NeuroImage 244 (2021) 118631

Contents lists available at ScienceDirect

NeuroImage

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

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

a r t i c l e i n f o

Keywords:
Guilt
Anger
Expectation violation
Ventral striatum
fMRI

a b s t r a c t

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-

* Corresponding authors.
E-mail addresses: gxx114455@gmail.com (X. Gao), hongbo.yu@psych.ucsb.edu (H. Yu),

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 e ects, 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

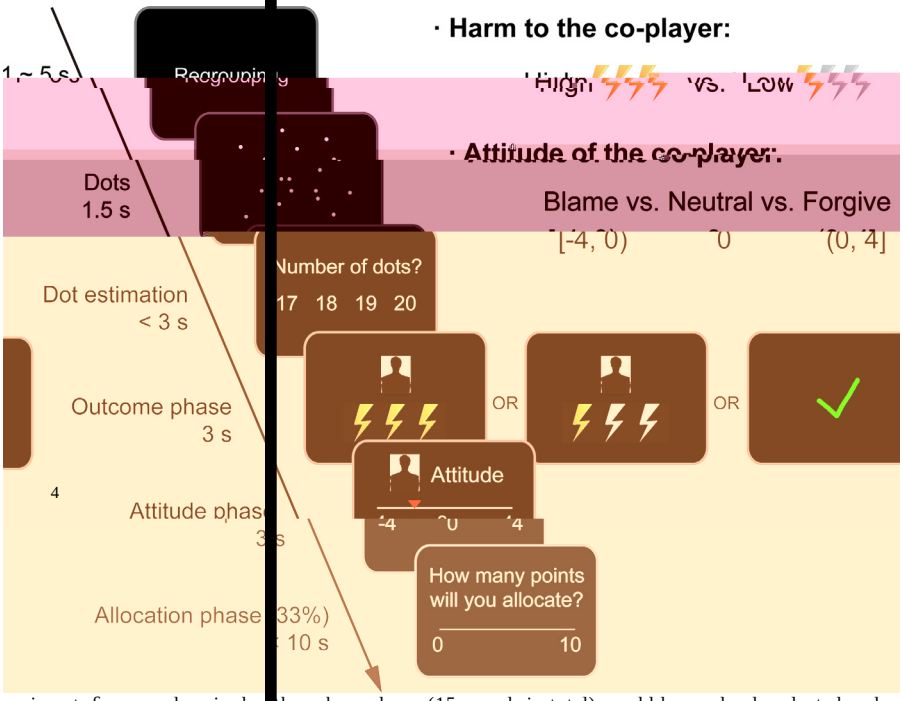


Fig. 1. Procedures of the interactive game. In each round, after being paired with a same sex anonymous co-player, the participant would see a picture of dots for 1.5 s and estimate the number of dots quickly by choosing one of the four numbers presented on the screen within 3 s (i.e., Dot estimation). After that the correctness of the estimation was revealed. If the estimation was correct, the current trial was terminated and the game entered the next round; otherwise, the co-player (the victim) in the current round would receive a pain stimulation, with either high or low intensity, randomly determined by the computer program (i.e., *Harm to the co-player*: High harm *vs.* Low harm; Outcome phase, 3 s). Then the participant would be presented with the co-player's attitude on the scale from -4 to 4 (i.e., Attitude phase, 3 s). Positive value stands for forgiveness (Forgive condition), negative value for blame (Blame condition), and zero for neutral attitude, i.e., neither blame nor forgiveness (Neutral condition). In one-third of the trials, at the end of each trial, the participant was asked to divide 10 points (1 point = 2 Yuan; 20 Yuan ≈ 3.1 USD) between him/herself and the co-player paired in this trial (i.e., Allocation phase, < 10 s), with the knowledge that the co-player was not aware of this procedure. In the remaining trials, the current trial terminated after the presentation of the co-player's attitude. 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. 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

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 di erent 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 e ective 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 di erence 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 di erence 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 e ect) for ratings of guilt, ratings of anger and amounts of allocation respectively. By-participant random slopes for each fixed e ect 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 coe cients.

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 e ect 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 e ects. 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 e ects 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 e ects. 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 e ect 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 e ects 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 coe cient for each path of the model was labeled on corresponding figures. For each path of the model, the values of *c* indicated the e ect of the corresponding independent variable on dependent variable before controlling for the e ect of mediator; the values of *c'* indicated the e ect of the corresponding independent variable on dependent variable after controlling for the e ect 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 e ect was included in this LMM. Results showed that neither the main e ect 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, a ording 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 mm^3 isotropic voxel MPRAGE) was acquired after functional imaging. Imaging processing was conducted following the standard preprocessing 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 di erence 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 cuto 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 Beeck, 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

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) × 3 (Attitude of the co-player: Blame *vs.* Forgive *vs.* Neutral) interaction e ect 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 e ect 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 di erential 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 di erences 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_Responsisble", "Both_Responsible", and "Self_Responsible". 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_Responsible) *vs.* Low (Both_Responsible) 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_Responsible) *vs.* Low (Both_Responsible) 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_Responsible) *vs.* Low (Both_Responsible) 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_Responsible, Both_Responsible, Co-player_Responsible) 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_Responsible) *vs.* Low (Both_Responsible) 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_Responsible, Both_Responsible, Co-player_Responsible) 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 e cient 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 shu ed 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 di erence between classification accuracies for reactive guilt and reactive anger for each parcel. Permutation *ps* for accuracy di erences were computed for the probability of the estimated accuracy di erences in permutations being greater than the observed accuracy di erences. 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 di erent 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 di erential 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 di erent 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 e ect respectively, with by-participant random slopes for the fixed e ect 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 di erential 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 di erential 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 di erential neural representations or just resulted from a limited power of analyses. To our knowledge, there is no standard way of calculating e ect 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-

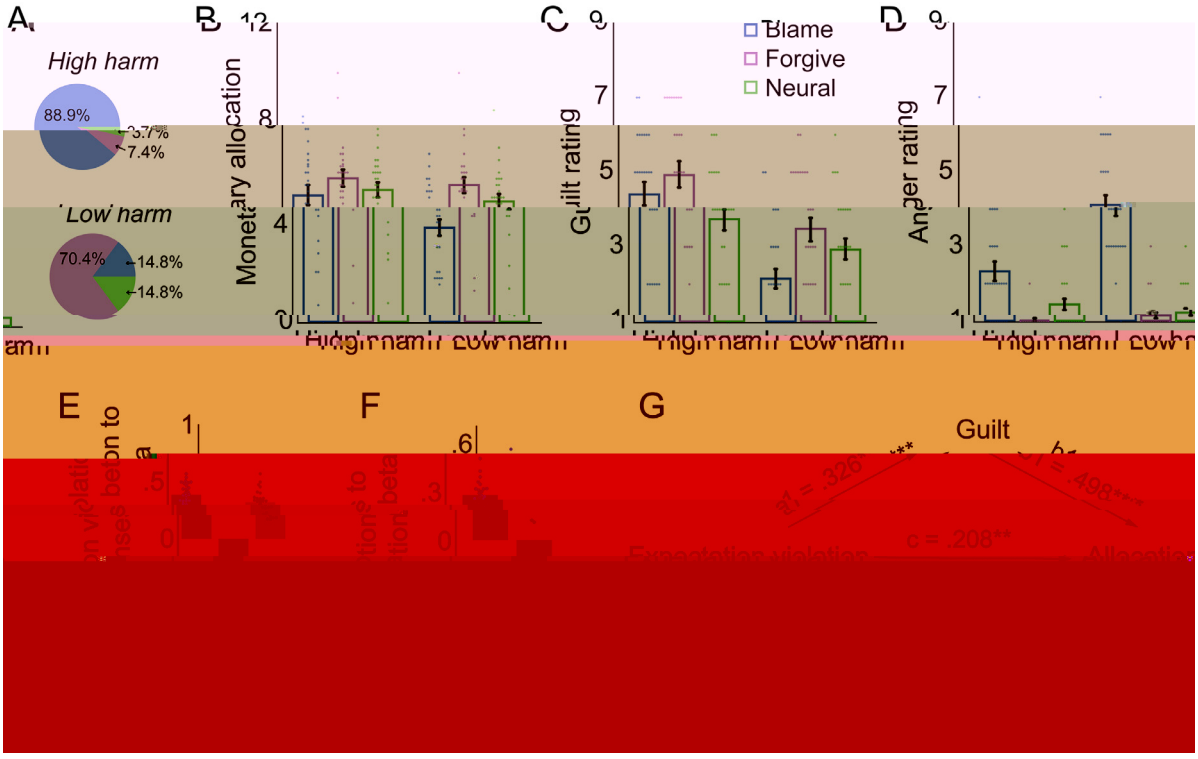


Fig. 2. Behavioral results. (A - D) Participants' pre-experiment prediction of the co-player's attitudes in High and Low harm conditions (A), amount of monetary allocation (B), post-experiment rating of guilt (C), and post-experiment rating of anger (D) in the six conditions. The scale of monetary allocation 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

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 e ects were observed on ratings of both guilt ($F_{2, 52}$ = 10.77, p < 0.001, $\eta^2_{partial}$ = 0.29, power = 0.97; Fig. 2C) and anger ($F_{2, 52}$ = 30.23, p < 0.001, $\eta^2_{partial}$ = 0.54, 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, *Cohen's d* = 0.91; anger: t_{26} = 10.79, p < 0.001, *Cohen's d* = 2.75) and Neutral condition (guilt: t_{26} = -2.95, p = 0.020, *Cohen's d* = 0.56; anger: t_{26} = 10.35, p < 0.001, *Cohen's d* = 2.66); such di erences were significantly reduced in High harm conditions (guilt Blame *vs.* Forgive: t_{26} = -1.25, p = 0.667, *Cohen's d* = 0.29; guilt Blame *vs.* Neutral: t_{26} = 1.84, p = 0.231, *Cohen's d* = 0.41; anger Blame *vs.* Forgive: t_{26} = 4.78, p < 0.001, *Cohen's d* = 1.30; anger Blame *vs.* Neutral: t_{26} = 3.25, p = 0.010, *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, *Cohen's d* = 0.69) and the feeling of anger was lower (t_{26} = -2.51, p = 0.056, *Cohen's d* = 0.67) in Forgive condition; such di erences were significantly reduced in Low harm conditions (guilt: t_{26} = 2.20, p = 0.110, *Cohen's d* = 0.36; anger: t_{26} = -0.57, p = 1.000, *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.

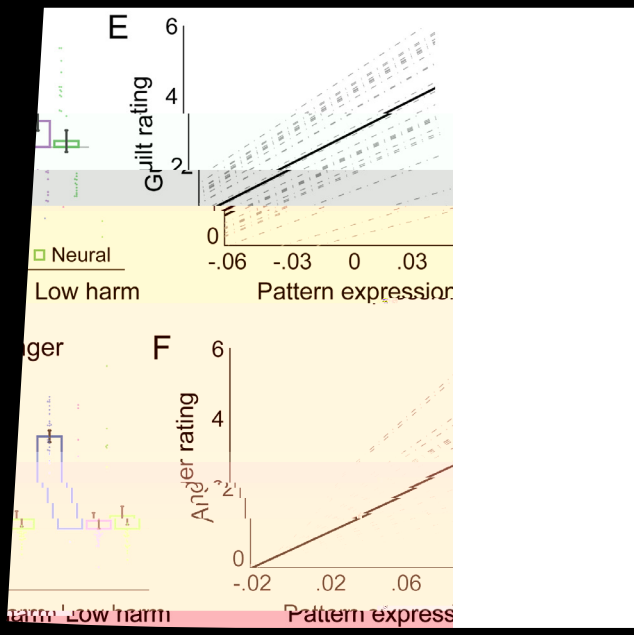
To directly examine the relationships between expectation violation and feelings of guilt and anger, we calculated expectation violation of co-player's attitude in each condition according to participants' predicted attitudes of co-player before the task. Consistent with our hypothesis, results of linear mixed modeling (Fig. 2E) revealed that while self-reported guilt was positively correlated with the level of expectation violation (β = 0.37 ± 0.07 (SE), t = 5.84, p < 0.001, power = 1.00), self-reported anger was negatively correlated with the level of expectation violation (β = -0.59 ± 0.07, t = -8.44, p < 0.001, power = 1.00). Moreover, the amount of allocation was positively with the level of expectation violation (β = 0.32 ± 0.05, t = 6.00, p < 0.001, 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 either 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 (β = 0.28 ± 0.06, t = 4.86, p < 0.001, power = 1.00; Fig. 2F), ratings of anger contributed negatively to allocation (β = -0.23 ± 0.07, t = -3.44, p = 0.002, power = 0.99; Fig. 2F).

Finally, we examined whether feelings of guilt and anger mediated the e ect 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 e ect of this model was significant, with normalized coe cient of overall mediating e ect = 0.238, p < 0.001, c = 0.208, p = 0.006, c' = -0.030, p = 0.724, a complete mediation. The normalized coe cients of the mediating e ect of guilt and anger were 0.162, p < 0.001 and 0.076, p = 0.115, respectively. No significant di erence was observed between the mediating e ect 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 χ^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



s. Error bars represent *SEs*.

0.001, p_{FDR} = 0.002, power =1.00; Fig. 3F and Fig. S3B). indicated that the classifiers (i.e., multivariate brain activations) of reactive guilt and anger accurately reflected neural each emotion in Attitude phase. Moreover, cross-emotion revealed that the pattern for guilt was not able to distinn High and Low anger

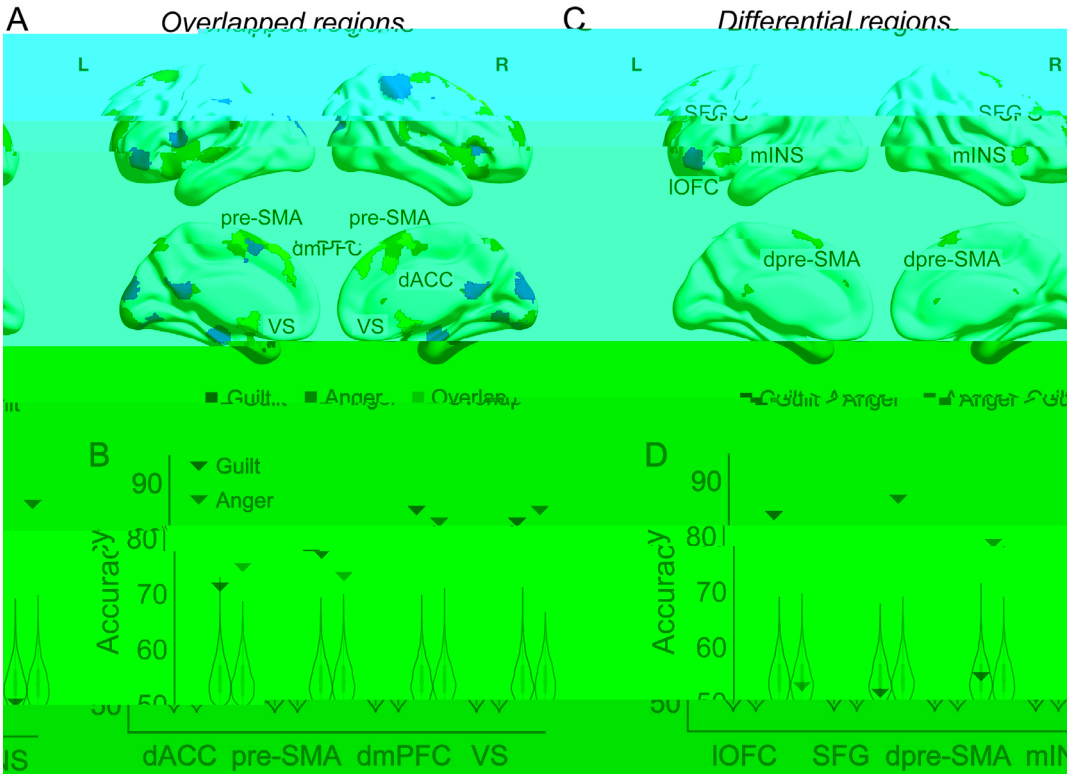


Fig. 4. Shared and differential neural local classifiers for reactive guilt and reactive anger after receiving the co-player's attitude feedbacks. (A) Local classifiers that significantly discriminating High *vs.* Low guilt conditions (blue) and High *vs.* Low anger conditions (red) in Attitude phase, 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 di erential 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 di erential classification abilities for reactive guilt and anger, and observed brain regions that showed di erential sensitivity to reactive guilt and anger. Results showed that while left lateral orbitofrontal cortex (lOFC) 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 di ered 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. Di erent 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 correlated 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 was 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

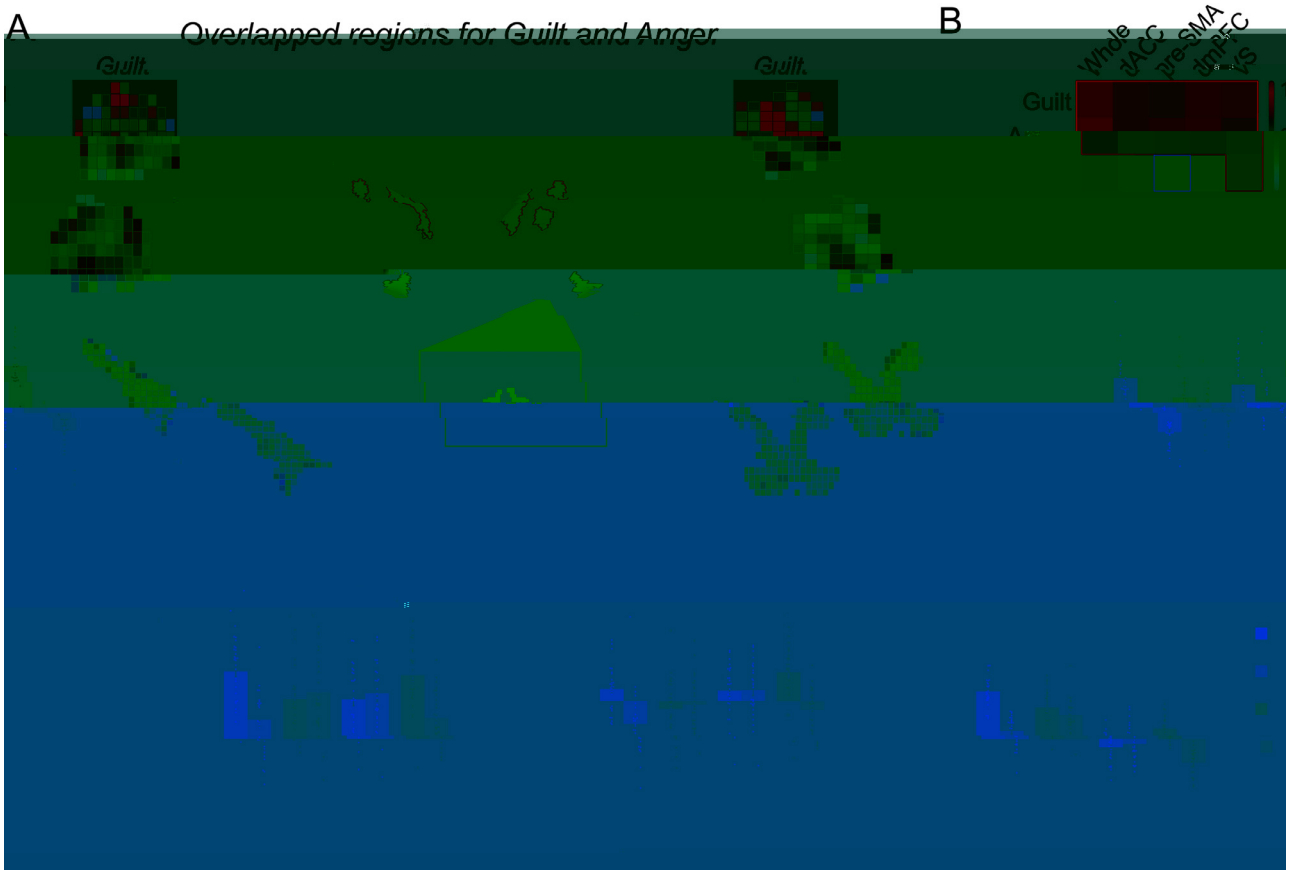


Fig. 5. Shared and differential neural representations for overlapping regions identified for reactive guilt and anger. (A) Local patterns for overlapping regions (i.e., dACC, pre-SMA, dmPFC and VS) for processing reactive guilt and anger. (B) Pattern expression values of reactive guilt and anger in 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 di erential 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, a ect, 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. Di erential whole-brain patterns for guilt before and after receiving the co-player's attitude

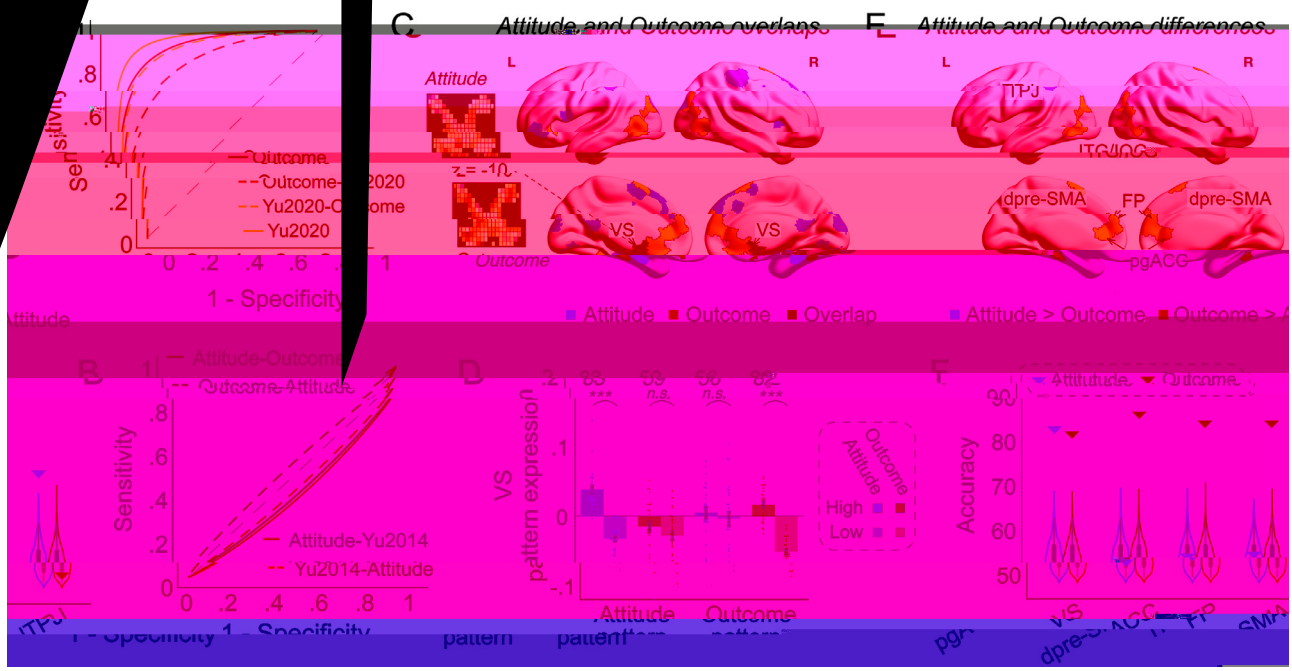


Fig. 6. Shared and differential neural representations for reactive guilt in Attitude phase and non-reactive guilt in Outcome phase. (A) ROCs for the two-choice forced-alternative accuracies for within-study and cross-study classifications using the data of Outcome phase in the current study and of Yu 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 form 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 di erential 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 w t yet available, sho eflect 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 001; Wager et al., 2013) to develop a whole-brain classifier to d minate High harm *vs.* Low harm conditions in Outcome phase. This classifier yielded an average classification accuracy of 85.2% (± 7.0% SE), $p < 0.001$, $p_{FDR} = 0.004$ (Fig. 6A). Results of cross-study classif (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% (± 15.3%),

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 di er 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 ± 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 indicate 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 e ects 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 di erential 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 ±

the victim's attitudes both in terms of brain regions involved and in terms of neural representations. While reactive guilt recruited lOFC, reactive anger recruited mINS and SFG. These results are consistent with previous studies showing the involvement of lOFC 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 threF1 1 Tfn7.9701 0 0 7.9701 56.78481 5617.6143 625.4191 Tmn.0003 Tcn(a101 0 0 7.9701 236.3429 635.886 Tm228 Tmn-.0

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 di erences 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 e ective 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 di erential 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,

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 a ect. 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 a ective 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:
g6,

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 o enses? 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 mergence 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. A ective 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. A ect. 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 o ender accounts of interpersonal conflict: autobiographic