger and gratitude as predictors, (3) guilt, anger and sadness as predictors, (4) guilt, anger and embarrassment as predictors, (5) gratitude, sadness and embarrassment as predictors (Table S3 in *Supplementary Materials*). Parameters estimated based on this winning model showed that while ratings of guilt contributed positively to allocation ($\beta = 0.28 \pm 0.06, t = 4.86, p < 0.001, \text{power} = 1.00$; Fig. 2F), ratings of anger contributed negatively to allocation ($\beta = -0.23 \pm 0.07, t = -3.44, p = 0.002, \text{power} = 0.99$; Fig. 2F).

Finally, we examined whether feelings of guilt and anger mediated the effect of expectation violation on the amount of allocation using multivariate mediation model analysis. A multiple mediation model with the ratings of guilt and anger as mediators simultaneously (Fig. 2G) showed that the total indirect effect of this model was significant, with normalized coefficient of overall mediating effect = 0.238, $p < 0.001, c = 0.208, p = 0.006, c' = -0.030, p = 0.724$, a complete mediation. The normalized coefficients of the mediating effect of guilt and anger were 0.162, $p < 0.001$ and 0.076, $p = 0.115$, respectively. No significant difference was observed between the mediating effect of guilt and anger, 0.087, $p = 0.180$. The Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy (Hair et al., 2006) and Bartlett's test of sphericity (Tobias and Carlson, 1969) showed that the current dataset was adequately sampled and met the criteria for structural equation modeling (KMO value = 0.61 > 0.60, Bartlett's test $\chi^2 = 139.42, df = 6, p < 0.001$). This model performed well with comparative fit indices exceeding the > 0.90 acceptable threshold (CFI = 1.00, TLI = 1.00), and the root mean square error of approximation and the standardized root mean squared residual were within the reasonable fit range of < 0.08 (RMSEA < 0.001, SRMR = 0.013) (Browne and Cudeck, 1992; Hu, 1995; West et al., 2012). These results were in line with the hypothesis that social expectation violation may serve as an important factor that influences the transgressor's feelings of guilt and anger after receiving the victim's attitude feedback, which in turn influences subsequent allocation behaviors. Note, the above behavioral results of fMRI experiment were independently replicated in an additional behavioral experiment (Table 1, Table S2, and Fig. S1).

3.2. Whole-brain pattern classifications for reactive guilt and anger in Attitude phase

At neural level, we first investigated the neural bases underlying participants' reactive guilt and anger after receiving victim's attitude (i.e., Attitude phase) using multivariate pattern analysis (MVPA). For each participant, we categorized the conditions in Attitude phase into "High guilt" and "Low guilt" groups based on post-experiment guilt ratings of this participant. Using the images of all participants, we applied linear support vector machine (SVM) (Friedman et al., 2001; Wager et al., 2013) to train a whole-brain multivariate pattern classifier for reactive guilt discriminating these two groups of maps. With a 5-folds cross-validation method in which the images from the same participant were held out together, we calculated the accuracy and significance of the

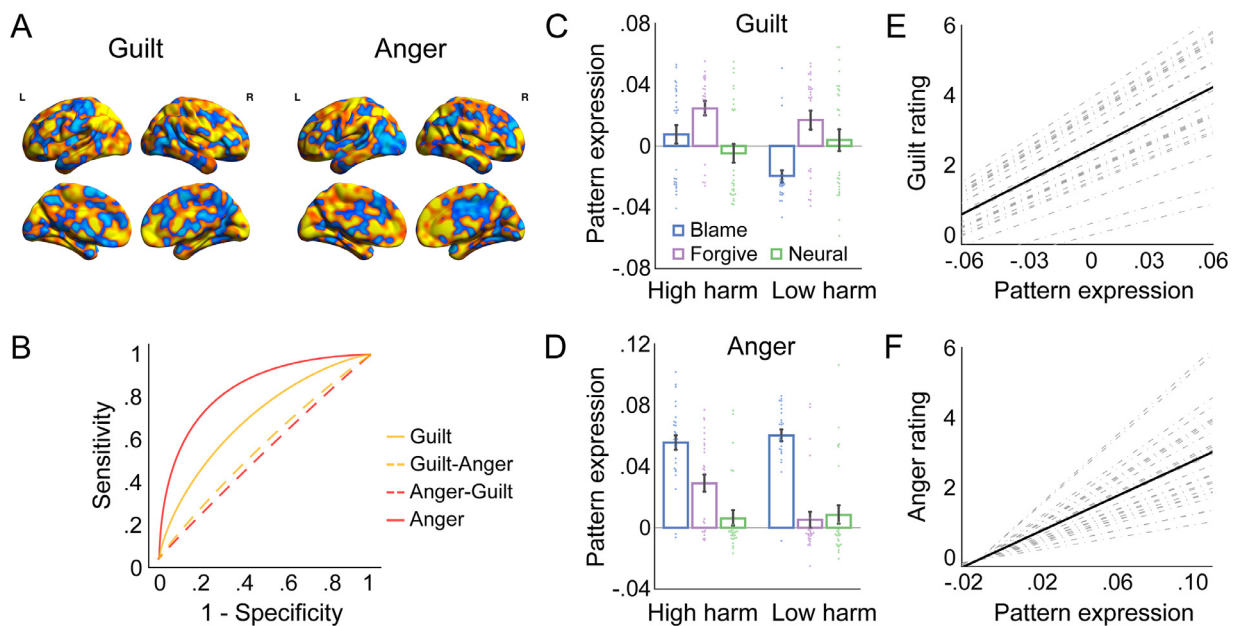


Fig. 3. Whole-brain multivariate pattern analyses for reactive guilt and anger in Attitude phase. (A) Whole-brain multivariate patterns discriminating High vs. Low guilt conditions and High vs. Low anger conditions in Attitude phase. (B) Receiver operating characteristic curves (ROCs) for within-emotion and cross-emotion classifications. Orange solid, cross-validations for High vs. Low guilt conditions in Attitude phase; orange dash, using High vs. Low guilt pattern to predict High vs. Low anger conditions; red solid, cross-validations for High vs. Low anger conditions in Attitude phase; red dash, using High vs. Low anger pattern to predict High vs. Low guilt conditions. (C and D) Pattern expression values of reactive guilt and reactive anger in the six conditions respectively. The modes of the pattern expressions of guilt and anger were consistent with the patterns of post-experiment ratings of guilt and anger. (E and F) Regression analyses showed that the condition-wise pattern expression values of guilt and anger were predictive of corresponding guilt and anger ratings. Error bars represent SEs.

SVM classifiers using the forced-choice discrimination test (see details in *Materials and Methods*, Chang et al., 2015; Wager et al., 2013; Woo et al., 2014). The same set of analyses was conducted to train a whole-brain multivariate pattern classifier for reactive anger by categorizing the conditions in Attitude phase into "High anger" and "Low anger" groups based on post-experiment anger ratings. To be noted, in the current study, the categorization of guilt groups were not correlated with that of anger groups ($\beta = 0.27 \pm 0.45$, $z = 0.61$, $p = 0.545$), which ensured that the multivariate pattern classification procedures for guilt and anger were independent from each other. Results showed that the whole-brain classifier for reactive guilt yielded an average classification accuracy of 70.4% ($\pm 9.0\%$), $p = 0.031$, $p_{FDR} = 0.031$, and the whole-brain classifier for reactive anger yielded an average classification accuracy of 81.5% ($\pm 7.6\%$), $p < 0.001$, $p_{FDR} = 0.002$ (Fig. 3, A and B).

To test whether the classifiers for reactive guilt could predict condition-wise post-experiment ratings of guilt, we obtained condition-wise pattern expression values of guilt for each participant by computing the dot product of the guilt classifier and each of the six contrast images in Attitude phase. This value reflects the distance between a given activation map and the classifier represented by a hyperplane in the feature space. The same analysis was conducted to obtain condition-wise pattern expression values of anger for each condition and each participant. Results showed that the modes of the pattern expressions of guilt and anger were consistent with the patterns of post-experiment ratings of guilt and anger respectively, with significant 2 (Harm: High vs. Low harm) * 3 (Attitude: Blame vs. Forgive vs. Neutral) interactions in pattern expressions of guilt and anger (guilt: Fig. 3C, $F_{2, 52} = 6.12$, $p = 0.004$, $p_{FDR} = 0.006$, $\eta^2_{partial} = 0.19$, power = 0.87; anger: Fig. 3D, $F_{2, 52} = 5.44$, $p = 0.007$, $p_{FDR} = 0.007$, $\eta^2_{partial} = 0.17$, power = 0.83). Regression analysis showed that these condition-wise pattern expression values of guilt were predictive of the corresponding guilt ratings ($\beta = 0.54 \pm 0.05$, $t = 10.20$, $p < 0.001$, $p_{FDR} = 0.003$, power = 1.00; Fig. 3E and Fig. S3A); so did the pattern expression values of anger ($\beta = 0.62 \pm 0.08$,

$t = 7.37$, $p < 0.001$, $p_{FDR} = 0.002$, power = 1.00; Fig. 3F and Fig. S3B). These results indicated that the classifiers (i.e., multivariate brain activation patterns) of reactive guilt and anger accurately reflected neural processing of each emotion in Attitude phase. Moreover, cross-emotion classification revealed that the pattern for guilt was not able to distinguish between High and Low anger conditions (accuracy = $63.0 \pm 12.1\%$, $p = 0.248$, $p_{FDR} = 0.372$), nor could the pattern for anger distinguish between High and Low guilt conditions in Attitude phase (accuracy = $63.0 \pm 12.1\%$, $p = 0.248$, $p_{FDR} = 0.372$) (Fig. 3B). The pattern expression values of guilt and anger did not significantly correlated with each other, $\beta = 0.03 \pm 0.10$, $t = 0.36$, $p = 0.725$, $p_{FDR} = 0.761$, power = 0.10. Power simulation indicated that this correlation, even if existed, was so small that it would be hard to detect (power = 0.2) even with a very large sample, e.g., $N = 200$ (Fig. S3C). These results indicated different neural representations of reactive guilt and reactive anger during Attitude phase at whole brain level.

3.3. Shared and differential local classifiers for reactive guilt and anger after receiving co-player's attitude feedbacks

Next, we searched for specific brain regions that were involved in reactive guilt and reactive anger processing respectively in Attitude phase, using an a priori 200-parcel whole-brain parcellation based on meta-analytic functional co-activation of the Neurosynth database (de la Vega et al., 2016; van Baar et al., 2019). Within each parcel, we applied linear SVM (Friedman et al., 2001; Wager et al., 2013) to train local classifiers discriminating High vs. Low guilt groups and discriminating High vs. Low anger groups, respectively. The use of a parcellation scheme has several advantages over the more conventional searchlight approach, such as less computationally expensive and higher homogeneity with functional neuroanatomy (Chang et al., 2021; Craddock et al., 2012; van Baar et al., 2019). On one hand, we observed that ventral striatum (VS), and regions in medial prefrontal cortex (mPFC), including dorso-

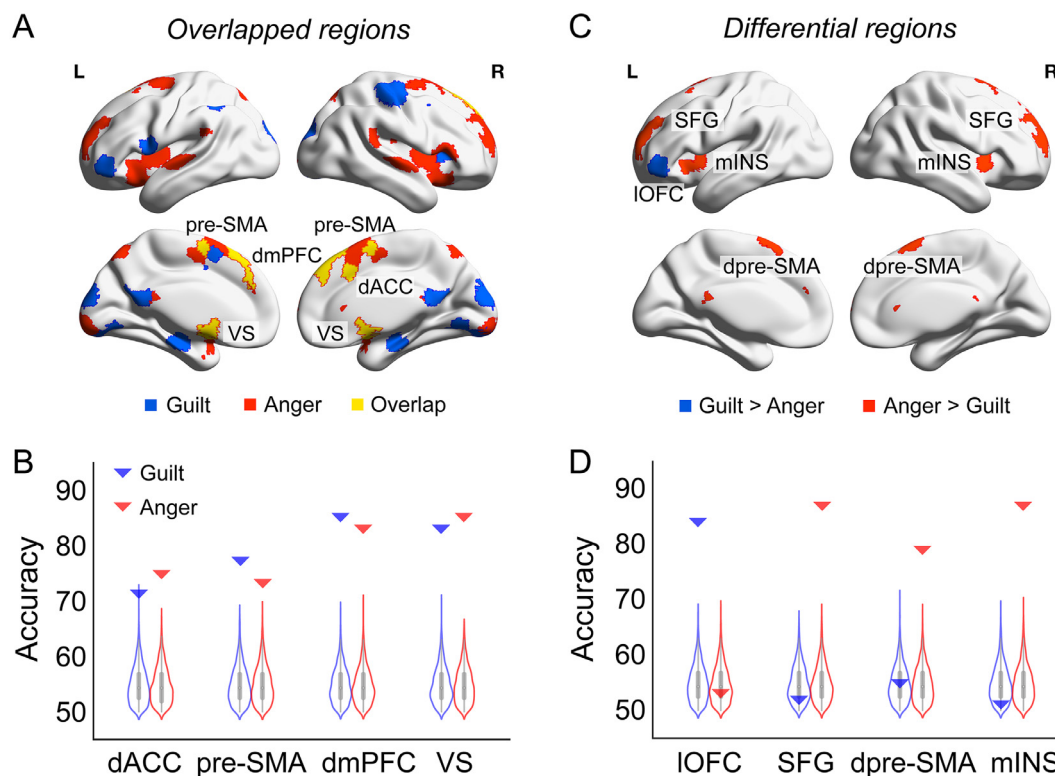


Fig. 4. Shared and differential neural local classifiers for reactive guilt and reactive anger after receiving the co-player's attitude feedbacks. (A) Local classifiers that significantly discriminating High vs. Low guilt conditions (blue) and High vs. Low anger conditions (red) in Attitude phase, with yellow parts indicating the overlapping regions for these two emotions. (B) Classifications accuracies for High vs. Low guilt conditions (blue triangle) and High vs. Low anger conditions (red triangle) in overlapping regions. Each violin plot indicates the accuracy distribution of permutation tests for each classification. (C) Regions more sensitive to guilt than to anger (blue) and more sensitive to anger than to guilt (red). (D) Classification accuracies for High vs. Low guilt conditions (blue triangle) and High vs. Low anger conditions (red triangle) in regions that showed differential sensitivities to reactive guilt and anger. Each triangle represents the accuracy for each classification. Each violin plot indicates the accuracy distribution of permutation tests for each classification. Results were thresholded at $q < 0.05$, FDR corrected, two-tailed.

medial prefrontal cortex (dmPFC), dACC/aMCC, and pre-supplementary motor area (pre-SMA), were sensitive to the processing of both reactive guilt and anger (Fig. 4, A and B; Table S5). On the other hand, we conducted 2500-fold permutation tests for brain regions that showed differential classification abilities for reactive guilt and anger, and observed brain regions that showed differential sensitivity to reactive guilt and anger. Results showed that while left lateral orbitofrontal cortex (IOFC) responded more sensitively to guilt than to anger, superior frontal gyrus (SFG), dorsal pre-supplementary motor area (dpre-SMA) and middle insula (mINS) responded more sensitively to anger than to guilt (Fig. 4, C and D; Table S5).

Although we observed the involvements of dACC, pre-SMA, dmPFC and VS in processing both reactive guilt and anger in Attitude phase, it is possible that the neural representations for guilt and anger in these overlapping regions differed from one another (see patterns for guilt and anger in these regions in Fig. 5A). To test this possibility, we computed pattern expression values for the six conditions in Attitude phase based on the guilt pattern and the anger pattern respectively for each of the four regions. If there existed shared neural representations of guilt and anger in a given brain region, then 1) the condition-wise pattern expression values obtained from the guilt pattern and the anger pattern should be positively correlated, and 2) the guilt classifier should be able to discriminate High vs. Low anger conditions, and the anger classifier should be able to discriminate High vs. Low guilt conditions. Different results were obtained for the four brain regions. First, regression analyses showed that the pattern expression values obtained from the guilt pattern and the anger pattern in VS were positively corre-

lated with each other (Fig. 5B; $\beta = 0.29 \pm 0.09$, $t = 3.21$, $p = 0.004$, $p_{FDR} = 0.012$, power = 0.88). In contrast, the pattern expression values obtained from the guilt pattern and the anger pattern were uncorrelated or negatively correlated with each other in dACC, pre-SMA and dmPFC (Fig. 5B; dACC: $\beta = -0.08 \pm 0.08$, $t = -1.01$, $p = 0.315$, $p_{FDR} = 0.441$, power = 0.04; pre-SMA: $\beta = -0.28 \pm 0.10$, $t = -2.71$, $p = 0.013$, $p_{FDR} = 0.020$, $p_{FDR} = 0.034$, power = 0.68; dmPFC: $\beta = -0.18 \pm 0.09$, $t = -1.98$, $p = 0.061$, $p_{FDR} = 0.116$, power = 0.47). Simulations of sample size and statistical power confirmed that our conclusions would not change if the sample size increased (see *Supplementary Materials* and Fig. S3, D-G). Second, for VS, the pattern expression values in High anger conditions were significantly higher than those in Low anger conditions based on the guilt pattern, with forced-choice classification accuracy of $59.3 \pm 4.7\%$, $p = 0.022$, $p_{FDR} = 0.046$; the pattern expression values in High guilt conditions were significantly higher than those in Low guilt conditions based on the anger pattern, with forced-choice classification accuracy of $59.3 \pm 4.7\%$, $p = 0.022$, $p_{FDR} = 0.046$ (Fig. 5C). In contrast, for dACC, pre-SMA and dmPFC, neither the pattern expression values based on the anger pattern could discriminate High vs. Low guilt conditions, nor the pattern expression values based on the guilt pattern could discriminate High vs. Low anger conditions (Fig. 5, D-F; dACC: guilt to anger accuracy = $53.1 \pm 4.2\%$, $p = 0.480$, $p_{FDR} = 0.560$, anger to guilt accuracy = $53.1 \pm 4.2\%$, $p = 0.480$, $p_{FDR} = 0.560$; pre-SMA: guilt to anger accuracy = $51.2 \pm 4.0\%$, $p = 0.813$, $p_{FDR} = 0.813$, anger to guilt accuracy = $53.1 \pm 4.2\%$, $p = 0.480$, $p_{FDR} = 0.560$; dmPFC: guilt to anger accuracy = $54.6 \pm 4.6\%$, $p = 0.104$, $p_{FDR} = 0.182$, anger to guilt accuracy = $52.4 \pm 4.1\%$, $p = 0.582$, $p_{FDR} = 0.643$). These two

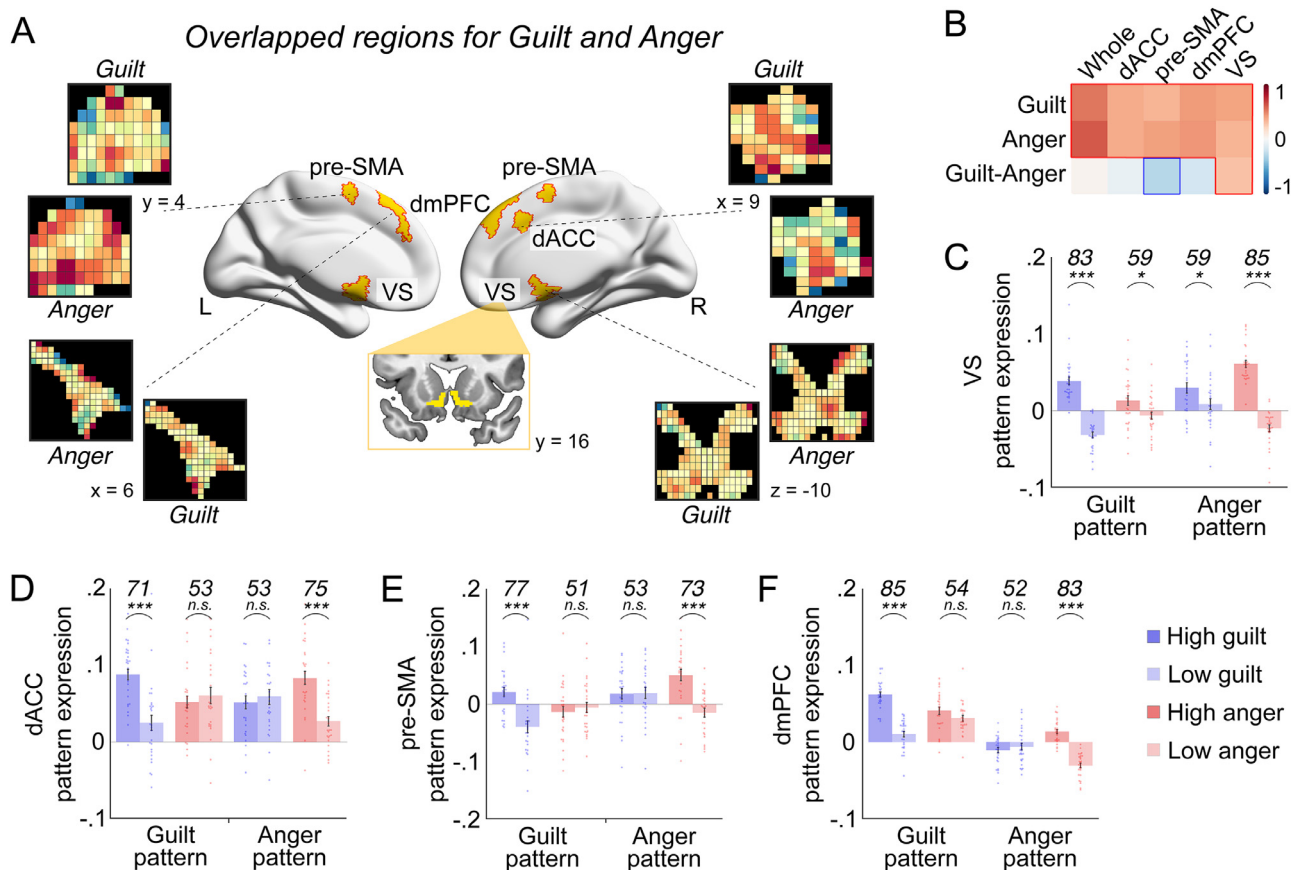


Fig. 5. Shared and differential neural representations for overlapping regions identified for reactive guilt and anger. (A) Local patterns for overlapping regions (i.e., dACC, pre-SMA, dmPFC and VS) for processing reactive guilt and anger. (B) Pattern expression values of reactive guilt and anger in the six conditions were obtained from the whole brain classifier and classifiers in dACC, SMA, dmPFC and VS, respectively. The first row refers to the regression betas capturing the relationships between pattern expression values of guilt and post-experiment guilt ratings in each region. The second row refers to the regression betas capturing the relationships between pattern expression values of anger and post-experiments anger ratings in each region. The third row refers to the regression betas capturing the relationships between pattern expression values of guilt and pattern expression values of anger ratings in each region. (C - F) Pattern expression values for High guilt, Low guilt, High anger and Low anger conditions obtained from guilt patterns and anger patterns in VS (C), dACC (D), pre-SMA (E) and dmPFC (F), respectively. The numbers above each paired bars indicate the forced-choice classification accuracy generated from pattern expression values of the corresponding two conditions. *n.s.* $p > 0.05$, * $p < .05$, ** $p < .01$, *** $p < .001$, *FDR* corrected. Error bars represent *SEs*.

lines of evidence indicated a shared neural representation for reactive guilt and anger in VS, but differential neural representations for reactive guilt and anger in dACC, pre-SMA and dmPFC during Attitude phase.

Since mPFC is associated with diverse psychological processes, including motor function, cognitive control, affect, and social cognition, to avoid the biased reverse inference, we mapped regions identified in the current study onto a mPFC template built by large-scale meta-analysis of human mPFC (de la Vega et al., 2016). This meta-analysis applied a meta-analytic data-driven approach to nearly 10,000 fMRI studies to identify putatively separable regions of mPFC. Multivariate classification analyses aimed at identifying the psychological functions most strongly predictive of activity in each region revealed a tripartite division within MFC, with each zone displaying a relatively distinct functional signature (de la Vega et al., 2016). Using this template, we found that the dmPFC identified in the current study located on the anterior zone of mPFC, with the corresponding sub-region associated preferentially with social information processing. The dACC identified in the current study located on the middle zone, with the corresponding sub-region associated preferentially with conflict, pain, and cognitive control related processing. In contrast, the pre-SMA identified in the current study located on the posterior zone, which was associated preferentially with motor functions (Fig. S2).

3.4. Differential whole-brain patterns for guilt before and after receiving the co-player's attitude feedbacks

Our behavioral results suggested that the participants' emotional responses (guilt and anger) after transgression were dependent not only on the level of harm they caused to the victim, but also on the victim's reactive attitude towards the harm, indicating a potential re-appraisal process on the transgressor's side after receiving social feedbacks. To formally test whether a re-appraisal of social emotions occurred at neural level, we conducted cross-phase and cross-study whole-brain multivariate classifications to examine (1) whether there existed similar or differential neural representations of guilt before (i.e., non-reactive guilt in Outcome phase) and after (i.e., reactive guilt in Attitude phase) receiving the victim's attitude feedback, and (2) whether these neural representations in these two phases were similar to or different from the multivariate brain pattern for guilt in the situation where the victim's attitude was not provided (i.e., non-reactive guilt) in Yu et al. (2020a).

First, previous studies have suggested that a transgressor's guilt is positively correlated with the extent of harm (Berndsen and McGarty, 2010; Renetzky, 2015). This was supported by our behavioral results that participants felt guiltier in High harm conditions than in Low harm conditions. Therefore, we predict that neural differences between High and Low harm conditions in Outcome phase, where the co-

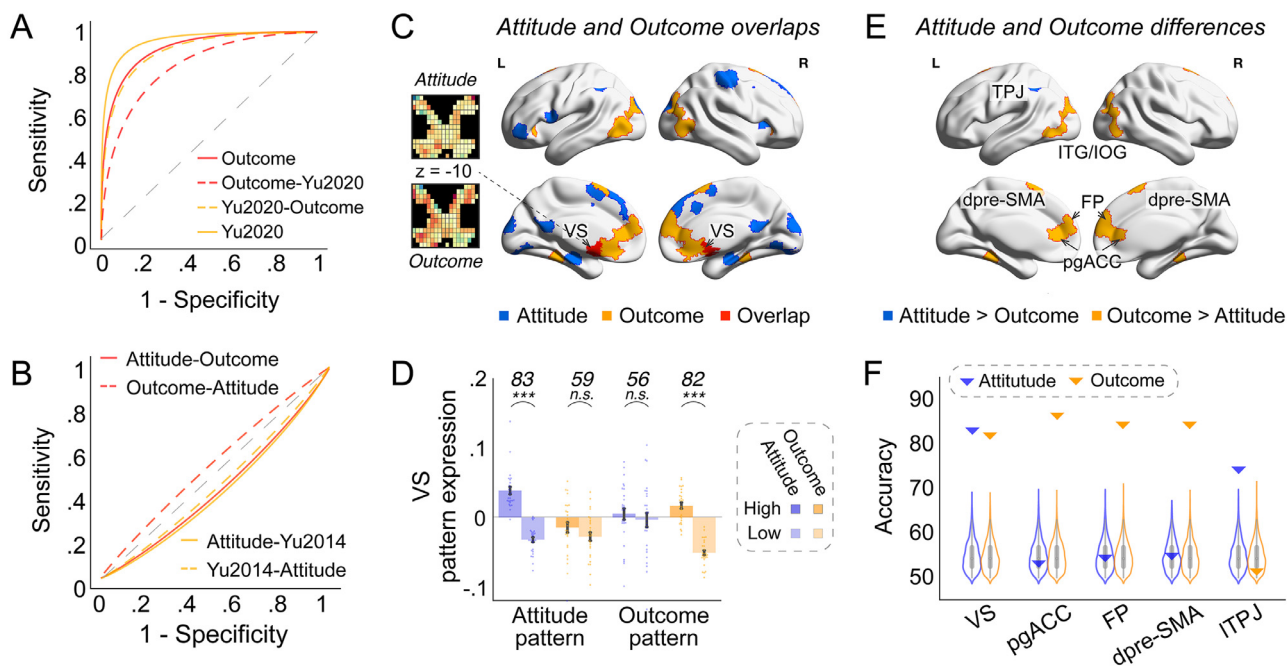


Fig. 6. Shared and differential neural representations for reactive guilt in Attitude phase and non-reactive guilt in Outcome phase. (A) ROCs for the two-choice forced-alternative accuracies for within-study and cross-study classifications using the data of Outcome phase in the current study and of Yu et al. (2020a). Red solid, cross-validations for High vs. Low Harm in Outcome phase; red dash, using Outcome High vs. Low Harm pattern to predict High vs. Low guilt conditions in Yu et al. (2020a); orange solid, cross-validations for High vs. Low guilt in Yu et al. (2020a); orange dash, using Yu et al. (2020a) High vs. Low guilt pattern to predict High vs. Low harm conditions in Outcome phase. (B) ROCs for cross-phase and cross-study classifications. Red solid, using Attitude High vs. Low guilt pattern to predict High vs. Low harm conditions in Outcome phase; red dash, using Outcome High vs. Low harm pattern to predict High vs. Low guilt conditions in phase Attitude; orange solid, using Attitude High vs. Low guilt pattern to predict High vs. Low guilt conditions in Yu et al. (2020a); orange dash, using Yu et al. (2020a) High vs. Low guilt pattern to predict High vs. Low guilt conditions in Attitude phase. (C) Local classifiers that significantly discriminating High vs. Low guilt conditions in Attitude phase (i.e., reactive guilt; blue) and High vs. Low harm conditions in Outcome phase (i.e., non-reactive guilt; orange), with red parts indicating the overlapping regions. Results were thresholded at $q < 0.05$, FDR corrected, two-tailed. (D) Pattern expression values for High guilt and Low guilt conditions in Attitude phase and High harm and Low harm conditions in Outcome phase generated from patterns of Attitude (reactive guilt) and Outcome (non-reactive guilt) phases in VS. The numbers above each paired bars indicate the forced-choice classification accuracy generated from pattern expression values of the corresponded two conditions. (E) Regions responded more sensitively to Attitude than Outcome phases (blue) and regions responded more sensitively to Outcome than Attitude phases (orange). Results were thresholded at $q < 0.05$, FDR corrected, two-tailed. (F) Classifications accuracies for High vs. Low guilt conditions in Attitude phase (blue triangle) and High vs. Low harm conditions in Outcome phase (orange triangle) in regions that showed differential sensitivities in two phases. Each triangle represents the accuracy for each classification. Each violin plot indicates the accuracy distribution of permutation tests for each classification. *n.s.* $p > 0.05$, * $p < .05$, ** $p < .01$, *** $p < .001$, FDR corrected. Error bars represent SEs.

player's attitude was not yet available, should reflect the neural processing of non-reactive guilt, and should be similar to the neural signature of non-reactive guilt established in Yu et al. (2020a). To test this prediction, we used linear SVM (Friedman et al., 2001; Wager et al., 2013) to develop a whole-brain classifier to discriminate High harm vs. Low harm conditions in Outcome phase. This classifier yielded an average classification accuracy of 85.2% ($\pm 7.0\%$, SE), $p < 0.001$, $p_{FDR} = 0.004$ (Fig. 6A). Results of cross-study classification (Fig. 6A) demonstrated that the whole-brain classifier discriminating High harm vs. Low harm conditions in Outcome phase could discriminate High (Self Responsible) vs. Low (Both Responsible) guilt conditions in Yu et al. (2020a), accuracy = 75.0% ($\pm 15.3\%$), $p = 0.023$, $p_{FDR} = 0.031$. The neural signature of guilt in Yu et al. (2020a) could also distinguish between High harm vs. Low harm conditions in Outcome phase in the current study, accuracy = 70.3% ($\pm 13.5\%$), $p = 0.048$, $p_{FDR} = 0.048$. These findings were further supported by the results of pattern expression values. When the classifier for harm in Outcome phase developed on the basis of the current dataset was applied to brain activation maps reported in Yu et al. (2020a), the yielded pattern expression values were able to predict the self-reported guilt in different conditions in Yu et al. (2020a, 2014); $\beta = 0.32 \pm 0.09$, $t = 3.57$, $p = 0.002$, $p_{FDR} = 0.004$, power = 0.91 (Fig. S3H). These results suggested that guilt processing occurring during Outcome phase, which was similar to guilt processing in previous

neuroimaging studies where no reactive attitude was involved (i.e., non-reactive guilt) (Yu et al., 2020a).

Second, the reactive guilt pattern classifier in Attitude phase performed at chance level in discriminating High vs. Low harm conditions in Outcome phase (accuracy = 40.4 \pm 7.8%, $p = 0.442$, $p_{FDR} = 0.885$) and High vs. Low guilt conditions in Yu et al. (2020a) (accuracy = 52.1 \pm 7.5%, $p = 0.885$, $p_{FDR} = 0.885$) (Fig. 6B). Moreover, the pattern classifier for harm in Outcome phase and the guilt signature in Yu et al. (2020a) performed at chance in discriminating High vs. Low guilt conditions in Attitude phase (Outcome: accuracy = 55.5 \pm 10.7%, $p = 0.701$, $p_{FDR} = 0.885$; Yu et al. (2020a): accuracy = 37.0 \pm 7.1%, $p = 0.581$, $p_{FDR} = 0.885$) (Fig. 6B). These null effects were also observed in the pattern expression values. We found that the pattern expression values for the six conditions in Attitude phase generated from the pattern classifier for harm in Outcome phase ($\beta = 0.02 \pm 0.08$, $t = 0.25$, $p = 0.806$, $p_{FDR} = 0.885$, power = 0.03) and from the guilt signature in Yu et al. (2020a) ($\beta = -0.01 \pm 0.07$, $t = -0.17$, $p = 0.868$, $p_{FDR} = 0.885$, power = 0.03) could not significantly predict the guilt ratings in Attitude phase. The pattern expression values for different conditions in Yu et al. (2020a) generated from the pattern classifier for guilt in Attitude phase could not significantly predict the guilt ratings in the corresponding conditions in Yu et al. (2020a; see also Yu et al. (2014)), $\beta = -0.23 \pm 0.13$, $t = -1.76$, $p = 0.102$, $p_{FDR} = 0.405$, power = 0.24. Simula-

tions of sample size and statistical power confirmed that our conclusions would not change if the sample size increased (see *Supplementary Materials* and Fig. S3, I-K). These results suggested that after the co-player's attitude feedback, the neural representation of guilt (i.e., reactive guilt) might differ from the representation of guilt at whole-brain level when the co-player's attitude is not involved (i.e., non-reactive guilt).

As a supplementary analysis, we examined whether the representation of reactive anger in Attitude phase could be distinguished by the multivariate pattern of anger identified in a previous meta-analysis (Wager et al., 2015). It turns out that the meta-analytical anger pattern discriminated High vs. Low reactive anger conditions in the current study with an accuracy of $59.0 \pm 5.3\%$, $p = 0.034$. Given that previous studies on anger mainly focused on the neural bases of the victim's anger in response to transgressions (Blair, 2012; Chang and Smith, 2015; Denson et al., 2009; Klimecki et al., 2018), this result indicates that there might exist shared neural representations underlying the processing of the victim's anger and the transgressor's reactive anger. Unfortunately, since we did not have access to the behavioral and neuroimaging data of any study on the victim's anger, we could not use reactive anger pattern built in the current study to predict the victim's anger in the other study. This deficiency calls for further studies to specifically address the issue.

It is worth noting that the whole-brain patterns for physical pain (Woo et al., 2014), social rejection (Woo et al., 2014), vicarious pain (Krishnan et al., 2016), empathic distress and empathic care (Ashar et al., 2017) or skin conductance and heart rate (Eisenbarth et al., 2016) established in previous studies performed at chance or lower than chance in discriminating High vs. Low harm conditions in Outcome phase, neither were these patterns able to discriminate High vs. Low guilt conditions or High vs. Low anger conditions in Attitude phase. These null effects indicated that the results reported above, to a large extent, were not driven by other processes possibly related to guilt or anger, such as pain or empathy (Table S6).

3.5. Shared and differential neural local classifiers for reactive guilt and non-reactive guilt

Finally, we identified specific brain regions that were involved in non-reactive guilt processing (e.g., High vs. Low harm) during Outcome phase and compared this processing with that of reactive guilt during Attitude phase using the 200-parcel whole-brain parcellation template (de la Vega et al., 2016; van Baar et al., 2019). Results demonstrated the involvements of VS in the processing of guilt during both Outcome phase and Attitude phase (Fig. 6C; outcome accuracy = $81.9 \pm 7.0\%$, $p < 0.001$, attitude accuracy = $83.0 \pm 7.1\%$, $p < 0.001$; permutation $ps < 0.001$). Using the pattern in Attitude phase to predict High vs. Low harm conditions in Outcome phase resulted in a forced-choice classification accuracy of $59.3 \pm 4.7\%$, $p = 0.022$, $p_{FDR} = 0.198$. The pattern in Outcome phase performed at chance level in discriminating High vs. Low guilt conditions in Attitude phase, with forced-choice classification accuracy of $56.2 \pm 4.4\%$, $p = 0.135$, $p_{FDR} = 0.405$ (Fig. 6D). Moreover, we observed differential neural bases between the processing of non-reactive guilt and reactive guilt. Specifically, pgACC, frontal pole, and dpre-SMA showed higher classification accuracy for the non-reactive guilt related processing in Outcome phase (i.e., High vs. Low harm), than for the reactive guilt related processing in Attitude phase (Fig. 6, E and F; see Table S7 for statistics). In contrast, left temporoparietal junction (ITPJ) showed higher classification accuracy for guilt in Attitude phase than in Outcome phase (Fig. 6, E and F; Table S7).

4. Discussion

The mutuality of social emotions have been a focus of debates in moral philosophy (Helm, 2017; Strawson, 1974), but its neurocognitive bases have been largely understudied in psychology and neuroscience

(Schilbach et al., 2013). Here, we demonstrated that the victim's reactive attitudes interacted with the extent of harm to modulate both the strength and the type of the transgressor's social emotions, with expectation violation serving as an important mediator. While unexpected forgiveness following severe harm amplified the participants' guilt, unexpected blame following minor harm reduced the participants' guilt and increased their anger. Our findings demonstrated that social emotions are not only elicited by what are there in the physical world (e.g., our action and its consequences), but are also tuned towards the mental states of our social interactions. This mutuality of social emotions helps us attune to adaptive information that would otherwise not be factored in when we navigate our social world (Sznycer and Lukaszewski, 2019). The mutuality of social emotions is constituted by and enacted through a mutually binding web of expectations – what we as agents are expected to act and what we can expect our social partners would respond to our actions. Being sensitive to these mutual expectations, adjusting our attitudes, emotions and behaviors accordingly are thus a crucial constituent of being a respectful member of a “community of respect” (Helm, 2017).

4.1. The role of social expectation violation in transgressor's reactive guilt and anger

Philosophical theories posit that mutual expectations and reactive attitudes are important driving forces for social emotions arising in dynamic interactions (Darwall, 2013; Strawson, 1974). Consistently, empirical studies reveal that individuals can form expectations about others' attitudes and behaviors according to social norms and past experiences, and learn from expectation violations, enabling them to maximize the benefits with respect to social goals (for reviews, see Olsson et al., 2018; Olsson et al., 2020). Based on this framework, recent theories suggest that social emotions can function as informational error signals that make expectation violations salient to the agent (Chang and Jolly, 2017; Miceli and Castelfranchi, 2014). However, few studies have empirically examined this notion to date. Here we present evidence of how the transgressor's guilt and anger may serve as error signals that enable the transgressor to achieve the social goal of maintaining equity (Decety and Yoder, 2017; Fehr and Schmidt, 1999).

Specifically, consistent with previous studies on non-reactive guilt, when a transgression occurs, the transgressor's guilt enables him/her to recognize this violation (i.e., being in an unjustifiable advantageous status) and implement compensations to restore the balance (Adams, 1965; Baumeister et al., 1994; Blair, 2012; Donnerstein and Hatfield, 1982; Homans, 1974; Kubany and Watson, 2003; Walster et al., 1978). One step further, our results show that the expectation violation derived from the victim's attitude may function as a social affective momentum, which readjusts the tilted balance between self and the other. From the agent's (i.e., transgressor's) perspective, the extra benefit of unexpected forgiveness may amplify the transgressor's advantageous status and lead to reactive guilt and subsequent compensation to eliminate this inequity. In contrast, the unjustified harshness of unexpected blame may reverse the direction of inequity (i.e., disadvantageous inequity), which causes reactive anger that aims to eliminate this inequity by reducing compensation or seeking even more aggressive behaviors. Neurally, the MVPA results revealed a shared neural representation for reactive guilt and anger in VS, a region that is commonly implicated in processing both social and non-social expectation violations in previous studies (for reviews, see Olsson et al., 2020; Ruff and Fehr, 2014; Seo and Lee, 2012). This neural finding, although preliminary, suggests that the expectation violation may function as a potential shared mechanism for reactive guilt and anger, consistent with our behavioral results.

4.2. Differential neural representations for transgressor's reactive guilt and anger

At neural level, in addition to the neural commonalities in VS, we identified differential neural bases of guilt and anger after receiving

the victim's attitudes both in terms of brain regions involved and in terms of neural representations. While reactive guilt recruited IOFC, reactive anger recruited mINS and SFG. These results are consistent with previous studies showing the involvement of IOFC in non-reactive guilt related processing (Wagner et al., 2011; Zhu et al., 2018) and the involvements of SFG and insula in victims' anger related processing (Blair, 2012; Denson et al., 2009; Zhu et al., 2020). More importantly, although overlaps were observed in mPFC, including dmPFC, dACC/aMCC, and pre-SMA, the neural patterns of reactive guilt and anger in these regions could not predict each other, suggesting differential neural representations for reactive guilt and anger. By mapping these three regions identified in the current study onto a mPFC template built by large-scale meta-analysis (10,000 fMRI studies) (Fig. S2), we found that the three regions are located on separate zones in mPFC, which have been associated with different psychological functions (de la Vega et al., 2016): 1) dmPFC is most strongly associated with social processing and mentalizing (Isoda and Noritake, 2013; Schurz and Perner, 2015; Schurz et al., 2014); 2) dACC is located on the mPFC sub-region that is associated preferentially with conflict, pain, and cognitive control related processing (Cavanagh and Shackman, 2015; Heilbronner and Hayden, 2016; Shackman et al., 2011); 3) pre-SMA is associated with motor functions, which are crucial for connecting cognition with action (Nachev et al., 2008). Indeed, the processes of assessing the extent of harm, understanding the victim's attitude, monitoring expectation violation, and executing cognitive control to guide behaviors reflected by the involvements of these three regions are the key psychological components underlying reactive guilt and anger (Helm, 2017). These results indicate that while these processes are implemented in anatomically overlapping areas for reactive guilt and anger, differential neural representations in these regions may contribute to different emotions and the associated behavioral tendencies.

In the current study, we focused on the general processing of reactive guilt and anger, which may cover several psychological components, including appraisals of related contextual information (e.g., the extent of harm, the victim's attitude, and expectation violation), emotional responses, and emotion-specific behavioral tendencies. It is important to test to what extent there exist commonalities and differences between reactive guilt and anger at specific components or levels of processing. Future studies specifically designed to dissociate such psychological components are needed to address this question. Moreover, although we provide evidence showing the shared and differential neural bases underlying reactive guilt and anger, we acknowledge the possibility that there may be effects that failed to survive the threshold in our study given the current sample size.

4.3. Differential neural representations for non-reactive guilt and reactive guilt

Extending previous findings on non-reactive guilt manipulating factors that were independent of the victim's reactions (Basile et al., 2011; Kédia et al., 2008; Koban et al., 2013; Morey et al., 2012; Takahashi et al., 2004; Wagner et al., 2011; Yu et al., 2014; Yu et al., 2020a), our results of cross-phase classifications indicate that the neural representation of reactive guilt after receiving the victim's attitude (Attitude phase) might differ from that of non-reactive guilt before receiving the victim's attitude (Outcome phase). Using functional parcellation-based MVPA, we showed that while pgACC, frontal pole and dorsal pre-SMA showed higher sensitivity to non-reactive guilt, TPJ showed higher sensitivity to reactive guilt. Given that previous studies have consistently shown the involvement of pgACC in self-efficacy related computations (Lockwood and Wittmann, 2018), such as self-esteem (Will et al., 2017) and self-ability processing (Wittmann et al., 2016), and given that the non-reactive guilt is commonly interpreted as a self-conscious and self-blaming emotion (Baumeister et al., 1994; Hoffman, 2000; Tangney et al., 2002; Tangney et al., 2007; Vaish and Hepach, 2019), it is conceivable that we observed activation of pgACC for the non-reactive

guilt. In contrast, TPJ, identified here for reactive guilt processing, is primarily related to mentalizing (Schurz and Perner, 2015; Schurz et al., 2014). The finding of TPJ activation for reactive guilt demonstrates participants' transformation from self-conscious related processes to other-regarding processes after the victim's reactions were known. These contrasting results further highlight the importance of investigating guilt from the perspective of interpersonal interactions (second-person engagement; Schilbach et al., 2013) rather than treating it as purely intrapsychic processes.

Here we analyzed the process of non-reactive guilt by comparing the neural differences between High and Low harm conditions in Outcome phase. One potential concern might be that in the task, participants are directly responsible for the incorrectness of their responses, but not to the level of harm; from this perspective, the extent of harm upon the co-players was incidental. However, theoretically, such incidental causality itself is sufficient to elicit guilt and remorse (Baumeister et al., 1994; McGraw, 1987). In social interactions, taking responsibility and feeling guilty for incidental harm, which one is not strictly responsible for but nevertheless brings about, is a natural response and an indicator of care and prosociality (Anderson et al., 2021; Mason, 2019). The other potential concern might be whether the results of harm-induced non-reactive guilt here could be generalized to those of studies applying other manipulations to induce non-reactive guilt, e.g., high vs. low responsibility for transgression (Baumeister et al., 1994). Results of cross-study neural predictions mitigate these concerns: at whole-brain level, the non-reactive guilt related processes in Outcome phase could be generalized to and predicted by the previous study on non-reactive guilt that manipulated the responsibility for transgression (Yu et al., 2020a), but could not be predicted by other harm-related psychological components (e.g., physical pain, social rejection, vicarious pain, empathic distress and empathic care, or skin conductance and heart rate). Given that the experimental design and stimuli in the current study are different from those in the previous study on non-reactive guilt (Yu et al., 2020a), this cross-study generalization also indicates that our findings, to a large extent, are not simply driven by the differences in experimental design and the processing of stimuli on the screen.

Moreover, by comparing neural representations in Outcome phase and Attitude phase, we provide novel evidence suggesting a re-appraisal process of guilt after receiving the victim's reactive attitude. Two potential accounts for the neural re-appraisal process of guilt in the two phases can be proposed based on our findings. One is that these neural differentiations reflect a reappraisal process of reactive guilt in Attitude phase that is independent from the non-reactive guilt processing in Outcome phase; reactive and non-reactive guilt are generated independently in correspondence to contextual information. Another account is that the reactive guilt in Attitude phase is generated based on the appraisal of non-reactive guilt in Outcome phase. The observed neural findings in Attitude phase may reflect the process of how guilt was evolved from Outcome to Attitude phase based on appraisals on the victim's reactive attitude. Further studies are needed to distinguish these two accounts.

4.4. Implications for future studies

In the current study, given the concern that asking participants to self-report their emotions during the task may explicitly give away the goal of our study and may invite participants to display these emotions due to social desirability (Larsen and Fredrickson, 1999; Nisbett and Wilson, 1977), we used monetary allocation to capture the guilt-induced compensation or anger-induced aggressive behaviors during scanning and used post-experiment ratings as the direct index for guilt and anger. This method has been widely used and proven effective in previous studies on social emotions, such as guilt (Chang et al., 2011; Gao et al., 2018; Li et al., 2020; Yu et al., 2014; Zhu et al., 2018) and gratitude (Liu et al., 2020; Yu et al., 2017; Yu et al., 2018; Zhu et al., 2018). Using this method, here we collect novel evidence regarding the neurocognitive bases of reactive guilt and anger, and identify the neural re-appraisal of

guilt after receiving the victim's attitudes. Our behavioral results were replicated in an independent sample, indicating that our results were not confounded by individual differences in the inability in emotion introspection (Larsen and Fredrickson, 1999; Nisbett and Wilson, 1977). We suggest that the current study opens venues for future investigations and technical developments.

Specifically, by manipulating the victim's reactive attitudes and the extent of harm, we examined the relationship between condition-wise expectation violation and guilt and anger obtained from subjective ratings. However, although we obtained novel evidence to demonstrate the importance of expectation violation, this condition-wise measurement is not sensitive enough to capture how expectation violation, social emotions and the corresponding brain responses vary in dynamic social interactions. Relatedly, due to the need for balancing ecological validity and experimental controllability, the current task is interactive only from the perspective of the participants (i.e., or "reactive" as some researchers define it; Hari et al., 2015), but not from the perspective of the co-players, whose attitude feedbacks were pre-determined by the computer program. We acknowledge that paradigms involving real social interactions between participants are vital for deeper understanding of social emotions and behaviors.

First, the establishments of effective and predictive physical (e.g., facial expressions) and physiological (e.g., skin conductance responses, pupil dilation) measures are needed to monitor the variations of complex social emotional responses (Antony et al., 2021; Chang et al., 2021). Second, recent theoretical work suggest that formalizing social emotions using computational models is critical for characterizing their impact on behaviors and identifying neural and physiological substrates during dynamic social interactions (Chang and Jolly, 2017; Chang and Smith, 2015; Jolly and Chang, 2019). In the current study, we deliberately attempted to minimize the participant's learning on the victim's attitudes or characters by making the co-players anonymous throughout the task. Yet, how expectation violation influences social emotions when the transgressor repeatedly receives and learns about the victim's reactive attitudes is an interesting and theoretically important question (Olsson et al., 2020; Siegel et al., 2018). Further studies combining social learning tasks with quantitative computational modeling are needed to address this question.

Although the view of expectation violation and equity maintaining provides a general framework for understanding inconsistent findings in the literature regarding how the victim's reactive attitudes modulate the transgressor's guilt and anger, there might be other factors that contribute to the reappraisal process but were not measured in the current study. For example, from the interpersonal perspective, the victim's forgiveness or blame may shorten or increase the social distance between the victim and the transgressor, which could in turn modulate the transgressor's social emotions and subsequent behavioral tendencies (Baumeister et al., 1994; Wallace et al., 2008). This social distance perceived from the victim's attitude may serve as another, but possibly correlated, mediator or modulator of the relationship between expectation violation and the transgressor's social emotions. Moreover, expectation violation and the perceived inequity derived from the victim's attitudes may modulate the transgressor's perception of self-responsibility for the harm and hence their emotional responses (Baumeister et al., 1990; Lemay Jr et al., 2012; León et al., 2009). Future studies are needed to distinguish these psychological components.

To conclude, by manipulating the victim's attitudes towards the transgressor's wrongdoings, the current study uncovered the psychological and neural bases underlying the transgressor's reactive guilt and anger, as well as the differential neural representations underlying reactive guilt and non-reactive guilt. These findings demonstrate the mutuality of social emotions and highlight the importance of understanding social emotions from the perspective of interpersonal interaction. Our approach of combining interactive game with multivariate pattern analysis opens a venue for investigating the neurocognitive bases of how

other human social emotions (e.g., gratitude, shame and indebtedness) and behaviors arise and evolve during social interactions.

Declaration of Competing Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Credit authorship contribution statement

Xiaoxue Gao: Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Funding acquisition, Writing – original draft, Writing – review & editing. **Hongbo Yu:** Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Funding acquisition, Writing – original draft, Writing – review & editing. **Lu Peng:** Conceptualization, Methodology, Investigation, Formal analysis. **Xiaoliang Gong:** Resources. **Yang Xiang:** Resources. **Changjun Jiang:** Resources. **Xiaolin Zhou:** Supervision, Funding acquisition, Writing – original draft, Writing – review & editing.

Acknowledgments

This work was supported by National Natural Science Foundation of China (31900798, 31630034, 71942001) and China Postdoctoral Science Foundation (2019M650008). The authors thank Ms. Zhewen He and Ms. Siyi Gong for the preparation of the manuscript, and three anonymous reviewers for their advice on the write-up of this article.

Data and code availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the *Supplementary Materials*. Original materials are available on OSF (https://osf.io/mj42y/?view_only=b7791dbf124f4d209757e68b2c340e03).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118631.

Reference

- Adams, J.S., 1965. Inequity in social exchange. In: Berkowitz, L. (Ed.), *Advances in Experimental Social Psychology*. Academic Press, pp. 267–299.
- Anderson, R.A., Kamtekar, R., Nichols, S., Pizarro, D.A., 2021. False positive" emotions, responsibility, and moral character. *Cognition* 214.
- Antony, J.W., Hartshorne, T.H., Pomeroy, K., Gureckis, T.M., Hasson, U., McDougle, S.D., Norman, K.A., 2021. Behavioral, physiological, and neural signatures of surprise during naturalistic sports viewing. *Neuron* 109, 377–390 e377.
- Ashar, Y.K., Andrews-Hanna, J.R., Dimidjian, S., Wager, T.D., 2017. Empathic care and distress: predictive brain markers and dissociable brain systems. *Neuron* 94, 1263–1273 e1264.
- Basile, B., Mancini, F., Macaluso, E., Caltagirone, C., Frackowiak, R.S., Bozzali, M., 2011. Deontological and altruistic guilt: evidence for distinct neurobiological substrates. *Hum. Brain Mapp.* 32, 229–239.
- Bates, D., Mächler, M., Bolker, B., Walker, S., 2014. Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.
- Baumeister, R.F., Stillwell, A., Wotman, S.R., 1990. Victim and perpetrator accounts of interpersonal conflict: Autobiographical narratives about anger. *J. Pers. Soc. Psychol.* 59, 994–1005.
- Baumeister, R.F., Stillwell, A.M., Heatherton, T.F., 1994. Guilt: an interpersonal approach. *Psychol. Bull.* 115, 243–267.
- Berndsen, M., McGarty, C., 2010. The impact of magnitude of harm and perceived difficulty of making reparations on group-based guilt and reparation towards victims of historical harm. *Eur. J. Soc. Psychol.* 40, 500–513.
- Blair, R.J.R., 2012. Considering anger from a cognitive neuroscience perspective. *Wiley Interdisc. Rev.-Cognit. Sci.* 3, 65–74.
- Browne, M.W., Cudeck, R., 1992. Alternative ways of assessing model fit. *Sociol. Methods Res.* 21, 230–258.
- Calhoun, V.D., Wager, T.D., Krishnan, A., Rosch, K.S., Seymour, K.E., Nebel, M.B., Mostofsky, S.H., Nyalakanai, P., Kiehl, K., 2017. The impact of T1 versus EPI spatial normalization templates for fMRI data analyses. *Hum. Brain Mapp.* 38, 5331–5342.

- Cavanagh, J.F., Shackman, A.J., 2015. Frontal midline theta reflects anxiety and cognitive control: meta-analytic evidence. *J. Physiol. Paris* 109, 3–15.
- Chang, L.J., Gianaros, P.J., Manuck, S.B., Krishnan, A., Wager, T.D., 2015. A sensitive and specific neural signature for picture-induced negative affect. *PLoS Biol.* 13, e1002180.
- Chang, L.J., Jolly, E., 2017. Emotions as computational signals of goal error. *Nat. Emotion* 343–348.
- Chang, L.J., Jolly, E., Cheong, J.H., Rapuano, K.M., Greenstein, N., Chen, P.A., Manning, J.R., 2021. Endogenous variation in ventromedial prefrontal cortex state dynamics during naturalistic viewing reflects affective experience. *Sci. Adv.* 7.
- Chang, L.J., Smith, A., 2015. Social emotions and psychological games. *Curr. Opin. Behav. Sci.* 5, 133–140.
- Chang, L.J., Smith, A., Dufwenberg, M., Sanfey, A.G., 2011. Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron* 70, 560–572.
- Ci, J., 2006. *The Two Faces of Justice*. Harvard University Press.
- Craddock, R.C., James, G.A., Holtzheimer 3rd, P.E., Hu, X.P., Mayberg, H.S., 2012. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Hum. Brain Mapp.* 33, 1914–1928.
- Darwall, S., 2013. *Morality, Authority, and Law: Essays in Second-Personal Ethics I*. OUP Oxford.
- De Hooge, I.E., Nelissen, R., Breugelmans, S.M., Zeelenberg, M., 2011. What is moral about guilt? Acting “prosocially” at the disadvantage of others. *J. Pers. Soc. Psychol.* 100, 462.
- de la Vega, A., Chang, L.J., Banich, M.T., Wager, T.D., Yarkoni, T., 2016. Large-scale meta-analysis of human medial frontal cortex reveals tripartite functional organization. *J. Neurosci.* 36, 6553–6562.
- Decety, J., Yoder, K.J., 2017. The emerging social neuroscience of justice motivation. *Trends Cogn. Sci.* 21, 6–14.
- Denson, T.F., Pedersen, W.C., Ronquillo, J., Nandy, A.S., 2009. The angry brain: neural correlates of anger, angry rumination, and aggressive personality. *J. Cogn. Neurosci.* 21, 734–744.
- Donnerstein, E., Hatfield, E., 1982. Aggression and Inequity. In: Greenberg, J., Cohen, R.L. (Eds.), *Equity and Justice in Social Behavior*. Academic Press, pp. 309–336.
- Eisenbarth, H., Chang, L.J., Wager, T.D., 2016. Multivariate brain prediction of heart rate and skin conductance responses to social threat. *J. Neurosci.* 36, 11987–11998.
- Exline, J.J., Worthington, E.L., Hill, P., McCullough, M.E., 2003. Forgiveness and Justice: a Research Agenda for Social and Personality Psychology. *Pers. Soc. Psychol. Rev.* 7, 337–348.
- Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. *Q. J. Econ.* 114, 817–868.
- Friedman, J., Hastie, T., Tibshirani, R., 2001. *The elements of statistical learning*. Springer series in statistics, New York.
- Friston, K.J., Glaser, D.E., Henson, R.N., Kiebel, S., Phillips, C., Ashburner, J., 2002. Classical and Bayesian inference in neuroimaging: applications. *Neuroimage* 16, 484–512.
- Gao, X., Yu, H., Saez, I., Blue, P., Zhu, L., Hsu, M., Zhou, X., 2018. Distinguishing neural correlates of context-dependent advantageous- and disadvantageous-inequity aversion. *PNAS* 115, E7680–E7689.
- Gassin, A., Elizabeth, 1998. Receiving Forgiveness as Moral Education: a theoretical analysis and initial empirical investigation. *J. Moral Educ.* 27, 71–87.
- Green, P., MacLeod, C.J., 2016. SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods Ecol. Evol.* 7, 493–498.
- Gross, J.J., 2002. Emotion regulation: affective, cognitive, and social consequences. *Psychophysiology* 39, 281–291.
- Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., Tatham, R., 2006. *Multivariate data analysis*. Uppersaddle River.
- Hari, R., Henriksson, L., Malinen, S., Parkkonen, L., 2015. Centrality of social interaction in human brain function. *Neuron* 88, 181–193.
- Heilbronner, S.R., Hayden, B.Y., 2016. Dorsal Anterior Cingulate Cortex: a Bottom-Up View. *Annu. Rev. Neurosci.* 39, 149–170.
- Helm, B., 2017. *Communities of Respect: Grounding Responsibility, Authority, and Dignity*. Oxford University Press.
- Hendriks, M.H.A., Daniels, N., Pegado, F., Op de Beeck, H.P., 2017. The effect of spatial smoothing on representational similarity in a simple motor paradigm. *Front. Neurol.* 8, 222–222.
- Hoffman, M.L., 2000. *Empathy and moral development: Implications for caring and justice*. Cambridge University Press, New York, NY, US.
- Homans, G.C., 1974. *Social behavior: Its elementary forms*, Revised ed. Harcourt Brace Jovanovich, Oxford, England.
- Hu, L., 1995. Evaluating model fit. In: *Structural equation modelling: concepts, issues and applications*, pp. 76–99.
- Inui, K., Tran, T.D., Hoshiyama, M., Kakigi, R., 2002. Preferential stimulation of Adelta fibers by intra-epidermal needle electrode in humans. *Pain* 96, 247–252.
- Isoda, M., Noritake, A., 2013. What makes the dorsomedial frontal cortex active during reading the mental states of others? *Front. Neurosci.* 7, 232.
- Jennings, D.J., Worthington, E.L., Van Tongeren, D.R., Hook, J.N., Davis, D.E., Gartner, A.L., Greer, C.L., Mosher, D.K., 2016. The transgressor’s response to denied forgiveness. *J. Psychol. Theor.* 44, 16–27.
- Johnstone, T., Ores Walsh, K.S., Greischar, L.L., Alexander, A.L., Fox, A.S., Davidson, R.J., Oakes, T.R., 2006. Motion correction and the use of motion covariates in multiple-subject fMRI analysis. *Hum. Brain Mapp.* 27, 779–788.
- Jolly, E., Chang, L.J., 2019. The flatland fallacy: moving beyond low-dimensional thinking. *Top. Cognit. Sci.* 11, 433–454.
- Kédia, G., Berthoz, S., Wessa, M., Hilton, D., Martinot, J.L., 2008. An agent harms a victim: a functional magnetic resonance imaging study on specific moral emotions. *J. Cogn. Neurosci.* 20, 1788–1798.
- Klimecki, O.M., Sander, D., Vuilleumier, P., 2018. Distinct brain areas involved in anger versus punishment during social interactions. *Sci. Rep.* 8, 10556.
- Koban, L., Corradi-Dell’Acqua, C., Vuilleumier, P., 2013. Integration of error agency and representation of others’ pain in the anterior insula. *J. Cogn. Neurosci.* 25, 258–272.
- Krishnan, A., Woo, C.W., Chang, L.J., Ruzic, L., Gu, X., López-Solà, M., Jackson, P.L., Pujol, J., Fan, J., Wager, T.D., 2016. Somatic and vicarious pain are represented by dissociable multivariate brain patterns. *Elife* 5.
- Kubany, E.S., Watson, S.B., 2003. Guilt: elaboration of a multidimensional model. *Psychol. Record* 53, 51–90.
- Larsen, R.J., Fredrickson, B.L., 1999. *Measurement issues in emotion research. Well-being: The foundations of hedonic psychology*. Russell Sage Foundation.
- Lemay Jr., E.P., Overall, N.C., Clark, M.S., 2012. Experiences and interpersonal consequences of hurt feelings and anger. *J. Pers. Soc. Psychol.* 103, 982–1006.
- León, I., Hernández, J.A., Rodríguez, S., Vila, J., 2009. When head is tempered by heart: heart rate variability modulates perception of other-blame reducing anger. *Motiv. Emot.* 33, 1–9.
- Lewandowsky, S., Farrell, S., 2010. *Computational modeling in cognition: Principles and practice*. Sage.
- Li, Z., Yu, H., Zhou, Y., Kalenscher, T., Zhou, X., 2020. Guilty by association: how group-based (collective) guilt arises in the brain. *Neuroimage* 209, 116488.
- Liu, G., Cui, Z., Yu, H., Rotshtein, P., Sui, J., 2020. Neural responses to intention and benefit appraisal are critical in distinguishing gratitude and joy. *Sci. Rep.* 10, 7864.
- Lockwood, P.L., Wittmann, M.K., 2018. Ventral anterior cingulate cortex in social decision-making. *Neurosci. Biobehav. Rev.* 92, 187–191.
- Louie, K., De Martino, B., 2013. *The Neurobiology Of Context-Dependent Valuation And Choice. Neuroeconomics: Decision Making and the Brain*, ed. PW a. FE Glimcher, pp. 455–476.
- Malle, B.F., 2021. Moral judgments. *Annu. Rev. Psychol.* 72, 293–318.
- Mason, E., 2019. *Ways to be Blameworthy*. Oxford University Press, Oxford.
- Mcgraw, K., 1987. Guilt following transgression: an attribution of responsibility approach. *J. Pers. Soc. Psychol.* 53, 247–256.
- McNulty, J.K., 2010. Forgiveness increases the likelihood of subsequent partner transgressions in marriage. *J. Fam. Psychol.* 24, 787–790.
- McNulty, J.K., 2011. The Dark Side of Forgiveness: the tendency to forgive predicts continued psychological and physical aggression in marriage. *Pers. Soc. Psychol. Bull.* 37, 770–783.
- Miceli, M., Castelfranchi, C., 2014. *Expectancy and Emotion*. OUP Oxford.
- Morey, R.A., McCarthy, G., Selgrade, E.S., Seth, S., Nasser, J.D., LaBar, K.S., 2012. Neural systems for guilt from actions affecting self versus others. *Neuroimage* 60, 683–692.
- Nachev, P., Kennard, C., Husain, M., 2008. Functional role of the supplementary and pre-supplementary motor areas. *Nat. Rev. Neurosci.* 9, 856–869.
- Nisbett, R.E., Wilson, T.D., 1977. Telling more than we can know: Verbal reports on mental processes. *Psychol. Rev.* 84, 231–259.
- Olsson, A., FeldmanHall, O., Haaker, J., Hensler, T., 2018. Social regulation of survival circuits through learning. *Curr. Opin. Behav. Sci.* 24, 161–167.
- Olsson, A., Knapska, E., Lindström, B., 2020. The neural and computational systems of social learning. *Nat. Rev. Neurosci.* 21, 197–212.
- Op de Beeck, H.P., 2010. Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? *Neuroimage* 49, 1943–1948.
- Parkinson, B., Illingworth, S., 2009. Guilt in response to blame from others. *Cognit. Emotion* 23, 1589–1614.
- Penny, W., Kiebel, S., Friston, K., 2003. Variational Bayesian inference for fMRI time series. *Neuroimage* 19, 727–741.
- Preacher, K.J., Hayes, A.F., 2008. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav. Res. Methods* 40, 879–891.
- Renetzky, H.C., 2015. *Guilt and moral compensation: relational or self-presentational?*, Psychology honors projects. Macalester College.
- Rosseel, Y., 2012. lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36.
- Ruff, C.C., Fehr, E., 2014. The neurobiology of rewards and values in social decision making. *Nat. Rev. Neurobiol.* 15, 549–562.
- Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., Vogeley, K., 2013. Toward a second-person neuroscience. *Behav. Brain Sci.* 36, 393–414.
- Schurz, M., Perner, J., 2015. An evaluation of neurocognitive models of theory of mind. *Front. Psychol.* 6, 1610–1610.
- Schurz, M., Radua, J., Aichhorn, M., Richlan, F., Perner, J., 2014. Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neurosci. Biobehav. Rev.* 42, 9–34.
- Seo, H., Lee, D., 2012. Neural basis of learning and preference during social decision-making. *Curr. Opin. Neurobiol.* 22, 990–995.
- Shackman, A.J., Salomons, T.V., Slagter, H.A., Fox, A.S., Winter, J.J., Davidson, R.J., 2011. The integration of negative affect, pain and cognitive control in the cingulate cortex. *Nat. Rev. Neurosci.* 12, 154–167.
- Siegel, J.Z., Mathys, C., Rutledge, R.B., Crockett, M.J., 2018. Beliefs about bad people are volatile. *Nat. Hum. Behav.* 2, 750–756.
- Strawson, P.F., 1974. *Freedom and resentment, and other essays (Vol. 595)*. Egmont Books.
- Szyczer, 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.
- Takahashi, H., Yahata, N., Koeda, M., Matsuda, T., Asai, K., Okubo, Y., 2004. Brain activation associated with evaluative processes of guilt and embarrassment: an fMRI study. *Neuroimage* 23, 967–974.
- Tangney, J.P., Dearing, R.L., Salovey, P., 2002. *Shame and guilt*. Guilford Press.
- Tangney, J.P., Stuewig, J., Mashek, D.J., 2007. Moral emotions and moral behavior. *Annu. Rev. Psychol.* 58, 345–372.

- Tobias, S., Carlson, J.E., 1969. Brief report: Bartlett's test of sphericity and chance findings in factor analysis. *Multivar. Behav. Res.* 4, 375–377.
- Vaish, A., Hepach, R., 2019. The development of prosocial emotions. *Emotion Rev.* 1754073919885014.
- van Baar, J.M., Chang, L.J., Sanfey, A.G., 2019. The computational and neural substrates of moral strategies in social decision-making. *Nat. Commun.* 10, 1–14.
- Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *Neuroimage* 145, 166–179.
- Wager, T.D., Atlas, L.Y., Lindquist, M.A., Roy, M., Woo, C.-W., Kross, E., 2013. An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med.* 368, 1388–1397.
- Wager, T.D., Kang, J., Johnson, T.D., Nichols, T.E., Satpute, A.B., Barrett, L.F., 2015. A Bayesian model of category-specific emotional brain responses. *PLoS Comput. Biol.* 11, e1004066.
- Wagner, U., N'Diaye, K., Ethofer, T., Vuilleumier, P., 2011. Guilt-specific processing in the prefrontal cortex. *Cereb. Cortex* 21, 2461–2470.
- Wallace, H.M., Exline, J.J., Baumeister, R.F., 2008. Interpersonal consequences of forgiveness: does forgiveness deter or encourage repeat offenses? *J. Exp. Soc. Psychol.* 44, 0–460.
- Walster, E., Walster, G.W., Berscheid, E., 1978. *Equity: theory and research*.
- West, S.G., Taylor, A.B., Wu, W., 2012. Model Fit And Model Selection In Structural Equation Modeling. *Handbook of Structural Equation Modeling*. The Guilford Press, New York, NY, US, pp. 209–231.
- Wilke, M., 2012. An alternative approach towards assessing and accounting for individual motion in fMRI timeseries. *Neuroimage* 59, 2062–2072.
- Will, G.-J., Rutledge, R.B., Moutoussis, M., Dolan, R.J., 2017. Neural and computational processes underlying dynamic changes in self-esteem. *Elife* 6, e28098.
- Wittmann, M.K., Kolling, N., Faber, N.S., Scholl, J., Nelissen, N., Rushworth, M.F., 2016. Self-other merge in the frontal cortex during cooperation and competition. *Neuron* 91, 482–493.
- Woo, C.-W., Koban, L., Kross, E., Lindquist, M.A., Banich, M.T., Ruzic, L., Andrews-Hanna, J.R., Wager, T.D., 2014. Separate neural representations for physical pain and social rejection. *Nat. Commun.* 5, 5380.
- 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.
- Young, L., Saxe, R., 2009. Innocent intentions: a correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia* 47, 2065–2072.
- 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.* 2944–2917.
- Yu, H., Hu, J., Hu, L., Zhou, X., 2014. The voice of conscience: neural bases of interpersonal guilt and compensation. *Soc. Cognit. Affect. Neurosci.* 9, 1150–1158.
- Yu, H., Koban, L., Chang, L.J., Wagner, U., Krishnan, A., Vuilleumier, P., Zhou, X., Wager, T.D., 2020a. A generalizable multivariate brain pattern for interpersonal guilt. *Cereb. Cortex* 30, 3558–3572.
- Yu, H., Koban, L., Crockett, M.J., Zhou, X., Wager, T.D., 2020b. Toward a brain-based bio-marker of guilt. *Neurosci. Insights* 15 263310552095763.
- Zechmeister, J.S., Romero, C., 2002. Victim and offender accounts of interpersonal conflict: autobiographical narratives of forgiveness and unforgiveness. *J. Pers. Soc. Psychol.* 82, 675–686.
- Zhu, R., Feng, C., Zhang, S., Mai, X., Liu, C., 2018. Differentiating guilt and shame in an interpersonal context with univariate activation and multivariate pattern analyses. *Neuroimage* 186, 476–486.
- Zhu, X., Wang, K., Cao, A., Zhang, Y., Qiu, J., 2020. Personality traits and negative affect mediate the relationship between cortical thickness of superior frontal cortex and aggressive behavior. *Neurosci. Lett.* 718, 134728.