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

Ruida Zhu a,b,c, Chunliang Feng a,b,c, Shen Zhang a,b,c, Xiaoqin Mai d,* , Chao Liu a,b,c,**

a State Key Laboratory of Cognitive Neuroscience and Learning & IDG, McGovern Institute for Brain Research, Beijing Normal University, 100875, Beijing, China
b Center for Collaboration and Innovation in Brain and Learning Sciences, Beijing Normal University, 100875, Beijing, China
c Beijing Key Laboratory of Brain Imaging and Connectomics, Beijing Normal University, 100875, Beijing, China
d Department of Psychology, Renmin University of China, 100872, Beijing, China

ARTICLE INFO

Keywords:
Guilt
Shame
Theory of mind
Cognitive control
Self-evaluation

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.

1. 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 and 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, Marshall and Gramzow, 1996b). 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 et al., 2016; Tangney and Dearing, 2003), so the capability of mentalizing and having a sense of self are thus required for these two emotions (Tangney and Dearing, 2003). In addition, guilt and shame are negative emotions, which evoke strong aversive feelings and psychological pain (Carni, Petrocchi, Miglio, Mancini, & Couyoumdjian, 2013; Tangney and 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 et al., 2011; Nelissen and 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 et al., 2006; Michl et al., 2014; Moll et al., 2007; Takahashi et al., 2004; Wagner, NDiaye,
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 and Dearing, 2003). Different foci often lead to different psychological processes and behavioral patterns. Compared with shame, guilt involves more other-oriented empathy (Tangney et al., 2007; Tangney and Dearing, 2003). It is not clear whether guilt involves more cognitive empathy (understanding 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 and Breugelmans, 2007; Howell et al., 2012; Yu et al., 2014; Zhu, Jin, et al., 2017a). 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, 2011; Tangney and Dearing, 2003; Zhu et al., 2018), which causes image-reparation behaviors such as withdrawal, hiding (avoiding by being directly criticized) and improvement of themselves (de Hooge et al., 2010; Gausel and 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 temporal 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 frontal gyrus, PCC, and parahippocampus. Pulcu et al. (2014) found shame compared to guilt increased activation in the amygdala and posterior 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 (dPFC)), 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).

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). Previous studies using univariate analysis methods found many brain regions activated similarly during different basic emotional states (Lindquist and Barrett, 2012; Pan et al., 2002; Vytal and Hamann, 2010) and the corresponding meta-analyses had difficulty in establishing unique neural correlates for different basic emotions (Lindquist and 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 et al., 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 and 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.

2. Methods

2.1. 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.

1 An example is presented in the supplementary materials to conceptually explain the difference between univariate activation analysis and multivariate pattern analysis.
2.2. 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 interpersonal guilt, Yu et al., 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 (Fig. 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.

2.3. 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 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 et al., 2013; Tangney and Dearing, 2003; Tracy and Robins, 2006). In 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 (Smith et al., 2002; Tangney and Dearing, 2003; Tracy and 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.

2.4. 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.

2.5. Image acquisition

Images were acquired on a 3 T S 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 x 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 x 256 mm).

2.6. fMRI data analysis

2.6.1. 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 x 3 x 3 mm³), 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.

2.6.2. Univariate activation analysis

At the individual level, we modeled the dots, the advice, the outcome, the advice of the other, 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). 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 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 et al., 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.

2.6.3. Multivariate pattern analysis

MPVA 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/ttdiencodingtoolbox/) (Hebart et al., 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 and Gong, 2018; Ledoit and Wolf, 2003; Mur et al., 2009). Considering many recent studies have demonstrated the reliability of SVM (Feng et al., 2016, 2017; Feng et al., 2018; Yu et al., 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 x 3 x 3 mm³), 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 5000 permutations. The resulting voxels were assessed for significance at 5% level with voxel-wise FWE correction, as determined by permuted datasets (see Nichols and 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.

3. Results

3.1. 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, η² p = .60) and “no response” (M = 0.07, SD = 0.25; F(1, 29) = 391.98, p < .001, η² p = .93), and the post-task ratings of guilt were significantly higher than the ratings of other emotions, all Fs > 27.98, all ps < .001, all η² ps > .49 (Fig. 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, η² p = .21) and “no response” (M = 0.03, SD = 0.18; F(1, 29) = 193.04, p < .001, η² p = .87), and the post-task ratings of shame were significantly higher than the ratings of other emotions, all Fs > 9.75, all ps < .004, all η² ps > .25. The guilt ratings were higher in the guilt than shame condition (F(1, 29) = 29.73, p < .001, η² p = .51) and the shame ratings were higher in the shame than guilt condition (F(1, 29) = 21.86, p < .001, η² p = .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, η² p = .67) and “no response” (M = 0.03, SD = 0.18; F(1, 29) = 437.64, p < .001, η² p = .94), and the ratings of happiness were significantly higher than the ratings of other emotions, all Fs > 21.12, all ps < .001, all η² ps > .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, η² p = .093), between the shame ratings in the guilt condition and the guilt ratings in the shame condition (F(1, 29) = 0.183, p = .672, η² p = .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) = 0.3.38, p = .706, η² p = .105). There was no significant difference in sadness, anger, happiness, or pride ratings between guilt and shame conditions either (all Fs < 2.98, all ps > .095, all η² ps < .093). The results demonstrated that the emotion intensity of participants’ guilt, shame, and happiness choices respectively in the guilt, shame and happiness conditions did not significantly change across three sessions (all Fs < 3.12, all ps > .051, all η² ps < .10), which implied that the target emotion in each condition was stable across three sessions (Figure S4).

2 In the main manuscript, we defined the guilt and shame conditions based on the outcomes (e.g., the participant's advice and the decision'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.
3.2. Neuroimaging results

3.2.1. 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 Fig. 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).

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 Fig. 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.

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 Fig. 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.

4. 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 and 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 et al., 2011; Ty et al., 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 and Dearing, 2003). Guilt but not shame promotes relationship-reparation behavior further implying that transgressors in guilt may have under-stood 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 and Yagi, 2015; Zhu et al., 2017a, b). 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 et al., 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 et al., 2006; Saxe and Kanwisher, 2003; Schurz et al., 2014), while the anterior portion of the TPJ is engaged in the attention orientation (Decety and Lamm, 2007; Lindquist and 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 et al., 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 incure punishment and relationship damage (Knoch et al., 2006; Knoch et al., 2009; Riva et al., 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 et al., 2006; Ghoshbani et al., 2013).

It is theoretically suggested that shame compared to guilt involves more devaluation of self (Tangney et al., 2007; Tangney and 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 et al., 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 (vIPFC and dIPFC). 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 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 et al., 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 processing 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 distributed neural representation (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 et al., 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 processing might be reflected in the multi-voxel neural patterns rather than regional-average activity responses of each single voxel in the self-related regions.
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 and Kanwisher, 2003). Besides, the TPJ is responsible for transient mental inference about others (Van Overwalle and 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 memory (e.g. hippocampus and parahippocampus), which were repeatedly reported in previous studies (Michl et al., 2014; Taka-hashi 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 et al., 2011). Shame rather than guilt has a strong effect on depression (Orth et al., 2006; Tangney et al., 1995). Theoretically, it could be attributed to the reason that shame is more associated with negative self-referential processing than guilt (Tangney and 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

Table 3

<table>
<thead>
<tr>
<th>Region</th>
<th>BA</th>
<th>MNI coordinates</th>
<th>T</th>
<th>Voxels</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Guilt &gt; Shame</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R vlPFC/OFC</td>
<td>11/10</td>
<td>30 54 6</td>
<td>5.71</td>
<td>349</td>
</tr>
<tr>
<td>R dorsolateral prefrontal cortex</td>
<td>45</td>
<td>45 33 24</td>
<td>5.44</td>
<td>84</td>
</tr>
<tr>
<td>L supramarginal gyrus/postcentral gyrus</td>
<td>40/2</td>
<td>−57 −21 30</td>
<td>5.20</td>
<td>109</td>
</tr>
<tr>
<td>R temporo-parietal junction</td>
<td>40/39</td>
<td>54 −51 33</td>
<td>4.73</td>
<td>72</td>
</tr>
</tbody>
</table>

**Shame > Guilt**

None.
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., 1996a). 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., 1996a). Nevertheless, a recent study showed that violation of moral standards is unnecessary for the experience of shame (Robertson et al., 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

### Table 4

Results of the multivariate analysis (p < .05, voxel-level with FWE correction, as determined by permutation distribution with 5000 permutations, cluster size > 10). L, left; R, right; dmPFC, dorsomedial prefrontal cortex; vACC, ventral anterior cingulate cortex.

<table>
<thead>
<tr>
<th>Region</th>
<th>BA</th>
<th>MNI coordinates</th>
<th>T score</th>
<th>Voxels</th>
</tr>
</thead>
<tbody>
<tr>
<td>L/R dmPFC</td>
<td>10/9</td>
<td>3     51       21</td>
<td>6.87</td>
<td>517</td>
</tr>
<tr>
<td>L/R vACC</td>
<td>32</td>
<td>0     48       6</td>
<td>5.18</td>
<td></td>
</tr>
<tr>
<td>R ventrolateral prefrontal cortex</td>
<td>45</td>
<td>42   18       12</td>
<td>5.56</td>
<td>11</td>
</tr>
<tr>
<td>L dorsolateral prefrontal cortex</td>
<td>8/6</td>
<td>−30  3        45</td>
<td>6.59</td>
<td>76</td>
</tr>
<tr>
<td>R temporo-parietal junction</td>
<td>40/39</td>
<td>57    −51      30</td>
<td>5.14</td>
<td>18</td>
</tr>
</tbody>
</table>

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., 1996a). 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., 1996a). Nevertheless, a recent study showed that violation of moral standards is unnecessary for the experience of shame (Robertson et al., 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 et al., 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 and 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 et al., 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 et al., 2017; Yu et al., 2017; Yu et al., 2014; Zhu, Wu, et al., 2017b) and explored the corresponding neural correlates (Leng et al., 2017; Yu et al., 2014; Zhu, Wu, et al., 2017b).

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 and 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 (supramarginal gyrus and TPJ). 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 (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...
Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2018.11.012.

References


