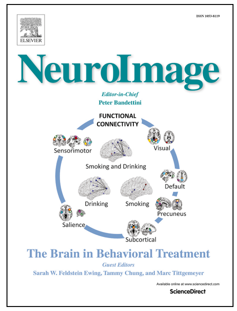
Accepted Manuscript



Differentiating guilt and shame in an interpersonal context with univariate activation and multivariate pattern analyses

Ruida Zhu, Chunliang Feng, Shen Zhang, Xiaoqin Mai, Chao Liu

PII: S1053-8119(18)32082-2

DOI: https://doi.org/10.1016/j.neuroimage.2018.11.012

Reference: YNIMG 15417

To appear in: *NeuroImage*

Received Date: 19 July 2017

Revised Date: 15 October 2018

Accepted Date: 9 November 2018

Please cite this article as: Zhu, R., Feng, C., Zhang, S., Mai, X., Liu, C., Differentiating guilt and shame in an interpersonal context with univariate activation and multivariate pattern analyses, *NeuroImage* (2018), doi: https://doi.org/10.1016/j.neuroimage.2018.11.012.

Differentiating Guilt and Shame in an Interpersonal Context with Univariate Activation and Multivariate Pattern Analyses

Ruida Zhu[1,2,3], Chunliang Feng[1,2,3], Shen Zhang[1,2,3], Xiaoqin Mai[4]* and Chao Liu[1,2,3]*

[1]State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, 100875, Beijing, China

[2]Center for Collaboration and Innovation in Brain and Learning Sciences, Beijing Normal University, 100875, Beijing, China

[3]Beijing Key Laboratory of Brain Imaging and Connectomics, Beijing Normal University, 100875, Beijing, China

[4]Department of Psychology, Renmin University of China, 100872, Beijing, China

*Correspondence to:

Xiaoqin Mai, Department of Psychology, Renmin University of China, Beijing, 100872, China

E-mail: maixq@ruc.edu.cn

Phone numbers: + 86 18910537082

Or

Chao Liu, State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, No. 59 Xinjiekouwai Street, Beijing, 100875, China

E-mail: liuchao@bnu.edu.cn

Phone numbers: +86 18911121354

Abstract. Guilt and shame are usually evoked during interpersonal interactions. However, no study has compared guilt and shame processing under such circumstances. In the present study, we investigated guilt and shame in an interpersonal context using functional magnetic resonance imaging (fMRI). Behaviorally, participants reported more "guilt" when their wrong advice caused a confederate's economic loss, whereas they reported more "shame" when their wrong advice were correctly refused by the confederate. The fMRI results showed that both guilt and shame activated regions related to the integration of theory of mind and self-referential information (dorsal medial prefrontal cortex, dmPFC) and to the emotional processing (anterior insula). Guilt relative to shame activated regions linked with theory of mind (supramarginal gyrus and temporo-parietal junction) and cognitive control (orbitofrontal cortex/ventrolateral prefrontal cortex and dorsolateral prefrontal cortex). Shame relative to guilt revealed no significant results. Using multivariate pattern analysis, we demonstrated that in addition to the regions found in the univariate activation analysis, the ventral anterior cingulate cortex and dmPFC could also distinguish guilt and shame. These results do not only echo previous studies of guilt and shame using recall and imagination paradigms but also provide new insights into the psychological and neural mechanisms of guilt and shame.

Keywords: guilt, shame, theory of mind, cognitive control, self-evaluation

Introduction

Guilt and shame, two typical moral emotions, often arise when social norms are violated (Haidt, 2003). They stop transgressors' further immoral behaviors by inhibiting their selfish impulses and making them concern others and blame themselves (Haidt, 2003). Guilt and shame play different roles in psychiatric disorders (Tangney & Dearing, 2003). Shame is positively related to various psychological problems, including depression, anxiety, and aggression, whereas guilt is not associated with most of these problems and even prevents the occurrence of aggression (Muris, 2015; Tangney, Wagner, Hill-Barlow, Marschall, & Gramzow, 1996). Considering their essential roles in norm compliance, large-scale cooperation, and psychiatric disorders, the past decade has witnessed a surge of interest in revealing the psychological and neural mechanisms underlying guilt and shame.

Guilt and shame share some similarities. In the experience of guilt and shame, transgressors need to understand others' suffering and blame themselves (Bastin, Harrison, Davey, Moll, & Whittle, 2016; Tangney & Dearing, 2003), so the capability of mentalizing and having a sense of self are thus required for these two emotions (Tangney & Dearing, 2003). In addition, guilt and shame are negative emotions, which evoke strong aversive feelings and psychological pain (Carnì, Petrocchi, Miglio, Mancini, & Couyoumdjian, 2013; Tangney & Dearing, 2003). These emotions could be so distressing that some transgressors punished themselves by putting their hands in ice water or giving themselves electric shock to attenuate them (Bastian, Jetten, & Fasoli, 2011; Nelissen & Zeelenberg, 2009). Consistently, a number of fMRI studies have found that both guilt and shame activated brain regions linked with theory of mind (e.g. superior temporal sulcus [STS] and temporo-parietal junction [TPJ]) (Finger, Marsh, Kamel, Mitchell, & Blair, 2006; Michl et al., 2014; Moll et al., 2007; Takahashi et al., 2004; Wagner, N'Diaye, Ethofer, & Vuilleumier, 2011), self-referential processing (e.g. anterior cingulate cortex [ACC] and posterior cingulate cortex [PCC]) (Michl et al., 2014; Moll et al., 2007; Shin et al., 2000; Yu, Hu, Hu, & Zhou, 2014), integration of theory of mind and self-referential information (e.g. dorsomedial prefrontal cortex [dmPFC]) (Finger et al., 2006; Fourie, Thomas, Amodio, Warton, & Meintjes, 2014; Michl et al., 2014; Moll et al., 2007; Shin et al., 2000), and emotional processing (e.g. anterior insula [AI] and amygdala) (Finger et al., 2006; Shin et al., 2000; Wagner et al., 2011; Yu et al., 2014).

In spite of those similarities, guilt and shame are also believed to be conceptually and theoretically different (Tangney, 1995, 1996). In guilt, transgressors focus on what they did to others and condemn their own immoral behavior (e.g. "I did a horrible thing"), whereas transgressors in shame focus on who they are and devalue themselves (e.g. "I am a bad person") (Lewis, 1971; Tangney & Dearing, 2003). Different foci often lead to different psychological processes and behavioral patterns. Compared with shame, guilt involves more other-oriented empathy (Tangney, Stuewig, & Mashek, 2007; Tangney, Stuewig, Mashek, & Hastings, 2011; Tangney & Dearing, 2003). It is not clear whether guilt involves more cognitive empathy (understand the others' mental state, also called theory of mind) or more emotional empathy (share others' emotion). However, findings that guilt (but not shame) facilitates relationship-reparation behaviors such as apology, compensation, and self-punishment could provide some clues (De Hooge, Zeelenberg, & Breugelmans, 2007; Howell, Turowski, & Buro, 2012; Yu et al., 2014; Zhu, Jin, et al., 2017). To form the motivation of relationship reparation, understanding the victims' state, such as dissatisfaction and potential revenge motivation, could be necessary (e.g. Nelissen, 2014). On the other hand, no study showed that guilt promotes individuals to feel the victims' feelings (e.g. anger or sadness). Compared with guilt, shame involves more self-oriented concerns about one's own negative image (Tangney et al., 2007; Tangney et al., 2011; Tangney & Dearing, 2003; Zhu et al., 2018), which causes image-reparation behaviors such as withdrawal, hiding (avoiding be directly criticized) and improvement of themselves (de Hooge, Zeelenberg, & Breugelmans, 2010; Gausel & Leach, 2011; Sznycer et al., 2016).

Although those theoretical distinctions between guilt and shame are quite clear, previous fMRI studies directly comparing guilt with shame found inconsistent results (Michl et al., 2014; Pulcu et al., 2014; Takahashi et al., 2004; Wagner et al., 2011). Three studies used imagination paradigms to induce target emotions by presenting participants hypothetical scenarios (Michl et al., 2014; Pulcu et al., 2014; Takahashi et al., 2004). Takahashi et al. (2004) showed that guilt compared to shame increased activation in the medial prefrontal cortex (mPFC), while shame compared to guilt increased activation in the middle temperal gyrus (MTG), hippocampus and visual cortex. On the contrary, Michl et al. (2014) revealed that guilt compared to shame increased activation in the MTG, insula and fusiform gyrus, whereas shame compared to guilt increased activation in the mPFC, dACC, inferior forntal gyrus, PCC, and parahippocampus. Pulcu et al. (2014) found shame compared to guilt increased activation in the amygdala and posterier insula in a major depressive disorder group, but not in a healthy control group. Another study used a recall paradigm to evoke target emotions by asking participants to recall personal experiences (Wagner et al., 2011). Results showed that guilt compared to shame activated the theory of mind network (e.g. dmPFC, STS, and temporal pole), the cognitive control network (e.g. orbitofrontal cortex (OFC) and dorsolateral prefrontal cortex (dlPFC)), the salience network (e.g. AI and amygdala), and other regions (e.g. cerebellum), but no significant effect was found when comparing shame to guilt (Wagner et al., 2011).

These inconsistent findings were probably caused by limitations of the existing experimental paradigms and analysis methods. As for the paradigms, the psychological processes of both imagination and recall are not necessary for guilt and shame (Bastin et al., 2016; Yu et al., 2014). The imagination and recall paradigms may casue some brain activations related to imagination and recall processing themselves rather than guilt and shame processing. Besides, individual differences in the ability to vividly create or recreate guilt and shame events in their mind could be another confounding variable. In addition, imagination and recall may not be able to completely capture the essential psychological processes of guilt and shame (Bastin et al., 2016). For example, a study directly comparing the recall and imagination paradigms to induce guilt suggests that the imagination paradigm may only induce some anticipatory thoughts but few emotional feelings (Mclatchie, Giner-Sorolla, &

Derbyshire, 2016). As for the analysis methods, previous studies merely used traditional univariate activation analysis to examine the neural correlates of guilt and shame. The univariate activation analysis, which is not as sensitive as other methods, such as multivariate pattern analysis (MVPA), may be unable to detect subtle differences between guilt and shame (Norman, Polyn, Detre, & Haxby, 2006; Pereira, Mitchell, & Botvinick, 2009).

Concerning the limitations above, the present study attempted to extend previous studies in two aspects. First, we developed a novel paradigm to induce guilt and shame in an interpersonal context. It enabled participants to directly experience guilt and shame during social interactions, which excluded unrelated psychological processes (e.g. imagination and recall). In fact, daily experience of guilt and shame (including thoughts and feelings) usually happens during interpersonal interactions but not imagination and recall (Yu et al., 2014). Combining fMRI techniques, we explored the neural correlates of interpersonal guilt and shame (with happiness, a non-moral emotion, as a control). Second, we did not only use the traditional univariate activation analysis, which enabled us to directly compare our results with the results of previous studies, but also for the first time conducted MVPA to explore the neural differences between guilt and shame. MVPA extracts and analyzes signals that are presented in the patterns of responses across multiple voxels and shows increased sensitivity compared to the univariate analysis (Norman et al., 2006)[1]. Previous studies using univariate analysis methods found many brain regions

[1] An example is presented in the supplementary materials to conceptually explain the difference between univariate activation analysis and multivariate pattern analysis.

activated similarly during different basic emotional states (Lindquist & Barrett, 2012; Phan, Wager, Taylor, & Liberzon, 2002; Vytal & Hamann, 2010) and the corresponding meta-analyses had difficulty in establishing unique neural correlates for different basic emotions (Lindquist & Barrett, 2012; Saarimäki et al., 2016). On the other hand, studies using MVPA have proved success in decoding emotional signals and revealing discrete neural signatures of basic emotions (Baucom, Wedell, Wang, Blitzer, & Shinkareva, 2012; Saarimäki et al., 2016). It suggests that at least some emotional signals in the brain are represented in multiple voxels instead of each single voxel. Therefore, we employed MVPA to identify brain regions that encode information about guilt and shame but show no regional-average activation changes in the contrasts between guilt and shame.

According to the existing theory and findings that guilt may involve more theory of mind processing, whereas shame may involve more self-referential processing (e.g. Lewis, 1971; Tangney & Dearing, 2003), we expected that the neural differences between guilt and shame would occur in the core regions linked with theory of mind and self-referential processing.

Methods

Participants

Thirty-three right-handed healthy students from Beijing Normal University participated in the experiment for payment. All participants provided written consent and reported no history of psychiatric, neurological, or cognitive diseases. Three participants were excluded due to excessive head motion (> 3 mm, one participant) or suspicion about the authenticity of the task (two participants), leaving thirty participants (17 females, M_{age} = 21.57 years, *SD* = 2.34) in final analyses. One male and one female students (both aged 22 years), who were strangers to the participants, were recruited as confederates. The study was approved by the Institutional Review Board at Beijing Normal University.

Task design

Upon arrival participants met a same sex confederate and were told that they would play an advice-decision game (adapted from a study on interpsersonal guilt, Yu, Hu, Hu, & Zhou, 2014) together via the internal network. Then they were led to different rooms and received instructions separately. In the advice-decision game, there were two roles, an advisor and a decider. During each trial, the advisor looked at a picture of dots (always containing 20 dots but in random positions) for 2 s and provided his or her advice about the number of the dots (more or less than 20) for the decider within 2 s. In the meantime, the decider looked at the same picture, but only for 1 s, and then decided whether to take the advice that he or she got from the advisor within 3 s. Then, the advisor and decider saw the outcomes of the advice and decision. Finally, two affective words emerged and the participants chose one word that precisely described their emotion at that time (Figure 1). Different words followed different outcomes (Table 1). The left and right positions of affective words were counterbalanced. Importantly, participants were clearly told that they did not have to respond if both words failed to match their current emotion. It was informed that when acting as the decider, participants received 1 Chinese yuan as reward for each right decision and lost 1 Chinese yuan as punishment for each wrong decision. When acting as the advisor, participants received 90 Chinese yuan as participation fee regardless of the correctness of their advice.

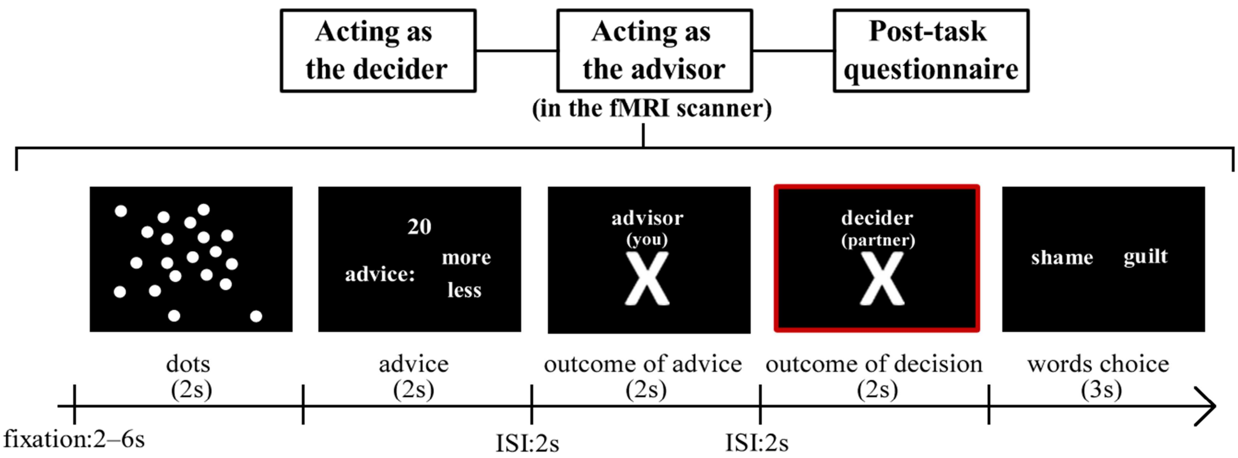


Figure 1. Timeline of the experimental procedure. We analyzed the fMRI data during the outcome stage of the decision (marked with a red frame in the figure). ISI: interstimulus interval.

Table 1. Affective words following different outcomes.

Roles	Conditions	Outcomes		Affective words
		Advice	Decision	
Decider		wrong	wrong	sadness or anger
		wrong	right	happiness or pride
		right	right	happiness or pride
		right	wrong	sadness or shame
Advisor	Guilt	wrong	wrong	guilt or shame
	Shame	wrong	right	guilt or shame
	Happiness	right	right	happiness or pride
	Uncertainty	right	wrong	happiness or pride

Procedure

Before acting as the advisor in the scanner, participants acted as the decider for 30 trials outside the fMRI scanner. The outcome of their decision was determined by following rules: If they adopted the advisor's advice, their probability of making a correct decision was 80%; otherwise, the probability was 20%. The feelings of guilt and shame were influenced by people's perception of responsibility and task difficulty (Hoffman, 1982). Such a manipulation highlighted the responsibility of the advisor and implied that the task of the advisor was not too difficult, which could strengthen the participants' guilt and shame when they acted as the advisor later.

During fMRI scanning, participants played the role of advisor for 96 trials (3 sessions, 32 trials in each session). In the 30 trials of the guilt condition, it was shown that the participant's advice and the decider's decision were wrong, which inferred that the participant's wrong advice, at least to some extent, caused the monetary loss of the decider. Indeed, bad outcomes and the responsibility for the bad outcomes cause guilt (Carnì, Petrocchi, Del Miglio, Mancini, & Couyoumdjian, 2013; Tangney & Dearing, 2003; Tracy & Robins, 2006). In the 30 trials of the shame condition, the advice was wrong but the decision was right, which implied that the decider had a better performance than the participants. It meant even though the decider had less time to look at the picture (1 s) than the participants (2 s), he or she correctly rejected the participant's wrong advice. The feelings of inability and rejection could induce shame (Leach, 2011; Smith, Webster, Parrott, & Eyre, 2002; Tangney & Dearing, 2003; Tracy & Robins, 2006). In 30 trials of the happiness condition (a control condition without guilt and shame), the advice and decision were right. In the remaining 6 trials of the uncertain condition, the advice was right and the decision was wrong. The number of this condition was set to be less than other conditions, because the results of a pilot experiment found that when the trial number of the uncertain condition was same as that of the shame condition, participants' feeling of shame was strongly weakened in the shame condition. If participants found that the decider correctly rejected the advice as many times as they wrongly rejected the advice, they thought the decider's good performance in the shame condition was just by luck and thus did not feel ashamed in the shame condition. Different trials were presented in a pseudo-random order, ensuring the trials of the same condition did not consecutively appear more than three times.

Post-task questionnaire and debriefing

After the game, the participants rated how strongly (1 = not at all, 9 = very strong) they felt each of six emotions (sadness, shame, happiness, guilt, anger, and pride) for different conditions and completed a test of instruction comprehension. All participants passed the test. In the end, the participants were debriefed and received 120 Chinese yuan as compensation.

Image Acquisition

Images were acquired on a 3 T Siemens Trio scanner with a 12 channel head coil at Beijing Normal University's Imaging Center, China. To acquire functional images, a T2-weighted functional images gradient-echo-planar imaging (EPI) sequence was used (number of slices = 33, TR = 2000 ms, TE = 30 ms, flip angle = 90°, slices thickness = 3.5 mm, gap between slices = 0.7 mm and FOV = 224 mm × 224 mm). High-resolution, whole brain, structural images were acquired by using a magnetization prepared rapid acquisition with gradient-echo (MPRAGE) sequence (number of slices = 144, TR = 2530 ms, TE = 3.39 ms, flip angle = 7°, slices thickness = 1.33 mm, gap between slices = 0.7 mm and FOV = 256 mm × 256 mm).

fMRI data analysis

Preprocessing

We focused on the behavioral and fMRI data when the participants acted as the advisor. Trials in which participants did not provide their advice were excluded from analyses. For neuroimaging data analyses, we used the Matlab based (The MathWorks, Inc) software SPM8 (http://www.fil.ion.ucl.ac.uk/spm). Preprocessing steps included slice timing correction, realignment, normalization to Montreal Neurological Institute (MNI) space (new voxel size = 3 × 3 × 3 mm^3), smoothing with an 6 mm full width at half maximum (FWHM) Gaussian kernel, and high-pass temporal filtering at 1/128 Hz to remove low frequency drifts.

Univariate activation analysis

At the individual level, we modeled the dots, the advice, the outcome of the advice, the outcome of the decision, the words choice, and the missing trials (participants did not give their advice) separately in the general linear model (GLM). The outcome of the decision event was further divided into four regressors corresponding to the four conditions (Guilt, Shame, Happiness, and Uncertainty)[2]. Only Guilt, Shame and Happiness conditions were analyzed. Six movement parameters were defined as nuisance regressors. All the regressors except for the nuisance regressors were convolved with canonical hemodynamic response function.

At the group level, contrasts of Guilt > Happiness, Shame > Happiness, Guilt > Shame, and Shame > Guilt were entered into a random effect analysis. The statistical

[2] In the main manuscript, we defined the guilt and shame conditions based on the outcomes (e.g., the participant's advice and the decider's decision were wrong). The guilt and shame conditions could also be defined based on the participant's self-report (e.g. the participant chose 'guilt' in the trial). In the supplementary materials, we illustrated why we defined the guilt and shame conditions according to the outcomes, but still showed the results of the univariate activation analysis when the guilt and shame conditions were defined based on the participant's self-report.

threshold was set at a threshold of $p < .001$ uncorrected at voxel level and an extent threshold of $p < .05$ with family-wise error (FWE) correction at cluster level (see Woo, Krishnan, & Wager, 2014).

To access common regions activated by guilt and shame conditions, we performed a conjunction analysis (Guilt > Happiness ∩ Shame > Happiness). The statistical threshold was same as the one used in the activation analysis.

Multivariate pattern analysis

MVPA was implemented on non-normalized and unsmoothed data. A GLM was built for each individual, which was identical to the one used in the univariate analysis, with the exception that trials were modeled separately here. The parameter estimates of the GLM were analyzed by a support vector machine (SVM) classifier embedded in the Decoding Toolbox (https://sites.google.com/site/tdtdecodingtoolbox/) (Hebart, Görgen, Haynes, & Dubois, 2015). The searchlight decoding analysis could be accomplished by using SVM or other machine-learning algorithms (e.g. linear discriminant analysis [LDA]). However, it has been suggested that SVM has lots of advantages compared to other algorithms (e.g. SVM deals with limited data in high-dimensional spaces gracefully and naturally and is less affected by data points shift far away from boundary) (Cui & Gong, 2018; Ledoit & Wolf, 2003; Mur, Bandettini, & Kriegeskorte, 2009). Considering many recent studies have demonstrated the reliability of SVM (Feng et al., 2016, 2017; Feng, Zhu, et al., 2018; Yu, Cai, Shen, Gao, & Zhou, 2016), SVM was chosen in our study. We performed a whole-brain searchlight decoding analysis using a sphere with a radius of four voxels. Using the data of voxels in each sphere, the SVM classifier was trained and then tested according to a leave-one-run-out cross-validation method. The classification accuracy of each sphere was assigned to the central voxel of the sphere, yielding a 3D map of classification accuracy. The map of each individual was normalized (to MNI space, voxel size = 3 × 3 × 3 mm^3), smoothed (6 mm FWHM Gaussian kernel) and entered into the group level analysis. To make inference, these maps were entered into a second-level permutation based analysis using the Statistical NonParametric Mapping toolbox (SnPM, http://warwick.ac.uk/snpm) with 5,000 permutations. The resulting voxels were assessed for significance at 5% level with voxel-wise FWE correction, as determined by permutated datasets (see Nichols & Holmes, 2002). Clusters containing more than 10 voxels were reported. We used the reported clusters as masks to extract the classification accuracy of the voxels within each cluster and calculated the mean accuracy for each cluster. The mean accuracy indicated the average percentage of correct guesses when the trained model used the signal of a sphere with a radius of four voxels within a certain cluster.

Results

Behavioral Results

In the guilt condition, "guilt" (mean [*M*] = 21.87, standard deviation [*SD*] = 5.92) was more frequently chosen than "shame" (*M* = 7.77, *SD* = 5.93; $F(1, 29) = 42.55$, $p < .001$, $\eta_p^2 = .60$) and "no response" (*M* = 0.07, *SD* = 0.25; $F(1, 29) = 391.98$, *p*

< .001, η_p^2 = .93), and the post-task ratings of guilt were significantly higher than the ratings of other emotions, all *F*s > 27.98, all *ps* < .001, all $\eta_p^2 s$ > .49 (Figures 2, S2 and S3). In the shame condition, “shame” (*M* = 18.63, *SD* = 7.31) was more frequently chosen than “guilt” (*M* = 11.17, *SD* = 7.28; *F*(1, 29) = 7.86, *p* = .009, η_p^2 = .21) and “no response” (*M* = 0.03, *SD* = 0.18; *F*(1, 29) = 193.04, *p* < .001, η_p^2 = .87), and the post-task ratings of shame were significantly higher than the ratings of other emotions, all *F*s > 9.75, all *ps* < .004, all $\eta_p^2 s$ > .25. The guilt ratings were higher in the guilt than shame condition (*F*(1, 29) = 29.73, *p* < .001, η_p^2 = .51) and the shame ratings were higher in the shame than guilt condition (*F*(1, 29) = 21.86, *p* < .001, η_p^2 = .43). In the happiness condition, “happiness” (*M* = 23.63, *SD* = 6.20) was more frequently chosen than “pride” (*M* = 6.20, *SD* = 6.27; *F*(1, 29) = 58.71, *p* < .001, η_p^2 = .67) and “no response” (*M* = 0.03, *SD* = 0.18; *F*(1,29) = 437.64, *p* < .001, η_p^2 = .94), and the ratings of happiness were significantly higher than the ratings of other emotions, all *F*s > 21.12, all *ps* < .001, all $\eta_p^2 s$ > .42. These results demonstrated that our manipulation successfully induced target emotion in each condition.

There was no significant difference between the guilt ratings in the guilt condition and the shame ratings in the shame condition (*F*(1, 29) = 2.99, *p* = .095, η_p^2 = .093), between the shame ratings in the guilt condition and the guilt ratings in the shame condition (*F*(1, 29) = .183, *p* = .672, η_p^2 = .006), or between the sum of the guilt and shame ratings in guilt condition and the sum of the guilt and shame ratings in the shame condition (*F*(1, 29) = .3.38, *p* = .076, η_p^2 = .105). There was no significant difference in sadness, anger, happiness, or pride ratings between guilt and shame conditions either (all *F*s < 2.98, all *ps* > .095, all $\eta_p^2 s$ < .93). The results demonstrated that the emotion intensity of participants were comparable between the guilt and shame conditions.

The number of participants' "guilt", "shame", and "happiness" choices respectively in the guilt, shame and happiness conditions did not significantly change across three sessions (all *F*s < 3.12, all *ps* > .051, all $\eta_p^2 s$ < .10), which implied that the target emotion in each condition was stable across three sessions (Figure S4).

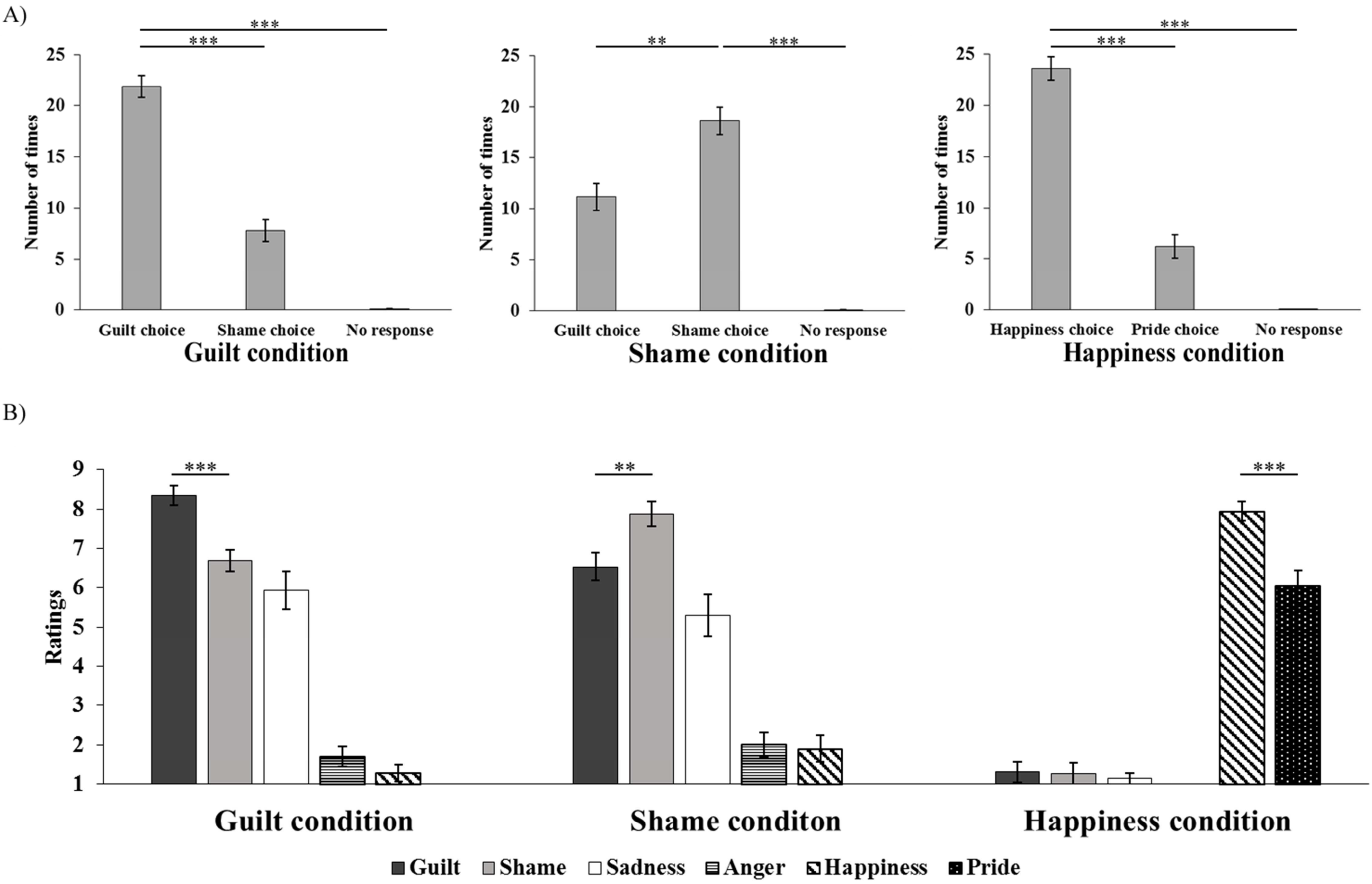


Figure 2. Behavioral results. A) Participants' choice of affective words in the guilt, shame and happiness conditions (means and standard errors). B) Participants'

post-task ratings of different emotions in the guilt, shame, and happiness conditions (means and standard errors). $^{**}p < .01$, $^{***}p < .001$.

Neuroimaging Results

Univariate activation analysis

The guilt condition relative to the happiness condition produced greater activation in the dmPFC, bilateral AI, right MTG, and cerebellum (Table 2 and Figure 3). The shame condition relative to the happiness condition elicited greater activation in the dmPFC and left AI. The conjunction analysis of the Guilt > Happiness and Shame > Happiness contrasts revealed two significant regions including dmPFC and left AI (Table 2).

Table 2. Brain activation in the guilt and shame conditions relative to the happiness condition and brain regions co-activated by the guilt and shame conditions ($p < .001$, uncorrected voxel-level and $p < .05$, cluster level with FWE correction). L, left; R, right.

	BA	MNI coordinates			*T* score	Voxels
Region		*x*	*y*	*z*		
Guilt > Happiness						
L/R dorsomedial prefrontal cortex	10/9	-9	51	21	7.02	746
L anterior insula	47	-30	18	-12	9.06	375
R anterior insula	47	30	18	-12	6.16	174
R middle temporal gyrus	21	54	-27	-9	5.74	90
L/R cerebellum		3	-51	-33	6.02	75
Shame > Happiness						
L/R dorsomedial prefrontal cortex	9	-9	51	18	5.33	148
L anterior insula	47	-30	18	-12	6.00	140
(Guilt > Happiness) ∩ (Shame > Happiness)						
L/R dorsomedial prefrontal cortex	9	-9	51	21	4.83	131
L anterior insula	47	-30	18	-12	6.29	152

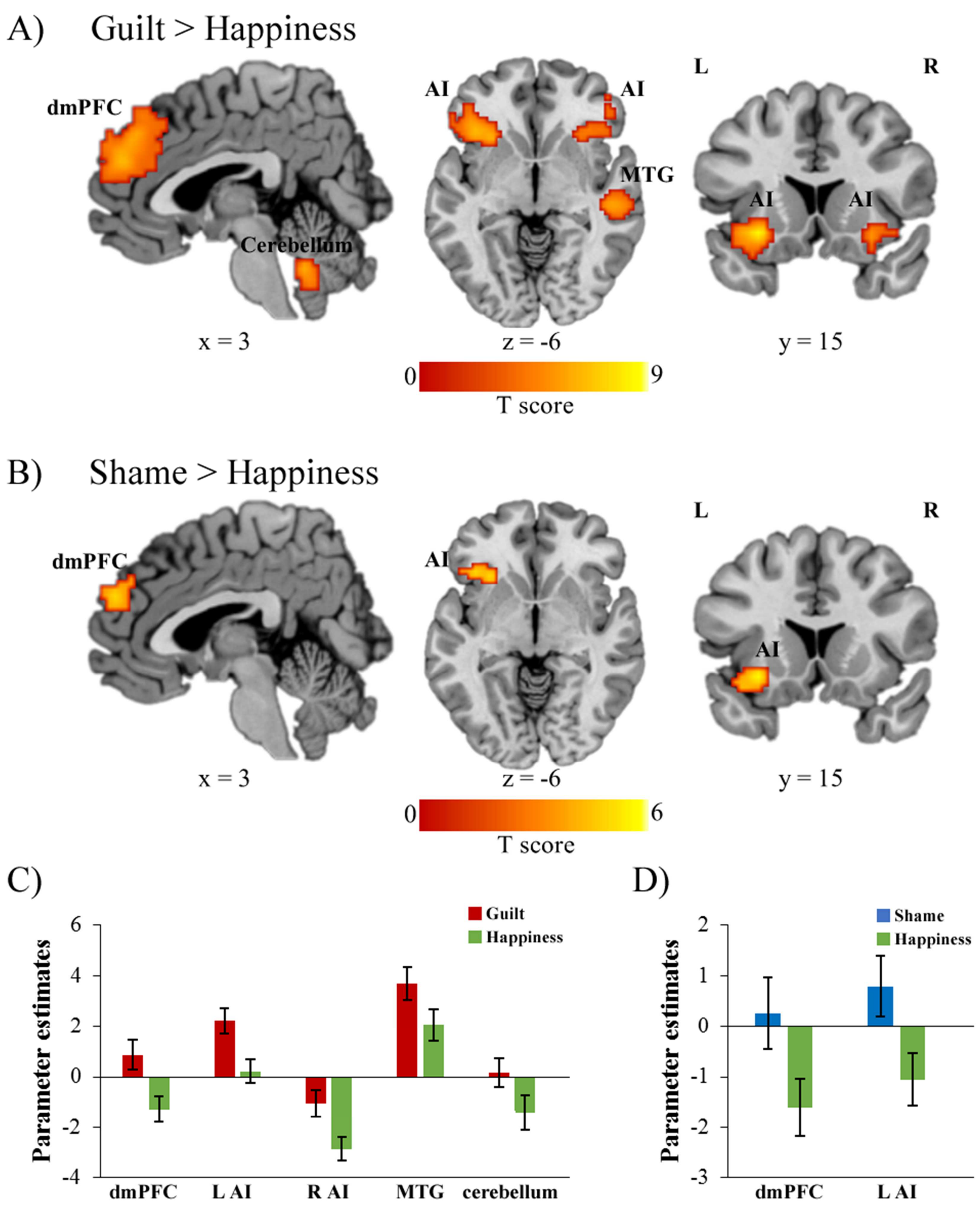


Figure 3. Brain activation in the guilt and shame conditions relative to the happiness condition. **A**) Guilt > Happiness. Activated regions were dmPFC, bilateral AI, MTG, and cerebellum. **B**) Shame > Happiness. Activated regions were dmPFC and left AI. **C**) The parameter estimates of the dmPFC, left AI, right AI, MTG and cerebellum in the Guilt > Happiness contrast (means and standard errors). **D**) The parameter

estimates of the dmPFC and left AI in the Shame > Happiness contrast (means and standard errors). L, left; R, right; dmPFC, dorsomedial prefrontal cortex; AI, anterior insula; MTG, middle temporal gyrus.

As expected, the guilt condition compared to the shame condition produced significant activation in brain regions related to theory of mind (left supramarginal gyrus and right TPJ) (Table 3 and Figure 4). In addition, the regions related to cognitive control (right vlPFC/OFC and right dlPFC) were also activated. Shame condition compared to guilt condition revealed no significant results under the predetermined threshold.

Table 3. Brain activation in the comparison between guilt and shame conditions ($p < .001$, uncorrected voxel-level and $p < .05$, cluster level with FWE correction). L, left; R, right. vlPFC, ventrolateral prefrontal cortex; OFC, orbitofrontal cortex.

	BA	MNI coordinates			*T* score	Voxels
Region		*x*	*y*	*z*		
Guilt > Shame						
R vlPFC/OFC	11/10	30	54	6	5.71	349
R dorsolateral prefrontal cortex	45	45	33	24	5.44	84
L supramarginal gyrus/postcentral gyrus	40/2	-57	-21	30	5.20	109
R temporo-parietal junction	40/39	54	-51	33	4.73	72
Shame > Guilt						
None.						

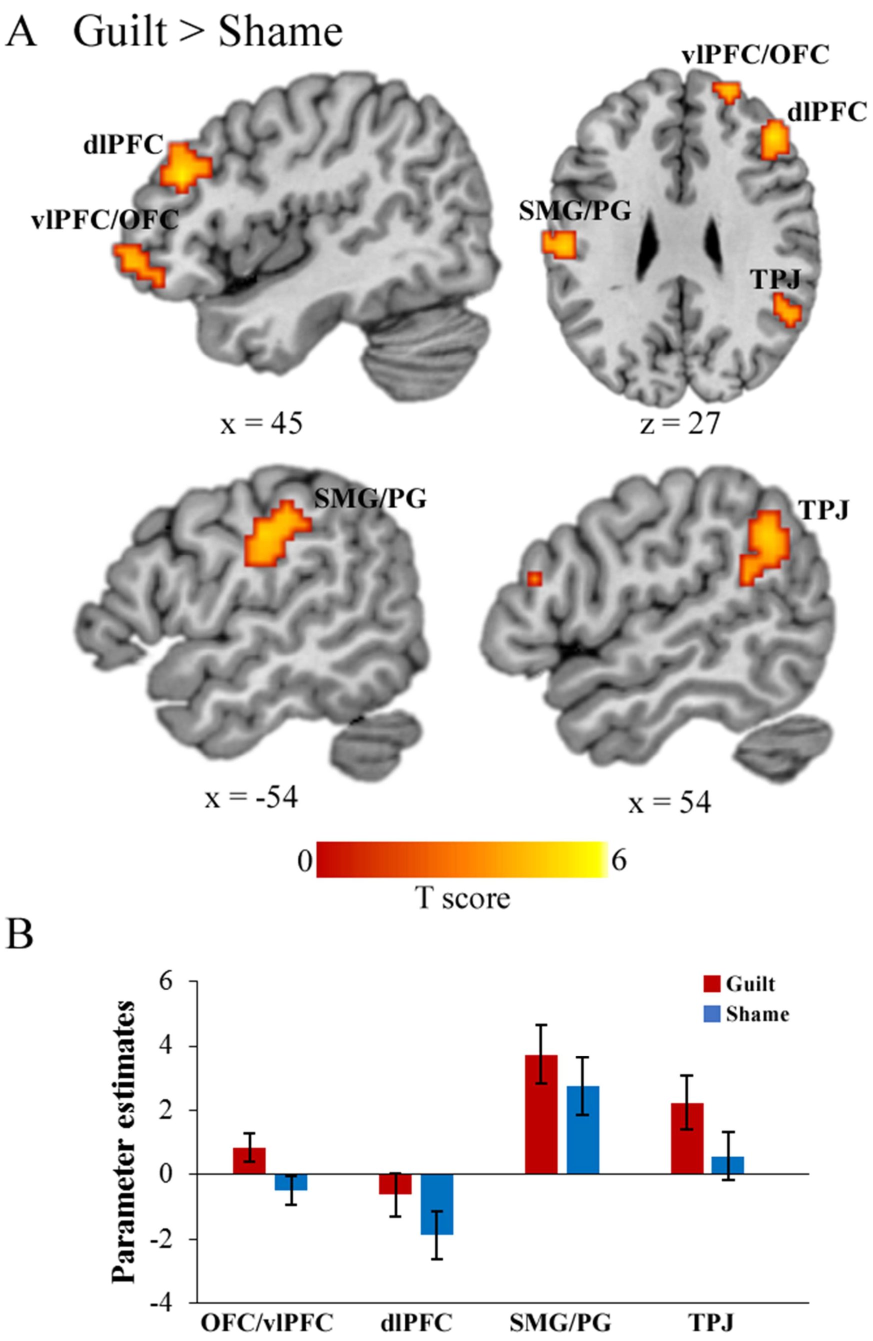


Figure 4. Brain activation in the comparison between guilt and shame conditions. **A**) Guilt > Shame contrast showed significant activation in the vlPFC/OFC, dlPFC, left supramarginal gyrus/precentral gyrus, and right TPJ. **B**) The parameter estimates of the vlPFC/OFC, dlPFC, left supramarginal gyrus/precentral gyrus, and right TPJ in

the Guilt > Shame contrast. L, left; R, right; vlPFC, ventrolateral prefrontal cortex; OFC, orbitofrontal cortex; dlPFC, dorsolateral prefrontal cortex; SMG, supramarginal gyrus, PG, postcentral gyrus, TPJ, temporo-parietal junction.

Multivariate pattern analysis

The MVPA results revealed that several regions exhibited differential multivariate representations of guilt vs. shame, comprising theory of mind related regions (right TPJ), cognitive control related regions (right vlPFC and left dlPFC), a self-referential processing related region (the vACC part of a large cluster), and a region related to both theory of mind and self-evaluation (the dmPFC part of a large cluster) (Table 4 and Figure 5). Among these regions, vlPFC, dlPFC, and TPJ were also identified with univariate analysis, whereas dmPFC and vACC did not show differences in the average regional activity between the guilt and shame conditions.

Table 4. Results of the multivariate analysis ($p < .05$, voxel-level with FWE correction, as determined by permutation distribution with 5,000 permutations, cluster size > 10). L, left; R, right; dmPFC, dorsomedial prefrontal cortex; vACC, ventral anterior cingulate cortex.

	BA	MNI coordinates			*T* score	Voxels
Region		*x*	*y*	*z*		
L/R dmPFC	10/9	3	51	21	6.87	517
L/R vACC	32	0	48	6	5.18	
R ventrolateral prefrontal cortex	45	42	18	12	5.56	11
L dosolateral prefrontal cortex	8/6	-30	3	45	6.59	76
R temporo-parietal junction	40/39	57	-51	30	5.14	18

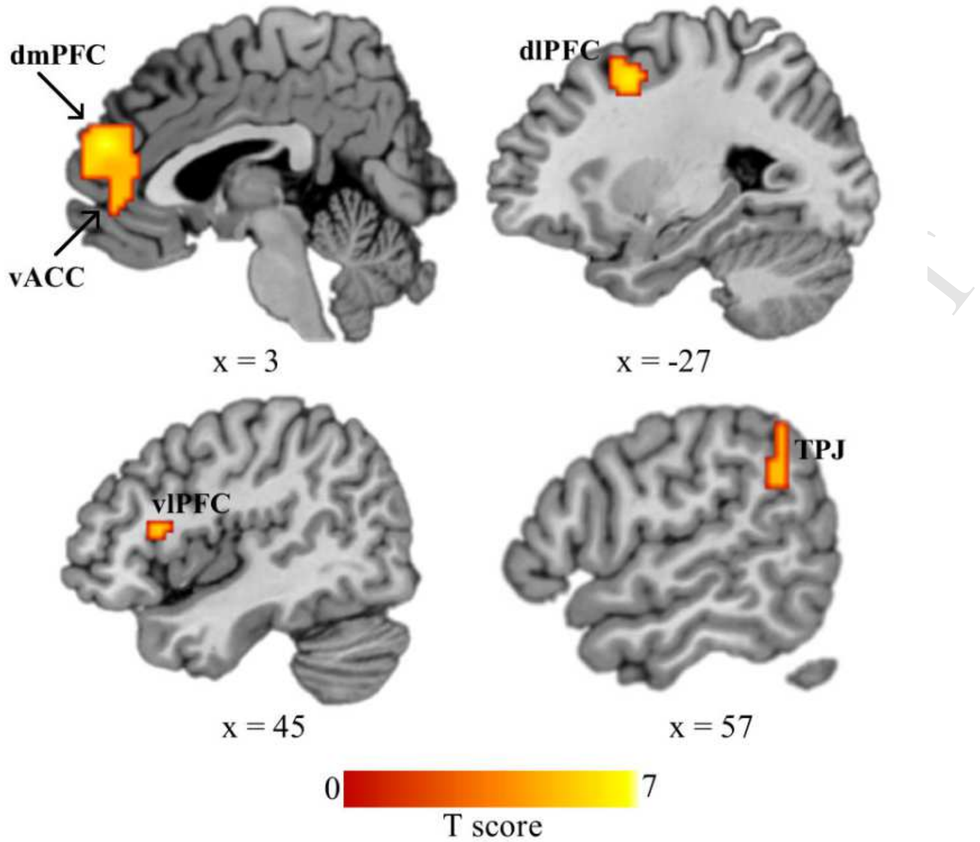


Figure 5. Results of the multivariate pattern analysis. Brain regions of dmPFC/vACC (*M* = 54.94%, *SE* = 0.76%), vlPFC (*M* = 53.67%, *SE* = 0.68%), dlPFC (*M* = 53.87%, *SE* = 0.63%) and TPJ (*M* = 53.95%, *SE* = 0.73%) exhibited significantly higher classification accuracy of guilt vs. shame than chance level (50%) (e.g. Schuck et al., 2015). dmPFC, dorsomedial prefrontal cortex; vACC, ventral anterior cingulate cortex; vlPFC, ventrolateral prefrontal cortex; dlPFC, dorsolateral prefrontal cortex; TPJ, temporo-parietal junction. M, mean accuracy; SE, standard error of accuracy.

Discussion

Our study investigated the neural correlates of guilt and shame in an interpersonal context. The behavioral results demonstrated that the target emotion was successfully evoked in each condition. Aligned with previous studies (Michl et al., 2014; Roth et al., 2014; Seara-Cardoso et al., 2016; Takahashi et al., 2004; Wagner et al., 2011), our results revealed that both guilt and shame elicited activation in the dmPFC and AI. The dmPFC is known as a core region in both the theory of mind network (for a review, see Schurz et al., 2014) and self-referential processing (for a review, see Northoff et al., 2006). It is believed to be a vital region where people integrate information of others' thoughts and emotion states with themselves' (D'Argembeau et al., 2007; Rebecca Saxe, Moran, Scholz, & Gabrieli, 2006). In the state of guilt and shame, the dmPFC may enable transgressors to understand others' suffering and negative attitudes toward them and to blame themselves. The AI is a key node in the salience network, which has a central role in detecting salient events (see a review, Uddin, 2015). It engages during experiencing various negative emotions, such as sadness and disgust (Craig, 2009). It is activated during the experience of both physical pain (e.g. receiving electric shock) and psychological pain (e.g. watching other's suffering or being excluded by others) (Gunther Moor et al., 2012; Singer et al., 2004). Moreover, the AI is more activated when individuals act morally than when they act immorally and is directly correlated with anticipatory guilt (Chang, Smith, Dufwenberg, & Sanfey, 2011; Ty, Mitchell, & Finger, 2017). These findings suggest that the AI may be involved in detecting salient social events in our study. Generally, the dmPFC and AI may respectively play important roles in cognitive processing and emotional processing during guilt and shame.

The theoretic work suggests that guilt compared to shame involves more other-oriented empathy (Tangney et al., 2007; Tangney & Dearing, 2003). Guilt but not shame promotes relationship-reparation behavior further implying that transgressors in guilt may have understood the victims dissatisfaction and potential revenge tendency (theory of mind processing) (De Hooge et al., 2007; Nelissen, 2014; Yu et al., 2014). Recent studies also showed that guilt is moderated by the relational utility of the victim, which also indirectly indicates transgressors in guilt do track the state of the victims (Nelissen, 2014; Ohtsubo & Yagi, 2015; Zhu et al., 2017). Supporting the hypothesis, we found that guilt evoked increased activity in the left supramarginal gyrus and right TPJ than shame. Both the supramarginal gyrus and TPJ belong to the theory of mind network (Schurz et al., 2014) and some researchers consider the supramarginal gyrus as a part of the TPJ (Gifuni, Kendal, & Jollant, 2016). It is worth noting that the TPJ is a relatively large and roughly characterized region. The posterior portion of the TPJ is implicated in the theory of mind (Aichhorn, Perner, Kronbichler, Staffen, & Ladurner, 2006; Saxe & Kanwisher, 2003; Schurz et al., 2014), while the anterior portion of the TPJ is engaged in the attention orientation (Decety & Lamm, 2007; Lindquist & Barrett, 2012). As our study did not localize the theory of mind network for each participant, it is not sure that the TPJ found in our task was related to the theory of mind or the attention orientation. However, based on the coordinates reported by a recent meta-analyses study of the theory of mind (the reported peak coordinates [56, -55, 27] of the right TPJ related to the theory of mind was within the right TPJ cluster found in our study, Figure S5), it is very likely the TPJ reported in our study played a role in the theory of mind (Schurz et al., 2014).

Accordingly, our results suggest that transgressors have more theory of mind processing when they feel guilty than ashamed.

Guilt relative to shame also increased the activity in cognitive control regions consisting of the OFC/vlPFC and dlPFC. These results are in line with a previous study using a recall paradigm to induce guilt and shame, which found similar neural activations (OFC and dlPFC) when comparing guilt to shame (Wagner et al., 2011). The vlPFC and dlPFC are implicated in controlling impulsive behaviors and optimizing social decisions (Feng, Luo, & Krueger, 2015; Koechlin, 2003). For example, brain stimulation studies have found that the disruption of the vlPFC or dlPFC, using transcranial magnetic stimulation or transcranial direct current stimulation, diminishes the ability to inhibit selfish or aggressive impulses, which could incur punishment and relationship damage (Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006; Knoch, Schneider, Schunk, Hohmann, & Fehr, 2009; Riva, Romero Lauro, DeWall, Chester, & Bushman, 2014; Strang et al., 2015). Therefore, in the state of guilt, the OFC/vlPFC and dlPFC may make transgressors curb their selfish impulses and bear some costs to make compensation in the future. Behavioral studies indeed have found that guilt is more likely to induce costly relationship-reparation behaviors than shame (Brown, González, Zagefka, Manzi, & Cehajic, 2008; Ghorbani, Liao, Çayköylü, & Chand, 2013).

It is theoretically suggested that shame compared to guilt involves more devaluation of self (Tangney et al., 2007; Tangney & Dearing, 2003). Nevertheless, in our results no region reached the predetermined threshold when comparing shame to guilt. This result is consistent with some previous observations that shame compared to guilt did not induce higher activity in brain regions involved in self-reference (Pulcu et al., 2014; Wagner et al., 2011). In fact, only one study identified self-referential processing regions (e.g. ACC and mPFC) that activated more for shame than guilt (Michl et al., 2014). Existing results thus suggest that it might be difficult for traditional univariate analysis, which only relies on the BOLD signal of each single voxel, to identify the difference between guilt and shame in the self-referential processing. The activity of the brain (e.g., neuronal firing) is in itself a way to exchange information among multiple neurons (Bray, Chang, & Hoeft, 2009). It has been shown that cognitive tasks could not be completed solely by the neurons within each single voxel (Bray et al., 2009; Fox et al., 2005). The neural information communication among distributed voxels also matters, especially for the complicated cognitive processing. Thus, the analysis method designed to learn spatially distributed patterns of neural activity may decode the neural representation that could not be captured by the univariate analysis (Bray et al., 2009).

Different from univariate analysis that focuses on each signal voxel, MVPA could extract and analyze the information spatially distributed among multiple voxels (Norman et al., 2006). In the present study, similar to the results of univariate analysis, MVPA showed that regions distinguishing guilt and shame were related to theory of mind (TPJ) and cognitive control (vlPFC and dlPFC). Importantly, MVPA additionally found that the multivariate neural patterns of the dmPFC and vACC, which revealed no significant regional-average activation differences in the contrast between guilt and shame, could distinguish guilt and shame. The unique MVPA results could be attributed to the relatively small activation difference of each signal voxel within the dmPFC/vACC cluster between the guilt and shame conditions, but that the activation pattern of multiple distributed voxels within the dmPFC/vACC cluster was different. Since the dmPFC is a region where theory of mind processing and self-referential processing interact (D'Argembeau et al., 2007; Rebecca Saxe et al., 2006), the MVPA results here imply that the dmPFC might put different weights on the theory of mind processing and self-referential processing when participants were in the state of guilt or shame. The vACC is one of the core regions involved in self-referential processing (see a review, Northoff et al., 2006). Different from the function of other self-related regions, such as reappraising self-related stimuli (e.g. dorsal ACC) and linking the self-referential stimuli to one's autobiographical memory (e.g. PCC and precuneus), the vACC relates current external stimuli to oneself and draw one's attention toward one's internal state (Northoff et al., 2006). Yoshimura et al. (2009) found that processing negative self-related stimuli activates vACC. Depressive patients who had a strong negative self-evaluation bias showed a high level of activation in vACC during self-referential processing (Yoshimura et al., 2010, 2014). Although we believe the activity of vACC represented self-referential processing based on the existing theory and the nature of our paradigm, we could not directly exclude the possibility that the vACC activity could reflect other functions, such as self-regulation (Allman, Hakeem, Erwin, Nimchinsky, & Hof, 2001; Fourie et al., 2014). Thus, we suggest the MVPA results of vACC provides preliminary evidence that the self-referential processing of shame is different from that of guilt. Our results of the Shame > Guilt contrast (no significant cluster) and the MVPA together suggested that the difference of guilt and shame in self-referential prcoessing might be reflected in the multi-voxel neural patterns rather than regional-average activity responses of each single voxel in the self-related regions.

An interesting question is that why the information related to guilt and shame in the dmPFC and vACC was represented by the multi-voxel pattern instead of each signal voxel. The dmPFC and vACC are related to the self-referential processing (Feng, Yan, Huang, Han, & Ma, 2018; Northoff et al., 2006). The self-referential processing is a kind of complex high-level cognitive processing, which is closely associated with both self-related and other-related information (Northoff et al., 2006; Schmitz, Kawahara-Baccus, & Johnson, 2004). We assumed that the multi-voxel distributed neural representation might be a more efficient way than the single-voxel isolated neural representation to integrate different types of information. Future studies are needed to demonstrate the assumption.

To the best of our knowledge, our study is the first to directly evoke and compare guilt and shame emotions in an interpersonal context. Our results did not only echo those findings identified in previous recall and imagination paradigms, but also revealed some novel results. While previous studies using the recall and imagination paradigms highlighted the role of the dmPFC in guilt compared to shame (Takahashi et al., 2004; Wagner et al., 2011), our study using the interpersonal paradigm identified the TPJ as an important region. It could be because our paradigm provided a real-time social interaction environment for the participants. The TPJ plays an important role in mentalizing in the social context but not the non-social context (Saxe & Kanwisher, 2003). Besides, the TPJ is responsible for transient mental inference about others (Van Overwalle & Baetens, 2009). Our results showed that the TPJ is a vital region to dissociate interpersonal guilt and shame. Our results did not find regions related to meomory (e.g. hippocamous and parahippocampus), which were repeatly reported in previous studies (Michl et al., 2014; Takahashi et al., 2004). This discrepancy could be owing to the reason that our design excluded some unnecessary psychological process induced by the recall and imagination paradigms, such as memory retrieval and mental imagery.

Differentiating the guilt and shame could provide insights on some psychiatric disorders, such as depression. Patients with depression symptoms are inclined to hold negative self-referential believes and repeatedly devalue themselves (see a review, Disner, Beevers, Haigh, & Beck, 2011). Shame rather than guilt has a strong effect on depression (Orth, Berking, & Burkhardt, 2006; Tangney, Burggraf, & Wagner, 1995). Theoretically, it could be attributed to the reason that shame is more associated with negative self-referential processing than guilt (Tangney & Dearing, 2003). Our study deepened this understanding at the neural level. For instance, the difference in neural activity patterns of the self-referential regions (e.g. vACC and dmPFC) between guilt and shame may explain the unique correlation between shame and depression.

Several limitations of our study should be noted. First, only two emotional words were provided for participants to choose in each trial. Nevertheless, we clearly informed the participants that they did not have to select any affective words if they had no such feelings, and self-reported ratings outside the scanner confirmed that target emotions were successfully induced. Relatedly, embarrassment, an emotion similar to shame, was not measured. The purpose of our study is not to differentiate shame from embarrassment. There are still disputes on whether shame and embarrassment are distinct emotional responses (Haidt, 2003; Kaufman, 2004; Lewis, 1971; Michl et al., 2014; Tangney, Miller, et al., 1996). A key proposed difference between shame and embarrassment is that shame is more associated with the moral violation than embarrassment (Haidt, 2003; Tangney, Miller, et al., 1996). Nevertheless, a recent study showed that violation of moral standards is unnecessary for the experience of shame (Robertson, Sznycer, Delton, Tooby, & Cosmides, 2018); instead, social devaluation is sufficient to evoke shame (Robertson et al., 2018). These findings further blur the boundary between shame and embarrassment. We suggest future studies on guilt and shame to measure participants' feeling of embarrassment (e.g. Fourie, Thomas, Amodio, Warton, & Meintjes, 2014).

Second, guilt and shame were not purely evoked in the guilt and shame conditions respectively, and the absolute difference of the guilt and shame ratings in the guilt and shame conditions was not very large,. These findings are in line with the conjecture that guilt and shame naturally coexist (Tangney & Dearing, 2003)(Michl et al., 2014; Takahashi et al., 2004; Wagner et al., 2011). Nevertheless, the fact that guilt and shame ratings are close in the guilt and shame conditions may make our reported neural results (e.g. the Shame > Guilt contrast) conservative to some extent.

Third, as to stimuli per se, the only difference between the guilt and shame condition was the outcome of the decision. The guilt and shame conditions could be respectively considered as negative and positive feedbacks, as the purpose of the participants was helping the confederate make a right decision. Some may wonder whether the neural activation difference between the guilt and shame conditions was merely caused by the negative and positive feedbacks. Studies on the feedback (prediction error) have provided compelling evidence that a negative feedback compared to a positive feedback increases the activation of midbrain (Aron, 2004) and dorsal anterior cingulate cortex (Bush et al., 2002; Holroyd et al., 2004; Nieuwenhuis, Holroyd, Mol, & Coles, 2004). However, our results did not reveal any significant results in the activity of those regions. It suggests the participants might combine the outcome of the decision with the rules of our study and form high-level cognition (guilt or shame). Besides, researchers have demonstrated that they successfully evoked moral emotions using similar feedback paradigms (Gao et al., 2018; Leng, Wang, Cao, & Li, 2017; Yu, Duan, & Zhou, 2017; Yu et al., 2014; Zhu, Wu, et al., 2017) and explored the corresponding neural correlates (Leng et al., 2017; Yu et al., 2014; Zhu, Wu, et al., 2017).

Fourth, constrained by the paradigm and the usage of fMRI scanner, the ecological validity of our study requires further investigation. For future studies on guilt and shame, there are two ways to improve the ecological validity. One is the virtual reality technique (Patil et al., 2018), and the other is the (portable) near-infrared spectroscopy system, which could be used to study face-to-face real social interaction (Piper et al., 2014; Tang et al., 2015).

In conclusion, using the fMRI technique during an advice-decision task, we evoked guilt and shame in the interpersonal context. Consistent with previous studies, we found that both guilt and shame activated regions related to the integration of theory of mind and self-referential processing (dmPFC) and to the emotional processing (AI). Supporting the theory that guilt involves more theory of mind processing (Tangney & Dearing, 2003), we showed that guilt relative to shame induced more activation in the regions related to theory of mind (supramarginal gyrus and TPJ). Our results also extended the theory by revealing that guilt relative to shame increased neural activity in the OFC/vlPFC and dlPFC, which suggests that guilt involves more cognitive control than shame. Consistent with the results of univariate analysis, the MVPA showed that regions dissociating guilt and shame include those related to theory of mind regions (TPJ) and cognitive control regions (vlPFC and dlPFC). Moreover, the MVPA also found differential neural patterns of the dmPFC and vACC in response to guilt and shame, which indicates that the self-referential processing of guilt and shame might be different. Our findings shed light on the psychological and neural mechanisms of interpersonal guilt and shame.

Funding

This work was supported by the National Key R&D Program of China (2017YFC0803402), the National Natural Science Foundation of China (NSFC) (31871094, 31771206), the Beijing Municipal Science and Technology Commission

(Z151100003915122), the National Program for Support of Top-notch Young Professionals, and the Research Funds of Renmin University of China (15XNLQ05).

Acknowledgements

We thank Yina Ma and Siyang Luo for constructive advice and thank Tao Jin, Huagen Wang and Rui Su for data collection.

References

Aichhorn, M., Perner, J., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Do visual perspective tasks need theory of mind? *Neuroimage*, *30*, 1059–1068.

Allman, J. M., Hakeem, a, Erwin, J. M., Nimchinsky, E., & Hof, P. (2001). The anterior cingulate cortex. The evolution of an interface between emotion and cognition. *Annals of the New York Academy of Sciences*, *935*, 107–117.

Aron, A. R. (2004). Human Midbrain Sensitivity to Cognitive Feedback and Uncertainty During Classification Learning. *Journal of Neurophysiology*, *92*, 1144–1152.

Bastian, B., Jetten, J., & Fasoli, F. (2011). Cleansing the soul by hurting the flesh: the guilt-reducing effect of pain. *Psychological Science*, *22*, 334–5.

Bastin, C., Harrison, B. J., Davey, C. G., Moll, J., & Whittle, S. (2016). Feelings of shame, embarrassment and guilt and their neural correlates: A systematic review. *Neuroscience and Biobehavioral Reviews*, *71*, 455–471.

Baucom, L. B., Wedell, D. H., Wang, J., Blitzer, D. N., & Shinkareva, S. V. (2012).

Decoding the neural representation of affective states. *NeuroImage*, *59*, 718–727.

Bray, S., Chang, C., & Hoeft, F. (2009). Applications of multivariate pattern classification analyses in developmental neuroimaging of healthy and clinical populations. *Frontiers in Human Neuroscience*, *3*, 1–12.

Brown, R., González, R., Zagefka, H., Manzi, J., & Cehajic, S. (2008). Nuestra culpa: collective guilt and shame as predictors of reparation for historical wrongdoing. *Journal of Personality and Social Psychology*, *94*, 75–90.

Bush, G., Vogt, B. A., Holmes, J., Dale, A. M., Greve, D., Jenike, M. A., & Rosen, B. R. (2002). Dorsal anterior cingulate cortex: a role in reward-based decision making. *Proceedings of the National Academy of Sciences*, *99*, 523–528.

Carnì, S., Petrocchi, N., Del Miglio, C., Mancini, F., & Couyoumdjian, A. (2013a). Intrapsychic and interpersonal guilt: a critical review of the recent literature. *Cognitive Processing*, *14*, 333–46.

Carnì, S., Petrocchi, N., Del Miglio, C., Mancini, F., & Couyoumdjian, A. (2013b). Intrapsychic and interpersonal guilt: A critical review of the recent literature. *Cognitive Processing*, *14*, 333–346.

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.

Craig, A. D. (2009). How do you feel—now? the anterior insula and human awareness. *Nature Reviews Neuroscience*, *10*.

Cui, Z., & Gong, G. (2018). The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *NeuroImage*, *178*, 622–637.

D'Argembeau, A., Ruby, P., Collette, F., Degueldre, C., Balteau, E., Luxen, A., … Salmon, E. (2007). Distinct regions of the medial prefrontal cortex are associated with self-referential processing and perspective taking. *Journal of Clinical Neuroscience*, *19*, 935–44.

de Hooge, I. E., Zeelenberg, M., & Breugelmans, S. M. (2010). Restore and protect motivations following shame. *Cognition and Emotion*, *24*, 111–127.

De Hooge, I. E., Zeelenberg, M., & Breugelmans, S. M. (2007). Moral sentiments and cooperation: Differential influences of shame and guilt. *Cognition & Emotion*, *21*, 1025–1042.

Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *Neuroscientist*, *13*, 580–593.

Disner, S. G., Beevers, C. G., Haigh, E. A. P., & Beck, A. T. (2011). Neural mechanisms of the cognitive model of depression. *Nature Reviews Neuroscience*, *12*, 467–477.

Feng, C., Azarian, B., Ma, Y., Feng, X., Wang, L., Luo, Y., & Krueger, F. (2017). Mortality salience reduces the discrimination between in□group and out□group interactions: A functional MRI investigation using multi□voxel pattern analysis. *Human Brain Mapping*, *38*, 1281–1298.

Feng, C., Deshpande, G., Liu, C., Gu, R., Luo, Y. J., & Krueger, F. (2016). Diffusion of responsibility attenuates altruistic punishment: A functional magnetic resonance imaging effective connectivity study. *Human Brain Mapping*, *37*, 663–677.

Feng, C., Luo, Y. J., & Krueger, F. (2015). Neural signatures of fairness-related normative decision making in the ultimatum game: A coordinate-based meta-analysis. *Human Brain Mapping*, *36*, 591–602.

Feng, C., Yan, X., Huang, W., Han, S., & Ma, Y. (2018). Neural representations of the multidimensional self in the cortical midline structures. *NeuroImage*, *183*, 291–299.

Feng, C., Zhu, Z., Gu, R., Wu, X., Luo, Y.-J., & Krueger, F. (2018). Resting-state functional connectivity underlying costly punishment: A machine learning approach. *Neuroscience*.

Finger, E. C., Marsh, A. A., Kamel, N., Mitchell, D. G. V, & Blair, J. R. (2006). Caught in the act: The impact of audience on the neural response to morally and socially inappropriate behavior. *NeuroImage*, *33*, 414–421.

Fourie, M. M., Thomas, K. G. F., Amodio, D. M., Warton, C. M. R., & Meintjes, E. M. (2014). Neural correlates of experienced moral emotion: An fMRI investigation of emotion in response to prejudice feedback. *Social Neuroscience*, *9*, 203–218.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). From The Cover: The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National*

Academy of Sciences, *102*, 9673–9678.

Gao, X., Yu, H., Sáez, I., Blue, P. R., Zhu, L., Hsu, M., & Zhou, X. (2018). Distinguishing neural correlates of context-dependent advantageous- and disadvantageous-inequity aversion. *Proceedings of the National Academy of Sciences*, 201802523.

Gausel, N., & Leach, C. W. (2011). Concern for self-image and social image in the management of moral failure: Rethinking shame. *European Journal of Social Psychology*, *41*, 468–478.

Ghorbani, M., Liao, Y., Çayköylü, S., & Chand, M. (2013). Guilt, Shame, and Reparative Behavior: The Effect of Psychological Proximity. *Journal of Business Ethics*, *114*, 311–323.

Gifuni, A. J., Kendal, A., & Jollant, F. (2016). Neural mapping of guilt: a quantitative meta-analysis of functional imaging studies. *Brain Imaging and Behavior*, 1–15.

Gunther Moor, B., Güroğlu, B., Op de Macks, Z. A., Rombouts, S. A. R. B., van der Molen, M. W., & Crone, E. A. (2012). Social exclusion and punishment of excluders: Neural correlates and developmental trajectories. *NeuroImage*, *59*, 708–717.

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 852–870). Oxford: Oxford University Press.

Hebart, M. N., Görgen, K., Haynes, J.-D., & Dubois, J. (2015). The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics*, *8*, 1–18.

Hoffman, M. L. (1982). Development of Prosocial Motivation: Empathy and Guilt. In N. Eisenberg (Ed.), *Development of prosocial motivation: empathy and guilt* (pp. 281–313). San Diego: Academic Press.

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R. B., Coles, M. G. H., & Cohen, J. D. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. *Nature Neuroscience*, *7*, 497–498.

Howell, A. J., Turowski, J. B., & Buro, K. (2012). Guilt, empathy, and apology. *Personality and Individual Differences, 53*, 917–922.

Kaufman, G. (2004). *The psychology of shame: Theory and treatment of shame-based syndromes*. Springer Publishing Company.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, *314*, 829–832.

Knoch, D., Schneider, F., Schunk, D., Hohmann, M., & Fehr, E. (2009). Disrupting the prefrontal cortex diminishes the human ability to build a good reputation. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 20895–20899.

Koechlin, E. (2003). The Architecture of Cognitive Control in the Human Prefrontal Cortex. *Science*, *302*, 1181–1185.

Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, *10*, 603–621.

Leng, B., Wang, X., Cao, B., & Li, F. (2017). Frontal negativity: An electrophysiological index of interpersonal guilt. *Social Neuroscience*, *12*, 649–660.

Lewis, H. B. (1971). Shame and guilt in neurosis. *Psychoanalytic Review*, *58*, 419.

Lindquist, K. A., & Barrett, L. F. (2012). A functional architecture of the human brain: Emerging insights from the science of emotion. *Trends in Cognitive Sciences*, *16*, 533–540.

Mclatchie, N., Giner-Sorolla, R., & Derbyshire, S. W. G. (2016). 'Imagined guilt' vs 'recollected guilt': implications for fMRI. *Social Cognitive and Affective Neuroscience*, *11*, 703–711.

Michl, P., Meindl, T., Meister, F., Born, C., Engel, R. R., Reiser, M., & Hennig-Fast, K. (2014). Neurobiological underpinnings of shame and guilt: A pilot fMRI study. *Social Cognitive and Affective Neuroscience*, *9*, 150–157.

Moll, J., de Oliveira-Souza, R., Garrido, G. J., Bramati, I. E., Caparelli-Daquer, E. M. a, Paiva, M. L. M. F., … Grafman, J. (2007). The self as a moral agent: linking the neural bases of social agency and moral sensitivity. *Social Neuroscience*, *2*, 336–352.

Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI - An introductory guide. *Social Cognitive and Affective Neuroscience*, *4*, 101–109.

Muris, P. (2015). Guilt, Shame, and Psychopathology in Children and Adolescents.

Child Psychiatry and Human Development, *46*, 177–179.

Nelissen, R. M. A. (2014). Relational utility as a moderator of guilt in social interactions. *Journal of Personality and Social Psychology*, *106*, 257–271.

Nelissen, R. M. a, & Zeelenberg, M. (2009). When guilt evokes self-punishment: Evidence for the existence of a Dobby Effect. *Emotion*, *9*, 118–122.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric Permutation Tests for {PET} functional Neuroimaging Experiments: A Primer with examples. *Human Brain Mapping*, *15*, 1–25.

Nieuwenhuis, S., Holroyd, C. B., Mol, N., & Coles, M. G. H. (2004). Reinforcement-related brain potentials from medial frontal cortex: origins and functional significance. *Neuroscience & Biobehavioral Reviews*, *28*, 441–448.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*, 424–430.

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain-A meta-analysis of imaging studies on the self. *NeuroImage*, *31*, 440–457.

Ohtsubo, Y., & Yagi, A. (2015). Relationship value promotes costly apology-making: Testing the valuable relationships hypothesis from the perpetrator's perspective. *Evolution and Human Behavior*, *36*, 232–239.

Orth, U., Berking, M., & Burkhardt, S. (2006). Self-conscious emotions and depression: Rumination explains why shame but not guilt is maladaptive.

Personality and Social Psychology Bulletin, *32*, 1608–1619.

Patil, I., Zanon, M., Novembre, G., Zangrando, N., Chittaro, L., & Silani, G. (2018). Neuroanatomical basis of concern-based altruism in virtual environment. *Neuropsychologia*, *116*, 34–43.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, *45*, 199–209.

Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002). Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage*, *16*, 331–348.

Piper, S. K., Krueger, A., Koch, S. P., Mehnert, J., Habermehl, C., Steinbrink, J., … Schmitz, C. H. (2014). A wearable multi-channel fNIRS system for brain imaging in freely moving subjects. *NeuroImage*, *85*, 64–71.

Pulcu, E., Lythe, K., Elliott, R., Green, S., Moll, J., Deakin, J. F. W., & Zahn, R. (2014). Increased amygdala response to shame in remitted major depressive disorder. *PLoS ONE*, *9*, 1–9.

Riva, P., Romero Lauro, L. J., DeWall, C. N., Chester, D. S., & Bushman, B. J. (2014). Reducing aggressive responses to social exclusion using transcranial direct current stimulation. *Social Cognitive and Affective Neuroscience*, 352–356.

Robertson, T. E., Sznycer, D., Delton, A. W., Tooby, J., & Cosmides, L. (2018). The true trigger of shame: social devaluation is sufficient, wrongdoing is unnecessary. *Evolution and Human Behavior*, *39*, 566–573.

Roth, L., Kaffenberger, T., Herwig, U., & Bruehl, A. B. (2014). Brain activation associated with pride and shame. *Neuropsychobiology, 69*, 95–106.

Saarimäki, H., Gotsopoulos, A., Jääskeläinen, I. P., Lampinen, J., Vuilleumier, P., Hari, R., … Nummenmaa, L. (2016). Discrete Neural Signatures of Basic Emotions. *Cerebral Cortex*, *26*, 2563–2573.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind." *Neuroimage*, *19*, 1835–1842.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*, 1835–1842.

Saxe, R., Moran, J. M., Scholz, J., & Gabrieli, J. (2006). Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Social Cognitive and Affective Neuroscience*, *1*, 229–234.

Schmitz, T. W., Kawahara-Baccus, T. N., & Johnson, S. C. (2004). Metacognitive evaluation, self-relevance, and the right prefrontal cortex. *NeuroImage*, *22*, 941–947.

Schuck, N. W., Gaschler, R., Wenke, D., Heinzle, J., Frensch, P. A., Haynes, J. D., & Reverberi, C. (2015). Medial prefrontal cortex predicts internally driven strategy shifts. *Neuron*, *86*, 331–340.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, *42*, 9–34.

Seara-Cardoso, A., Sebastian, C. L., McCrory, E., Foulkes, L., Buon, M., Roiser, J. P.,

& Viding, E. (2016). Anticipation of guilt for everyday moral transgressions: The role of the anterior insula and the influence of interpersonal psychopathic traits. *Scientific Reports, 6,* 36273.

Shin, L. M., Dougherty, D. D., Orr, S. P., Pitman, R. K., Lasko, M., MacKlin, M. L., … Rauch, S. L. (2000). Activation of anterior paralimbic structures during guilt-related script-driven imagery. *Biological Psychiatry*, *48*, 43–50.

Singer, T., Seymour, B., O'Doherty, J., Dolan, R. J., Kaube, H., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science (New York, N.Y.)*, *303*, 1157–62.

Smith, R. H., Webster, J. M., Parrott, W. G., & Eyre, H. L. (2002). The role of public exposure in moral and nonmoral shame and guilt. *Journal of Personality and Social Psychology*, *83*, 138.

Strang, S., Gross, J., Schuhmann, T., Riedl, A., Weber, B., & Sack, A. T. (2015). Be nice if you have to--the neurobiological roots of strategic fairness. *Social Cognitive and Affective Neuroscience*, *10*, 790–6.

Sznycer, D., Tooby, J., Cosmides, L., Porat, R., Shalvi, S., & Halperin, E. (2016). Shame closely tracks the threat of devaluation by others, even across cultures. *Proceedings of the National Academy of Sciences*, *113*, 2625–2630.

Takahashi, H., Yahata, N., Koeda, M., Matsuda, T., Asai, K., & Okubo, Y. (2004). Brain activation associated with evaluative processes of guilt and embarrassment: An fMRI study. *NeuroImage, 23*, 967–974.

Tang, H., Mai, X., Wang, S., Zhu, C., Krueger, F., & Liu, C. (2015). Interpersonal brain synchronization in the right temporo-parietal junction during face-to-face economic exchange. *Social Cognitive and Affective Neuroscience, 11*, 23–32.

Tangney, J. P. (1995). Recent advances in the empirical study of shame and guilt. *American Behavioral Scientist*, *38*, 1132–1145.

Tangney, J. P. (1996). Conceptual and methodological issues in the assessment of shame and guilt. *Behav. Res. Ther.*, *34*, 741–754.

Tangney, J. P., Burggraf, S. A., & Wagner, P. E. (1995). Shame-proneness, guilt-proneness, and psychological symptoms.

Tangney, J. P., & Dearing, R. L. (2003). Shame and guilt. New York: Guilford Press.

Tangney, J. P., Miller, R. S., Flicker, L., & Barlow, D. H. (1996). Are Shame, Guilt, and Embarrassment Distinct Emotions? *Journal of Personality and Social Psychology*, *70*, 1256–1269.

Tangney, J. P., Stuewig, J., Mashek, D., & Hastings, M. (2011). Assessing Jail Inmates' Proneness to Shame and Guilt: Feeling Bad About the Behavior or the Self? *Criminal Justice and Behavior*, *38*, 710–734.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology*, *58*, 345–72.

Tangney, J. P., Wagner, P. E., Hill-Barlow, D., Marschall, D. E., & Gramzow, R. (1996). Relation of shame and guilt to constructive versus destructive responses to anger across the lifespan. *Journal of Personality and Social Psychology*, *70*, 797–809.

Tracy, J. L., & Robins, R. W. (2006). Appraisal antecedents of shame and guilt:

Aupport for a theoretical model. *Personality and Social Psychology Bulletin*, *32*, 1339–51.

Ty, A., Mitchell, D. G. V, & Finger, E. (2017). Making amends: Neural systems supporting donation decisions prompting guilt and restitution. *Personality and Individual Differences*, *107*, 28–36.

Uddin, L. Q. (2015). Salience processing and insular cortical function and dysfunction. *Nature Reviews Neuroscience*, *16*, 55–61.

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage*, *48*, 564–584.

Vytal, K., & Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. *Journal of Cognitive Neuroscience*, *22*, 2864–2885.

Wagner, U., N'Diaye, K., Ethofer, T., & Vuilleumier, P. (2011). Guilt-Specific Processing in the Prefrontal Cortex. *Cerebral Cortex*, *21*, 2461–2470.

Woo, C.-W., Krishnan, A., & Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *Neuroimage*, *91*, 412–419.

Yoshimura, S., Okamoto, Y., Onoda, K., Matsunaga, M., Okada, G., Kunisato, Y., … Yamawaki, S. (2014). Cognitive behavioral therapy for depression changes medial prefrontal and ventral anterior cingulate cortex activity associated with self-referential processing. *Social Cognitive and Affective Neuroscience*, *9*, 487–493.

Yoshimura, S., Okamoto, Y., Onoda, K., Matsunaga, M., Ueda, K., Suzuki, S. ichi, &

ShigetoYamawaki. (2010). Rostral anterior cingulate cortex activity mediates the relationship between the depressive symptoms and the medial prefrontal cortex activity. *Journal of Affective Disorders*, *122*, 76–85.

Yoshimura, S., Ueda, K., Suzuki, S. ichi, Onoda, K., Okamoto, Y., & Yamawaki, S. (2009). Self-referential processing of negative stimuli within the ventral anterior cingulate gyrus and right amygdala. *Brain and Cognition*, *69*, 218–225.

Yu, H., Cai, Q., Shen, B., Gao, X., & Zhou, X. (2016). Neural substrates and social consequences of interpersonal gratitude: Intention matters. *Emotion*, *17*, 589–601.

Yu, H., Duan, Y., & Zhou, X. (2017). Guilt in the eyes: Eye movement and physiological evidence for guilt-induced social avoidance. *Journal of Experimental Social Psychology*, *71*, 128–137.

Yu, H., Hu, J., Hu, L., & Zhou, X. (2014). The voice of conscience: neural bases of interpersonal guilt and compensation. *Social Cognitive and Affective Neuroscience*, *9*, 1150–1158.

Zhu, R., Jin, T., Shen, X., Zhang, S., Mai, X., & Liu, C. (2017). Relational utility affects self-punishment in direct and indirect reciprocity situations. *Social Psychology*, *48*, 19–27.

Zhu, R., Wu, H., Xu, Z., Tang, H., Shen, X., Mai, X., & Liu, C. (2017). Early distinction between shame and guilt processing in an interpersonal context. *Social Neuroscience*, 1–14.

Zhu, R., Xu, Z., Tang, H., Liu, J., Wang, H., An, Y., … Liu, C. (2018). The effect of shame on anger at others: awareness of the emotion-causing events matters. *Cognition and Emotion*, 1–13.