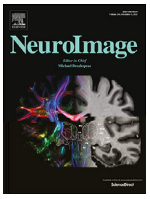
NeuroImage 244 (2021) 118631

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

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 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 di erences 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.

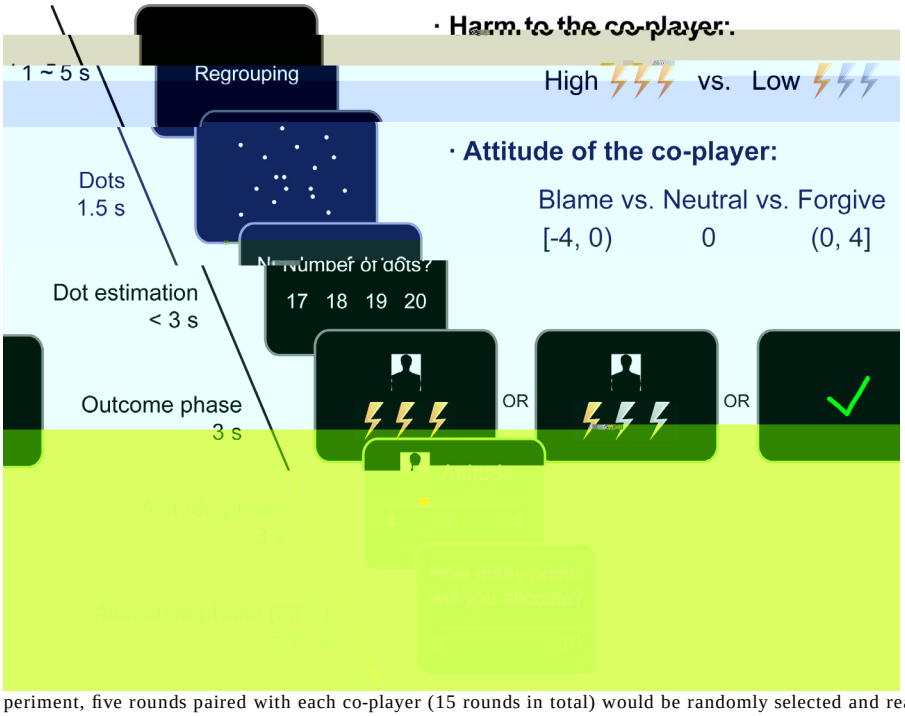


Fig. 1. Procedures of the interactive game. In each round, after being paired with a same sex anonymous co-player, the participant would see a picture of dots for 1.5 s and estimate the number of dots quickly by choosing one of the four numbers presented on the screen within 3 s (i.e., Dot estimation). 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

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 -2 (‘higher willingness to /F2 1 Tf 7.99461 0 0 6.3761 213.7168 667.269 Tmn0 Tcn( )Tjn/F14Tjn57.9f7.99461 0 701 49.863 593.87

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):

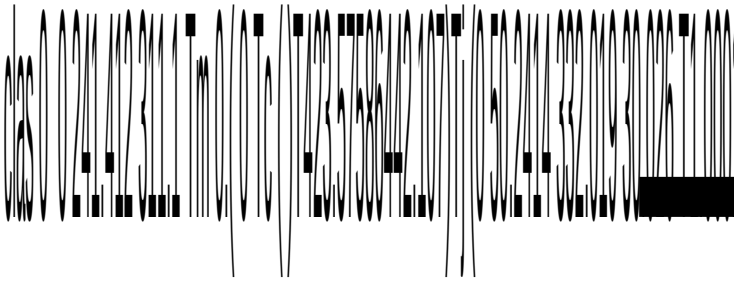
High_harm_Blame_Outcome, High_harm_Forgive_Outcome, High_harm_Neutral_Outcome, Low_harm_Blame_Outcome, Low_harm_Forgive_Outcome, Low_harm_Neutral_Outcome, High_harm_Blame_Attitude, High_harm_Forgive_Attitude, High_harm_Neutral_Attitude, Low_harm_Blame_Attitude, Low_harm_Forgive_Attitude, Low_harm_Neutral_Attitude (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

reactive guilt pattern



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 t

variable variable variable variable

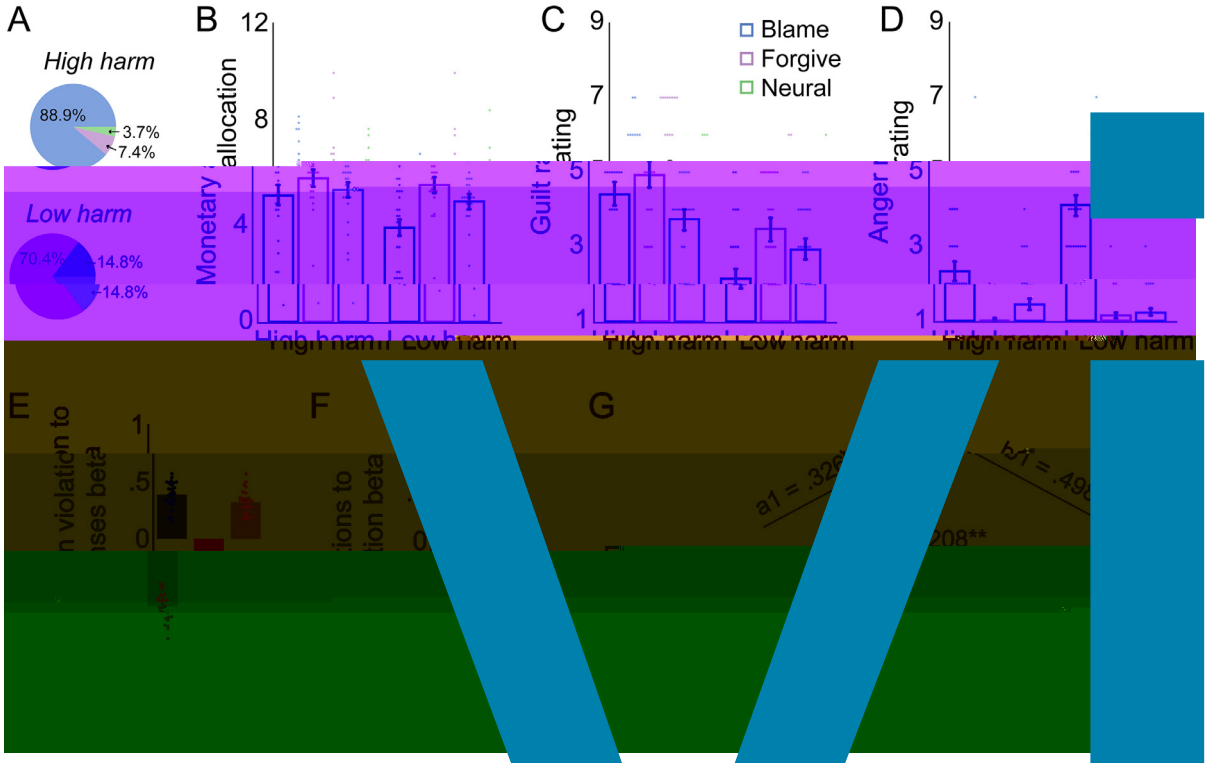


Fig. 2. Behavioral results. (A - D) Participants' pre-experiment predict -player's at gh and Low harm conditions (f monetary allocation (B), post-experiment rating of guilt (C), and post-experiment rat (D) in the ns. The scale of monetary alloca om 0 to 10 points (1 point = 2 Yuan). The scales of guilt and anger ratings ranged fro ll') to 7 g'). (E) The regression betas indi ntributions of expectation violation to the amount of monetary allocation, and post-exp ngs o nger. (F) The regression betas indi ntributions of post-experiment ratings of guilt and anger to the amount of monetary allo Post ratings of guilt and anger significa d the e ect of expectation violation on the amount of monetary allocation. *n.s.* $p > 0.05$, * p < .001. Error bars represent *SEs*.

ipants ranging from 0 to 200 using "simr" package in R (Green and MacLeod,

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	

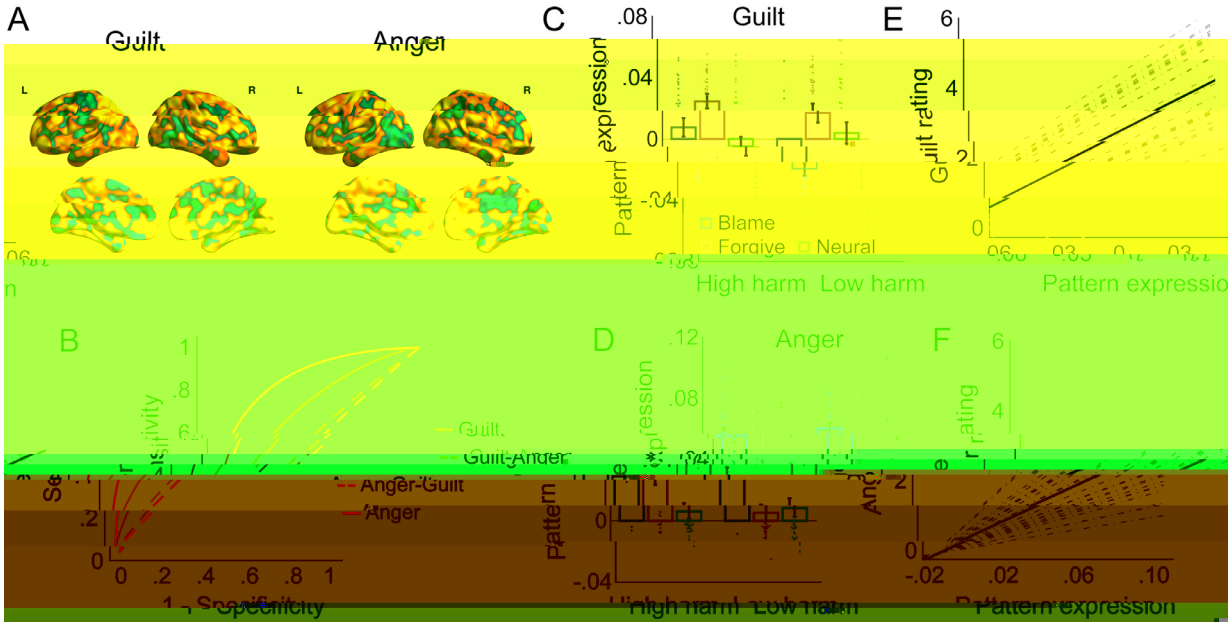


Fig. 3. Whole-brain multivariate pattern analyses for reactive guilt and anger in Attitude phase. (A) Whole-brain multivariate patterns discriminating High *vs.* Low guilt conditions and High *vs.* Low anger conditions in Attitude phase. (B) Receiver operating characteristic curves (ROCs) for within-emotion 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

Low guilt (

reedict

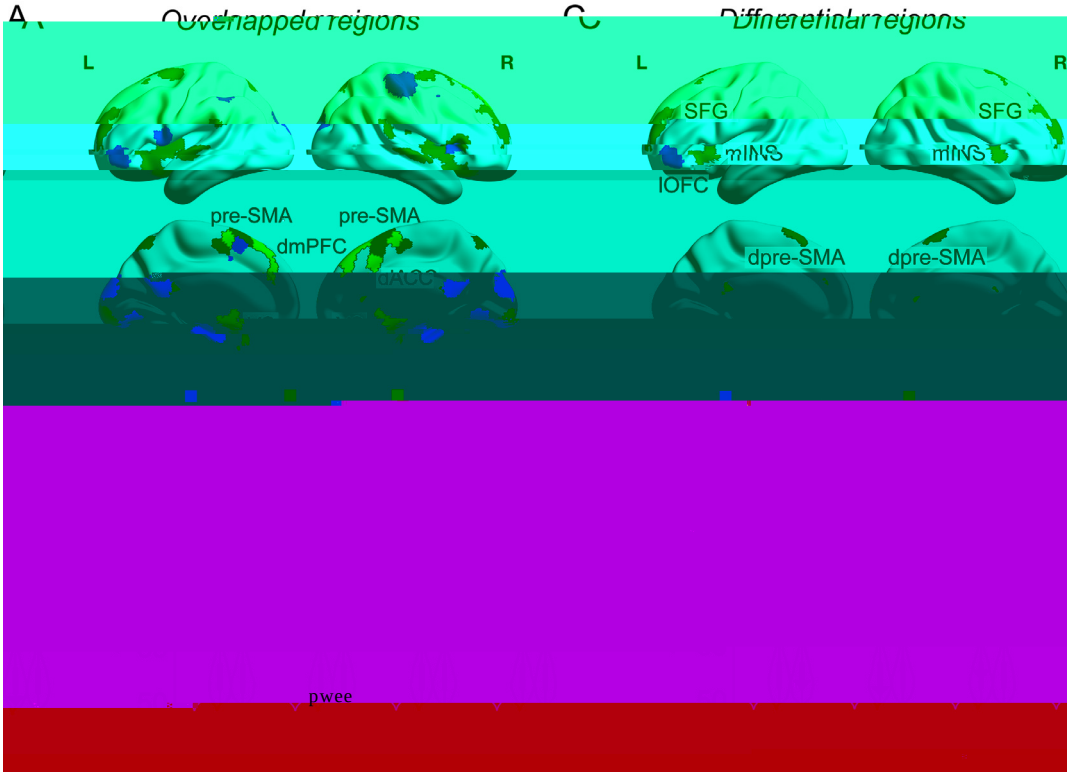


Fig. 4. Shared and di erential neural local classifiers for reactive guilt and reactive anger after receiving the co-player's attitude feedbacks. (A) Local classifiers that significantly discriminating High *vs.* Low guilt conditions (blue) and High *vs.* Low anger conditions (red) in Attitude phase, 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

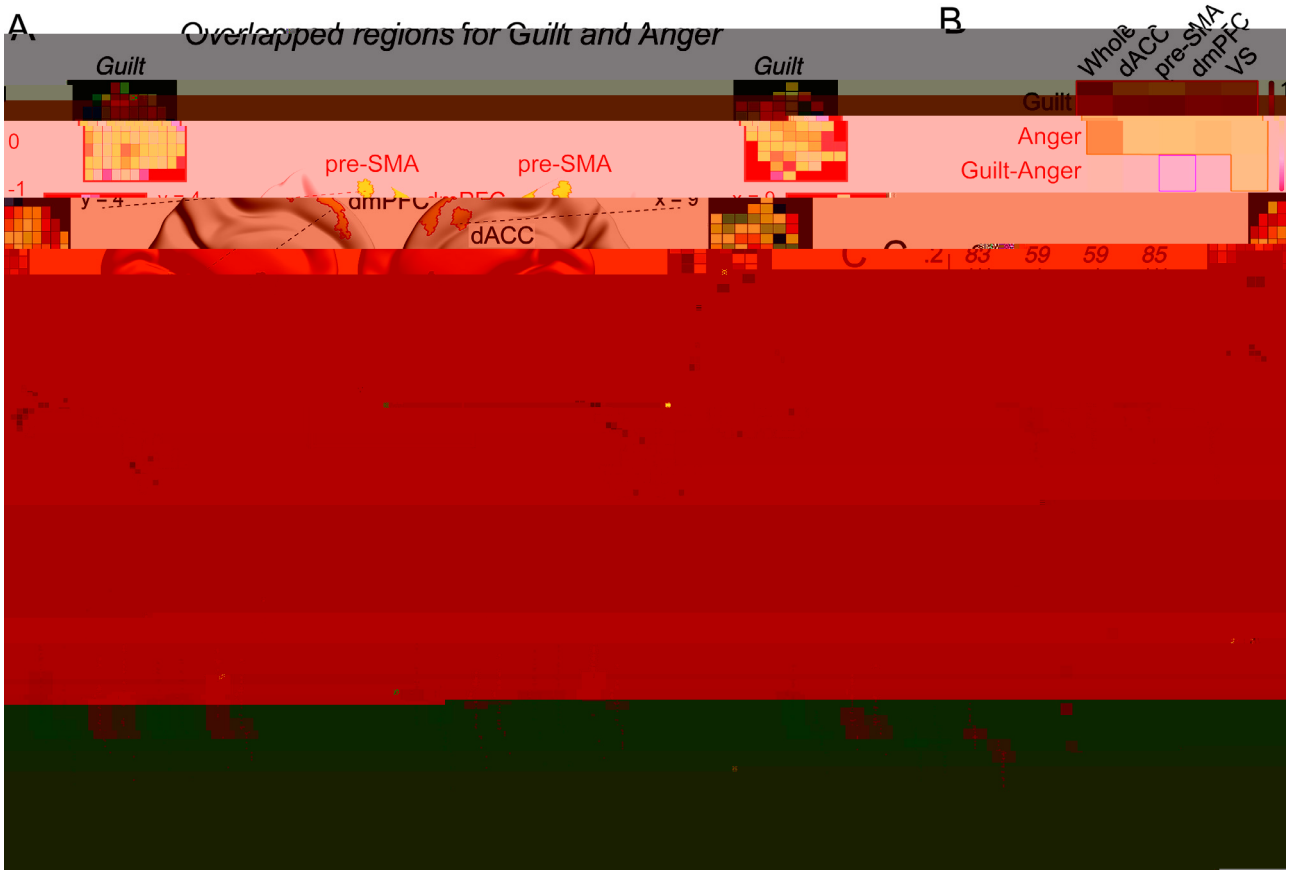


Fig. 5. Shared and di erential neural representations for overlapping regions identified for reactive guilt and anger. (A) Local patterns for overlapping regions (i.e., dACC, pre-SMA, dmPFC and VS) for processing reactive guilt and anger. (B) Pattern expression values of reactive guilt and anger in the 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

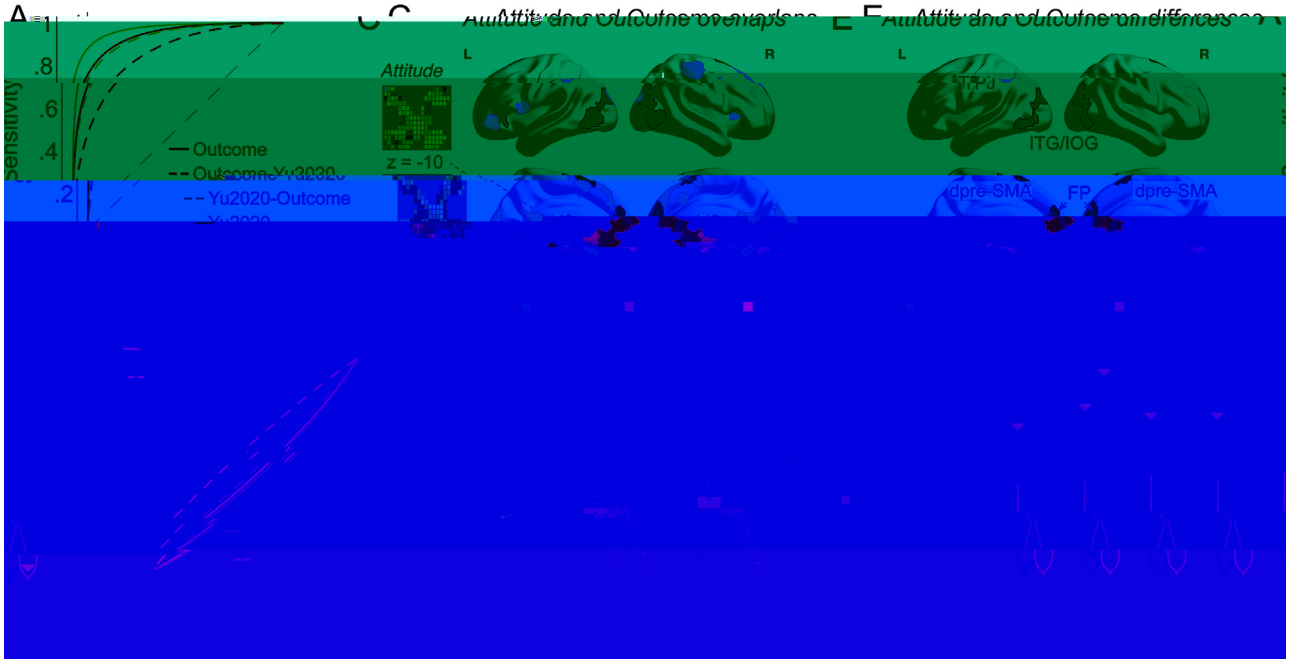


Fig. 6. Shared and di erential neural representations for reactive guilt in Attitude phase and non-reactive guilt in Outcome phase. (A) ROCs for the two-choice forced-alternative accuracies for within-study and cross-study classifications using the data of Outcome phase in the current study and of Yu 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 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% (± 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% (± 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% (± 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 di erent 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 ± 7.8%, $p = 0.442$, $p_{FDR} = 0.885$) and High *vs.*

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 fromude feedback, the neural representation of guilt (i.e., reactive guilt) might di er from the representation of guilt

fromfrom from

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 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 di erent 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 cogni-tion 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

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

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-directed behavior. Emotion 39, 76–99
Chang, L.J., Jolly, E., Cheong, J.H., Rapuano, K.M., Greenstein, N., Chen, , makes , , M.L.tice. M.L. , Anderson, S.R., 2002. equation equation

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