:10.1093/ a / . 090 SCAN (2014) 9, 1150–1158

The voice of conscience: neural bases of interpersonal guilt and compensation

Hongbo Yu,[1] Jie Hu,[1] Li Hu,[2] and Xiaolin Zhou[1,3,4]

[1]Center for Brain and Cognitive Sciences and Department of Psychology, Peking University, Beijing 100871, [2]Key Laboratory of Cognition and Personality (Ministry of Education) and School of Psychology, Southwest University, Chongqing 400715, [3]Key Laboratory of Machine Perception (Ministry of Education), and [4]PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, People's Republic of China

People feel bad for inflicting harms upon others; this emotional state is termed interpersonal guilt. In this study, the participant played multiple rounds of a dot-estimation task with anonymous partners while undergoing fMRI. The partner would receive pain stimulation if the partner or the participant or both responded incorrectly; the participant was then given the option to intervene and bear a proportion of pain for the partner. The level of pain voluntarily taken and the activations in anterior middle cingulate cortex (aMCC) and bilateral anterior insula (AI) were higher when the participant was solely responsible for the stimulation (Self_Incorrect) than when both committed an error (Both_Incorrect). Moreover, the gray matter volume in the aMCC predicted the individual's compensation behavior, measured as the difference between the level of pain taken in the Self_Incorrect and Both_Incorrect conditions. Furthermore, a mediation pathway analysis revealed that activation in a midbrain region mediated the relationship between aMCC activation and the individual's tendency to compensate. These results demonstrate that the aMCC and the midbrain nucleus not only play an important role in experiencing interpersonal guilt, but also contribute to compensation behavior.

Keywords: interpersonal guilt; anterior middle cingulate; fMRI; gray matter volume; mediation pathway analysis

INTRODUCTION

How would you feel if you lost the bicycle borrowed from your friend, which was the last present given to him by his grandmother before she died (de Hooge *et al.*, 2011)? This is an example of interpersonal guilt. In philosophy, guilt is understood as 'the moral feeling produced by conscience, itself the internalized voice of moral authority' (Griswold, 2007). The societal significance of guilt is wide and extensive (Baumeister *et al.*, 1994). It functions as a moral emotion, protecting and enhancing social relationships by punishing interpersonal wrongdoings and restoring equities (Baumeister *et al.*, 1994; Haidt, 2003). Moreover, the prospect of guilt prevents people from committing wrongful deeds (Chang *et al.*, 2011); the lack of guilt is a characteristic manifestation of psychopaths, who have normal moral knowledge but behave abnormally immoral (Blair, 2006; Kiehl, 2006).

Over the last decade, several neuroimaging studies have been carried out to investigate the neural mechanisms underlying the processing of guilt. These studies predominately used imagination or recall of a guilt-related situation as emotion-inducing stimuli (Shin *et al.*, 2000; Takahashi *et al.*, 2004; Berthoz *et al.*, 2006; Zahn *et al.*, 2009; Basile *et al.*, 2011). Since the script-based imagination and recall may require psychological processes that are nonessential to the experience of guilt, the results of these studies are mixed. Nonetheless, several brain regions have been consistently implicated and the activation of these regions have also been observed for the experience of negative affect, physical pain, and 'social pain' (Shackman *et al.*, 2011; Eisenberger, 2012). For instance, anticipating and imagining a guilt-evoking situation activate the anterior cingulate cortex (ACC), the anterior insula (AI) and the lateral orbitofrontal cortex (LOFC) (Shin *et al.*, 2000; Basile *et al.*, 2011; Chang *et al.*, 2011; Wagner *et al.*, 2011). These activations may reflect an 'unpleasant arousal akin to anxiety' (Tennen and Herzberger, 1987), such as the anxiety over being socially excluded (Baumeister *et al.*, 1994). This anxiety may in turn promote compensation and prosocial behaviors, thereby restoring the impaired social relationship (Griswold, 2007).

While imagining or recalling particular situations may, in fact, be successful in eliciting guilt-like feelings, the every-day experience of guilt is interpersonal and complex and needs to be analyzed in these more realistic settings (Baumeister *et al.*, 1994; Sanfey, 2007). Moreover, interpersonal guilt, as a type of moral emotion, or 'the voice of conscience', may stimulate moral behavior, such as compensation (Baumeister *et al.*, 1994; Griswold, 2007), the neural basis of which has not been examined in previous studies. In this study, we address three previously unanswered questions (but see Koban *et al.*, in press): (i) What are the brain responses to interpersonal guilt? (ii) What is the neural pathway through which the initial cortical response to interpersonal guilt is translated into compensation behaviors? (iii) What is the structural basis for guilt and compensation behavior?

We combined functional/structural MRI technique and an interpersonal game paradigm to investigate these three questions. While undergoing fMRI scanning, the participant was playing a dot-estimation task (Fliessbach *et al.*, 2007) with an anonymous partner. In each round of the game the partner, a confederate, was randomly chosen from three possible candidates. The participant underwent two scanning sessions. In the first session (Figure 1), the participant was told that a painful stimulation could befall their partner if one or both of them estimated incorrectly. There were four possible outcomes: both estimated correctly (Both_Correct), only the partner estimated incorrectly (Partner_Incorrect), only the participant estimated incorrectly (Self_Incorrect), and both estimated incorrectly (Both_Incorrect) (Table 1). Before stimulation, the participant indicated the level of pain he or she would be willing to take for the partner, which was used as an online measure of the level of guilt and compensation (Batson *et al.*, 1981; de Hooge *et al.*, 2011). We hypothesized that in the Self_Incorrect condition, where the participant's mistake is the

Received 23 November 2012; Revised 2 May 2013; Accepted 4 June 2013
Advance Access publication 26 July 2013

The authors thank Li Zhang, Chen Zhao, and Ping Xiao for technical support during fMRI data acquisition and thank Dr. Tor Wager for his help in fMRI data analysis. The authors are also grateful to the two anonymous reviewers for their constructive suggestions concerning the revision of the manuscript.

This study was supported by National Basic Research Program from the Ministry of Science and Technology of China (973 Program: 2010CB833904) and by grants from the National Natural Science Foundation of China (30110972, 91232708).

Correspondence should be addressed to Xiaolin Zhou, Department of Psychology, Peking University, 5 Yiheyuan Road, Beijing 100871, People's Republic of China. E-mail: xz104@pku.edu.cn

direct cause of the partner's suffering, the feeling of interpersonal guilt will be more intense than the Both_Incorrect and the Partner_Incorrect conditions. Therefore, the compensation behavior, i.e. taking painful stimulation for the partner, will be higher in the Self_Incorrect condition as compared with the other two conditions. In the second session, the participant was told that he/she would perform the same task with the same three partners. No pain stimulation was delivered to either side. This session was included to control for potential confounding factors such as social comparison (see below). After scanning, the participant was asked to rate separately, on a 9-point scale, their feelings of guilt, distress, anger and fear in the three pain-present conditions in the first session.

We first used a conventional general linear model-based analysis to identify guilt-related activations and brain correlates of individual's sensitivity to guilt. Then to identify the neural pathways through which the brain responses to guilt are translated to behavioral responses, we utilized a recently developed procedure to test mediation relationship and to locate multiple brain mediators (Mediation Effect Parametric Mapping, MEPM; Wager *et al.*, 2008, 2009a). With this procedure, Wager *et al.* (2009a) found that periaqueductal gray (PAG), a midbrain nucleus, mediated the brain processes of social threat and the physiological responses to the threat (see also Buhle *et al.*, in press). This finding highlights the importance of the cortical–subcortical interaction in the translation of brain processes of social-affective stimuli into physiological and experiential responses. If interpersonal guilt can be conceptualized as anxiety and social threat, as argued above, then it is possible that the midbrain nucleus mediates the relation between the cortical affective processing of interpersonal guilt and the behavioral responses to the guilt.

MATERIALS AND METHODS

Participants

Twenty-seven healthy right-handed graduate and undergraduate students took part in the fMRI scanning. Because of excessive head movements (>3 mm), 3 were excluded from data analysis, leaving 24 participants (mean age 22.0 years; age range: 19–24 years; 11 female) for data analysis. None of the participants reported any history of psychiatric, neurological or cognitive disorders. Informed written consent was obtained from each participant before scanning. The study was carried out in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Department of Psychology, Peking University.

Procedure

Each participant came to the scanning room individually. Upon arrival he/she met three confederates and was told that they would later play an interactive game together through an intranet, but in separate rooms. At least one confederate had the same sex as the participant and at least one had the opposite sex. An intra-epidermal needle electrode was attached to the left wrist of the participant for cutaneous electrical stimulation (Inui *et al.*, 2002). Participant-specific pain threshold was calibrated and three levels of pain stimulation were set as 1, 3 and 15 repeated pulses (with 0.5 ms duration of each pulse and a 10 ms interval between consecutive pulses) of epidermal electrical stimulation. The intensity of each pulse (in the unit of mA) was four times of the participant's pain threshold. The participant was then asked to rate the intensity of three levels of pain stimulation on a scale of 0 ('not painful') to 10 ('unbearable'). The mean intensity ratings were (mean $\pm$ SD) 2.5 ± 1.0, 5.1 ± 1.5 and 8.2 ± 1.5 for the low, medium and high intensity, respectively. All participants reported that the three levels of pain stimulation were clearly distinguishable. In the first scanning session, the participant was to perform the task described in Figure 1. There were four possible outcomes: both estimated correctly (Both_Correct), only the partner estimated incorrectly (Partner_Incorrect), only the participant estimated incorrectly (Self_Incorrect) and both estimated incorrectly (Both_Incorrect) (Table 1). The first fMRI scanning session consisted of 64 trials (16 for each experimental condition) and lasted for $\sim$25 min. Note that the estimation outcome (correct *vs* incorrect) was predetermined by a computer program such that each condition had an equal number of trials. This is for maximizing the fMRI statistical power and a postscan interview showed that no participant actually noticed this manipulation. The second scanning session was similar to the first except that no pain stimulation was delivered. The second scanning session consisted of 64 trials and lasted for about 17 min. After scanning, each s

Data acquisition

Images were acquired using a Siemens 3.0 Tesla Trio scanner with a standard head coil at the Key Laboratory of Cognition and Personality (Ministry of Education) of Southwest University, China. T2*-weighted functional images were acquired in 36 axial slices parallel to the AC–PC line with no interslice gap, affording full-brain coverage. Images were acquired using an EPI pulse sequence, with a TR of 2200 ms, a TE of 30 ms, a flip angle of 90°, an FOV of 220 mm × 220 mm and 3.4 mm × 3.4 mm × 3.5 mm voxels. A high-resolution, whole-brain structural scan (1 mm^3 isotropic voxel MPRAGE) was acquired after functional imaging.

GLM-based image analysis

Image preprocessing and analysis used the Statistical Parametric Mapping software SPM8 (Wellcome Trust Department of Cognitive Neurology, London, UK). Images were slice-time corrected, motion corrected, re-sampled to 2 × 2 × 2 isotropic voxel, normalized to Montreal Neurological Institute (MNI) space and spatially smoothed using an 8-mm FWHM Gaussian filter, and temporally filtered using a high-pass filter with 1/128 Hz cutoff frequency. The first-level (within-participant) statistical analysis was conducted with SPM8. Briefly, separate regressors in the GLM were specified for fMRI responses to the recasting cue, random dot presentation, estimation responses, outcome feedback, costly helping responses and pain delivery. Values for the 'Self_Incorrect > Both_Incorrect' contrast at the outcome stage were subjected to second-level random effects analysis using the one-sample *t* test in SPM8. The same contrast in the second session was defined as an exclusive mask ($P < 0.05$ uncorrected) for the first session results.

A robust regression analysis (Wager *et al.*, 2005) was carried out on the same contrast to locate, in an anatomically defined midbrain region of interest (ROI), regions in which activity correlated with the tendency to compensate (i.e. the Compensation Index; see 'Results: Behavioral result' section). In addition, we defined a parametric model to examine the neural correlate of intra-participant variation in the level of pain chosen. A new GLM was built in which the three error-containing conditions were combined in one regressor. This regressor was further modulated by a parametric regressor that contained trial-wise level of pain chosen by the participant.

Mediation pathway analysis

The MEPM analysis is based on a standard three-variable path model with a bootstrap test for the statistical significance of the product *a*b* (Wager *et al.*, 2008, 2009a). This analysis was conducted on 'Self_Incorrect *vs* Both_Incorrect' contrast values. For the ROI analysis in aMCC, we extracted the parameter estimates from a 6-mm edge cube around the peak voxel of the aMCC and of the midbrain revealed by the above second-level contrasts and used them as mediator and predictor in a mediation model. The Compensation Index was the dependent variable in this model. For the whole-brain exploratory analysis, the chosen threshold of $P < 0.017$ controlled the false positive rate (FDR) < 0.05 corrected in the whole brain.

Voxel-based morphometry analysis

The voxel-based morphometry (VBM) was conducted using the VBM toolbox developed by Christian Gaser (University of Jena, Department of Psychiatry, freely available online http://dbm.neuro.uni-jena.de/vbm/). Images were first segmented into gray matter (GM) and white matter. Coregistration of GM images across participants was achieved using the DARTEL (Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra) algorithm. The resulting template image was transformed to MNI stereotactic space using affine and nonlinear spatial normalization. Then the images were smoothed with a Gaussian kernel (full-width at half-maximum, FWHM = 10 mm). In a multiple regression to the GM volume in each voxel in an anatomically defined ACC mask we used the Compensation Index as covariate of interest, and used participants' age, gender and total brain volume as covariates of no interest were entered into a multiple regression to the GM volume in each voxel in the whole brain.

Multi-level kernel density analysis

Meta-analysis of neuroimaging studies is useful for testing the consistency of the recruitment of certain brain regions by certain psychological process and for evaluating the specificity of certain brain activation in indicating certain psychological computation (Wager *et al.*, 2009b). The meta-analytic approach adopted here, and used in a couple of recent papers (Ekin *et al.*, 2007; Wager *et al.*, 2007; Kober *et al.*, 2008; Wager *et al.*, 2008), is Multi-Level Kernel Density Analysis (MKDA) developed by Tor Wager and is freely accessible on his website (http://wagerlab.colorado.edu/). In MKDA, the coordinates in a standard stereotaxic space (i.e. Montreal Neurological Institute space) are treated as a sparse representation of activated locations. The question asked is whether the distribution of reported peaks shows any pattern or randomly distributed across the entire brain. Here we obtained a database of activation peaks from 15 published studies on emotion regulation and mapped the location of emotion regulation sources and sites (see Supplementary Data).

RESULTS

Behavioral results

Table 2 summarizes the level of pain voluntarily taken by the participants in the first scanning session and their self-reported feelings in the postscan manipulation check. Participants took the highest level of pain in the Self_Incorrect condition and took less in the Both_Incorrect condition and still less in the Partner_Incorrect condition, $F_{(2, 46)} = 65.09$, $P < 0.001$. To control for unwanted influences on the fMRI results, e.g. the influence from the participant's evaluation of his/her own performance in the dot-estimation task, we only compared the Self_Incorrect and Both_Incorrect conditions, in both of which the participant estimated incorrectly. To normalize the difference in the level of pain taken, we computed for each participant an index of tendency to compensation (i.e. Compensation Index) by dividing the difference between the level of pain taken in the Self_Incorrect and that in the Both_Incorrect by the sum of the two (mean Compensation Index = 0.12, $SD = 0.08$).

It should be noted that task difficulty and the sense of responsibility are directly related to whether individuals would actually feel guilty (Hoffman, 1982). We included a postscan question that asked the participants to indicate the perceived difficulty of the dot-estimation task. We found that the participants thought the task to be mildly difficult (5.0 ± 2.4, on 1–9 scale). Moreover, the participants felt more responsible, guilty and distressful for the partner's suffering in the Self_Incorrect than in the Both_Incorrect condition (Table 2). These findings confirmed the validity of our paradigm in guilt induction.

Guilt-related neural activation

We first identified the voxels in which activation was higher in the Self_Incorrect condition than in the Both_Incorrect condition in the first session (voxel-level $P < 0.001$, cluster-level $P_{FWE} < 0.05$). This contrast, after being exclusively masked with the same contrast in the second session (see 'Materials and Methods' section), revealed a single activation in the aMCC (peak voxel in MNI coordinates: 0, 34, 16; $k = 270$; Figure 2A). The application of the exclusive mask

ensured that the findings reported here cannot be explained in terms of social comparison (Fliessbach *et al.*, 2007), although the same pattern of brain activations was obtained without masking. To further ensure that the activation of aMCC was only present in the Pain but not in the No Pain (control) session, we extracted and plotted the regional signal change of aMCC in both sessions (Figure 2B). As can be seen, differential activations between the 'Self_Incorrect' and the 'Both_Incorrect' conditions were only present for the Pain session. Noted that we used the term 'aMCC' following the suggestion of Shackman *et al.* (2011) and Vogt (2009); where the cingulate cortex is parcellated on the basis of regional differences in microanatomy, connectivity, and physiology. In some previous publications, especially those related to 'social pain', this area is sometime included in a region called dorsal ACC (dACC; Eisenberger, 2012).

Tab 2 Behavioral results

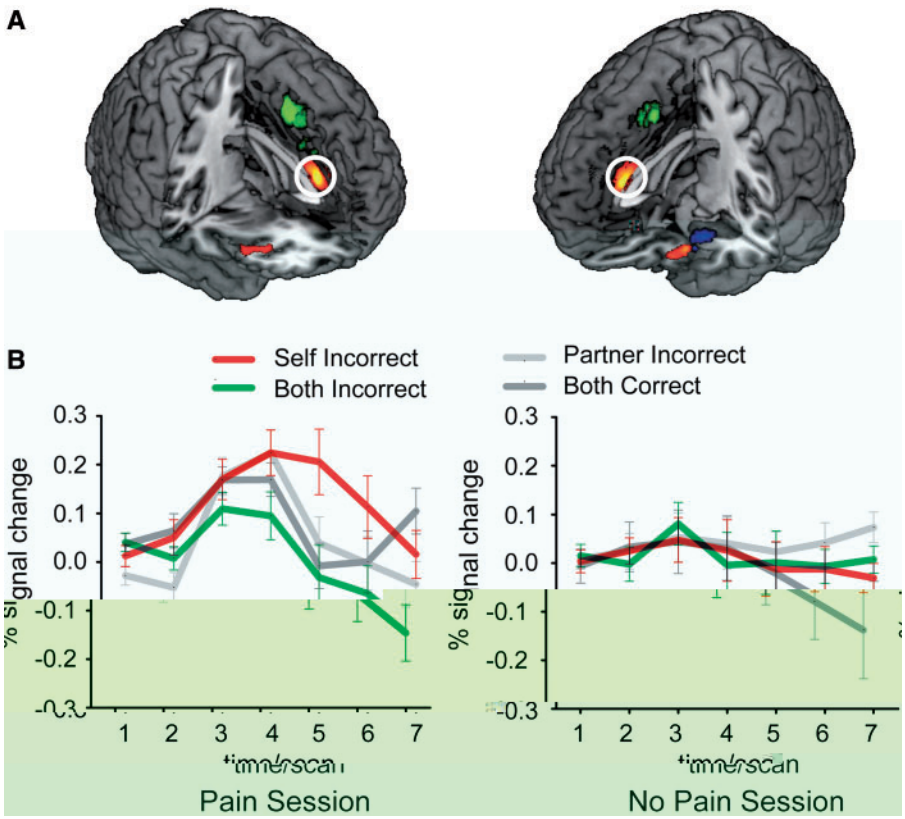
Item	Partner incorrect	Both incorrect	Self incorrect	F (2, 46)
Pain taken	2.0_a (0.6)	2.8_b (0.6)	3.1_c (0.6)	65.09^{***}
Responsibility	3.0_a (1.8)	4.6_b (1.6)	7.1_c (1.6)	35.31^{***}
Guilt	1.8_a (0.9)	3.4_b (1.7)	5.3_c (2.3)	33.43^{***}
Distress	2.0_a (1.5)	2.8_a (2.0)	4.0_b (2.4)	10.51^{***}
Fear	2.6 (1.8)	3.3 (2.3)	2.9 (1.9)	1.47
Anger	2.0 (1.6)	2.6 (2.3)	2.0 (1.3)	1.89

Note. Standard deviations are shown in parentheses. Significant differences (critical $\alpha < 0.05$, Bonferroni corrected) in pair-wise comparison are denoted by different subscripts. $^{***}P < 0.001$.

The bilateral insula failed to reach the whole-brain cluster level threshold. However, since bilateral AI/LOFC is consistently implicated in imagining and recalling guilt-related situations (Chang *et al.*, 2011; Wagner *et al.*, 2011), and since these areas in the present analysis did show activation in the whole-brain analysis at a slightly liberal cluster-level threshold ($P < 0.001$ uncorrected, $k > 50$), we conducted a spatially restricted analysis using anatomically defined ROI masks based on the automatic anatomical labeling (AAL) system (Maldjian *et al.*, 2003). The mask consisted of the insula and the inferior frontal gyrus (*pars orbitalis*). Activations were thresholding at $P_{FWE} < 0.05$ both at the voxel- and the cluster-level. Significant activations were found both in the left (−30, 16, −18; $k = 116$) and in the right insula (36, 30, −8; $k = 90$).

The intra-participant parametric analysis of outcome-related brain activity and pain chosen revealed that activations in right (i.e. contralateral to the stimulated hand) putamen, mid-insula and superior parietal cortex were *negatively* correlated with the level of pain chosen (Supplementary Figure S1). These regions have been implicated in anticipation and experience of pain (Hui *et al.*, 2000; Bingel *et al.*, 2002). It is conceivable that the anticipation of physical pain had prevented the participant from choosing higher level of pain, rendering them to behave more selfishly. The reversed contrast did not reveal any significant activation.

It may be argued that the activation increase observed in the aMCC and bilateral insula arose from the received and/or anticipated pain rather than the feeling of guilt *per se*. This is plausible because the participants generally selected higher pain stimulation in the 'Self_Incorrect' relative to the 'Both_Incorrect' condition. However,



F . 2 Results of the 'Self_Incorrect > Both_Incorrect' contrast and of the meta-analysis. (**A**) Results of the contrast 'Self_Incorrect > Both_Incorrect' in the first session is shown in yellow-to-red ($P < 0.001$ uncorrected, $k > 50$). Results of the meta-analysis of the emotion regulation literature are shown in green (regulation source) and blue (regulation site) ($P_{FWE} < 0.05$). The activations observed in this study are clearly distinct from the source of emotion regulation. (**B**) Timecourse of percent fMRI signal change in the aMCC ROI in the first (Pain, left) and second (No Pain, right) session. The timecourse is locked to the onset of the estimation outcome. Error bars indicate standard error (s.e.m.).

we found that during the pain delivery stage, activations in bilateral dorsal-posterior insula, dorsal middle cingulate cortex and bilateral primary sensory area positively correlated with the level of pain stimulation. Importantly, these brain regions are clearly separated from those observed in the contrast of 'Self_Incorrect > Both_Incorrect' (Supplementary Figure S2), suggesting that activations of aMCC and bilateral insula were not caused by pain stimulation *per se*.

Although our analysis of brain activations focused on the 'Self_Incorrect *vs* Both_Incorrect' contrast, we did check the activations of the 'Partner_Incorrect *vs* Both_Correct' contrast. The 'Partner_Incorrect > Both_Correct' contrast revealed activations in dACC, posterior cingulate cortex and left AI, i.e. the empathy network (Lamm *et al.*, 2011; Bernhardt and Singer, 2012). The reversed contrast revealed activations in the ventral striatum and bilateral amygdala, which are implicated in responding to salient positive outcome (Pessoa and Adolphs, 2000; Delgado *et al.*, 2008). These results confirmed the validity of our experimental manipulation, suggesting that the participants believed in the setup and were emotionally involved in the task.

We have focused our neuroimaging analysis on the outcome-stage brain activity rather than the decision-stage activity for two reasons. First, the participant could know the consequence of his/her guess performance at the outcome stage. The emergence of guilt, as soon as the bad consequence for the partner was clear, should be automatic and immediate. Thus it is conceivable to expect the neural processing of guilt to be initiated at the outcome stage. Second, the neural signals associated with the decision stage might suffer more from confounding factors than those associated with the outcome phase, such as motor responses and pain anticipation. When we looked into the data associated with the decision stage, we did not find any significant activation for the contrast 'Self_Incorrect > Both_Incorrect'. When we used a relatively liberal threshold ($P < 0.005$ uncorrected with 20 contiguous voxels), we found activations within the primary motor area, which probably reflect the motor component in the decision stage.

Functional specifity analysis

It could be argued that the aMCC activation observed in the above analysis arises from the suppression of the unpleasant feeling of guilt, rather than the experience of guilt *per se*. To test this hypothesis, we conducted a MKDA (Wager *et al.*, 2009b) based on 15 published studies on emotion regulation, including cognitive reappraisal, suppression, emotional conflict control, etc (see 'Materials and Methods' section). Activations were dichotomized into source and site, with the source being regions that increase activation during regulation and the site being regions that decrease activation during regulation. Results showed that the source of emotion regulation was predominantly located in the supplementary motor area (SMA; Figure 2A, green) while the site mostly was located in the left amygdala (Figure 2A, blue) ($P_{FWE} < 0.05$). This finding is consistent with a recent meta-analysis of emotion regulation that took into account a larger number of studies (Ochsner *et al.*, 2012). As can be seen, the aMCC activation in this study is clearly distinct from the regulation source, suggesting that aMCC plays a role other than emotion regulation. Taken together, we are confident to conclude that the activations in the aMCC observed here reflect the experience of interpersonal guilt.

Structural correlate of the individual sensitivity to guilt

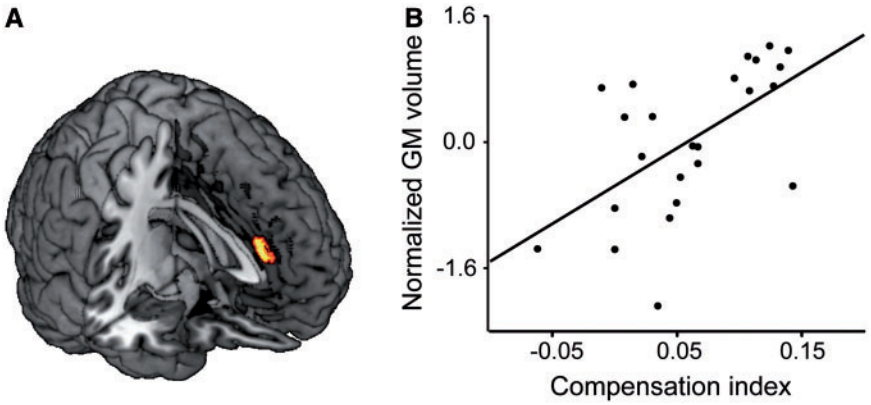
The aMCC activation reported above highlighted the physiological dynamics underlying the processing of interpersonal guilt. It is not clear whether this brain region is also structurally involved. We thus used voxel-based morphometry (VBM) (Ashburner and Friston, 2000) to test for a correlation between the GM volume of aMCC and the Compensation Index. We found that, over the 24 participants, the GM volume in an aMCC cluster (0, 34, 24; $k = 98$; $P_{FWE} < 0.05$ both at the voxel- and the cluster-level) positively correlated with their sensitivity to guilt (Figure 3; see 'Materials and Methods' section). No significant results were obtained outside the mask, in a whole-brain exploratory analysis.

Mediation pathway analysis

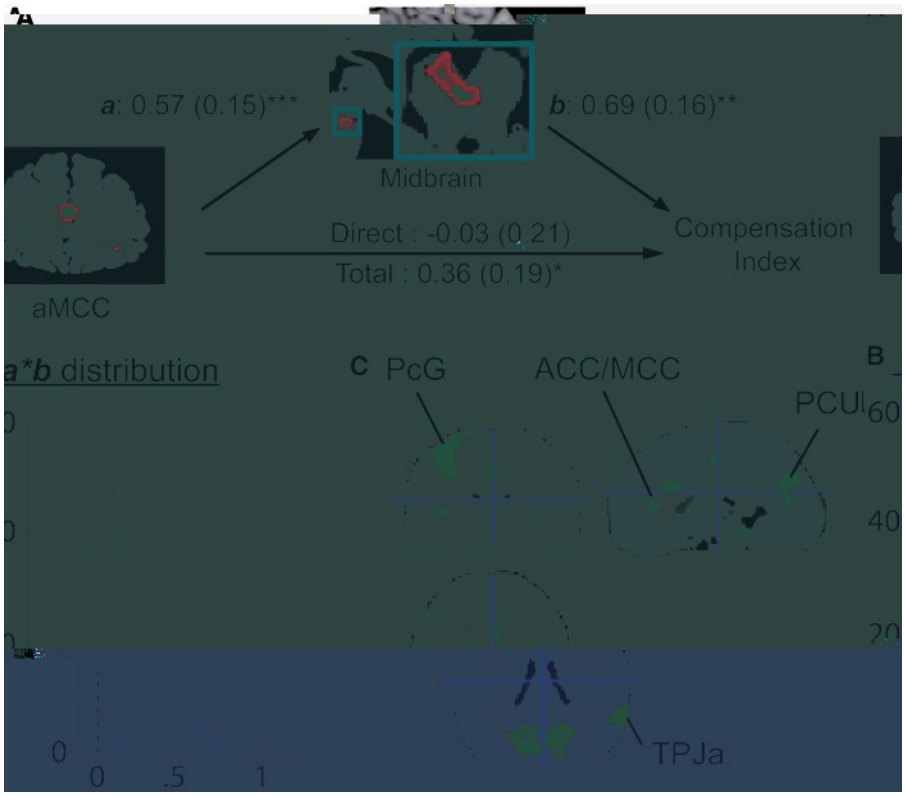
To investigate the neural pathway mediating the brain-to-behavioral translation of guilt-related responses, we need first to identify the brain correlates of the individuals' tendency to compensate. We thus carried out a voxel-wise robust regression analysis (Wager *et al.*, 2005) to correlate Compensation Index with guilt-related brain activation ('Self_Incorrect > Both_Incorrect'). Robust regression is useful for examining questions about individual differences because it down-weights potential outliers that could exert undue leverage on results (Ochsner *et al.*, 2009). A conventional statistical threshold for this analysis, false discovery rate (FDR) corrected $q_{FDR} < 0.05$, was used. The regression was conducted only within the midbrain region defined by anatomical mask (the AAL system, see above), and a significant cluster was found within this region (−2, −20, −20; $k = 7$; Figure 4A). This anatomical region was chosen because it has been implicated in representing social threat and in translating the brain processes of social threat to physiological responses to the threat (Wager *et al.*, 2009a).

It should be noted that we do not claim this area to be the PAG, which was identified in a couple of previous neuroimaging studies on physical and social threat (for a review, see Buhle *et al.*, in press). Due to the low spatial resolution of BOLD signal (relative to the size of the midbrain nucleus), we cannot say anything decisive about what this activation is. A speculation could be that this activation is centered in the caudal ventral tegmental area (VTA) and/or rostral medial tegmental nucleus, both of which play important roles in aversive processes (Laviolette *et al.*, 2002; Laviolette and van der Kooy, 2003; Jhou *et al.*, 2009). The whole-brain explorative analysis of robust regression did not yield significant results after cluster-level FDR correction. Nevertheless, when we used a relatively liberal criterion, i.e. a minimum of 20 contiguous voxels each significant at $P < 0.001$, we found that an empathy-related network, including the cuneus, precuneus, and superior parietal lobule (Supplementary Figure S3), whose activations were positively correlated with the Compensation Index.

If interpersonal guilt can be conceptualized as a type of anxiety and social threat, as argued above, then it is possible that the midbrain mediates the relation between the cortical affective processing of interpersonal guilt and the behavioral responses to the guilt. We thus carried out a recently developed procedure to test the hypothesis that the midbrain mediates the relation between the cortical affective processing of interpersonal guilt and the experiential-behavioral responses to the guilt (MEPM) (Wager *et al.*, 2008, 2009a). First, we tested whether the aMCC exerts indirect influence, via the mediation of midbrain, on the sensitivity to guilt. Parameter estimates corresponding to the estimation outcome were extracted from a 6-mm edge cube around the peak voxel of the aMCC revealed by the contrast 'Self_Incorrect > Both_Incorrect' and around the midbrain revealed by the robust regression. With the aMCC activation as predictor, the midbrain activation as mediator, and the Compensation Index as outcome, the MEPM estimated the strength and significance of the mediation relationship. Confirming our prediction, the mediation effect of midbrain was significant ($P = 0.004$; nonparametric bootstrap test for whether the mediation strength is significantly different from zero) such that the relationship between aMCC activation and Compensation Index was fully mediated by midbrain activation



F . 3 Results of the multiple regression analysis on the cingulate GM volume. The GM volume in an aMCC cluster positively correlated with the individual sensitivity to guilt (Compensation Index).



F . 4 Mediation analysis results. (**A**) Path diagram shows the relationships between regions in the path model. The predictor region (aMCC) is shown on the left, which predicts activations in a midbrain nucleus. This is the *a* path for the mediator region. The mediator region's connection to individual sensitivity to guilt (Compensation Index) is the *b* path. It was calculated controlling for aMCC activity. The lines are labeled with path coefficients, and standard errors are shown in parentheses. The direct path was calculated controlling for the mediator. (**B**) The bootstrapped mediation effect (path *a*b*) for the aMCC. (C) Results of the whole-brain exploratory MEPM analysis for indirect cortical influence of Compensation Index, whose effect is mediated by the midbrain. Activation clusters are shown at $q_{FDR} < 0.05$. Confirming the ROI-based analysis, the ACC/MCC is within the cortical network that influences Compensation Index indirectly via the midbrain. $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$, two-tailed.

(Figure 4A, B and Supplementary Figure S4). We then conducted a whole-brain exploratory search for indirect influence of Compensation Index, whose relationship was mediated by midbrain, with a $q_{FDR} < 0.05$ threshold. Confirming the ROI-based analysis, the aMCC was found significantly activated in the whole-brain analysis (Figure 3C). In addition, an empathy-related network (Lamm *et al.*, 2011; Bernhardt and Singer, 2012), including the right anterior temporoparietal junction (TPJa), the precentral gyrus (PcG) and the precuneus (PCU), was found to contribute to individual's tendency to compensation via the mediation of midbrain (Table 3 and Figure 3C).

DISCUSSION

Past research has not adequately investigated the brain and behavioral responses of social emotions in life-like circumstances; our study on interpersonal guilt in social interactive context is an important extension to this past research. Due to its interpersonal nature (Baumeister *et al.*, 1994), the feeling of and the responses to guilt is most natural in a social interactive context. Utilizing a game paradigm and functional/structural MRI, we showed the recruitment of the aMCC-AI network in response to an interpersonal guilt situation. This finding is in line with a recent neuroimaging study using a similar behavioral paradigm as the current one (Koban *et al.*, in press). Moreover, the structural

variability (GM volume) in the aMCC predicted individual difference in the sensitivity to guilt. Furthermore, we discovered that the cortical processing of guilt was translated by the midbrain to behavioral responses, suggesting that the human neural system processes the interpersonal guilt as social threat and anxiety (Wager *et al.*, 2009a) and attempts to minimize it through compensation behaviors.

The cingulate cortex and the insula are known to participate in a multitude of sensory, affective, cognitive and motivational processes. Their coactivations are seen in experiencing physical and social pain, empathy for pain, disgust, taste, etc (Lamm *et al.*, 2011; Bernhardt and Singer, 2012; Bernhardt *et al.*, in press). Theories have been proposed concerning the functional interplay between AI and ACC/MCC in various cognitive and affective processing. Craig (2002, 2009) argued that insular cortex plays a major role in interoception, i.e. translating bodily states to conscious emotional feeling states; ACC/MCC, in turn, forms the motivational and action-related output. Similarly, Medford and Critchley (2010) suggested that while the AI forms an input region of a system that is based on self-awareness, the global emotional feeling states are ultimately re-represented in cingulate cortex to generate and regulate appropriate responses. Viewed in this context, activations of AI and aMCC here may arise from participants' increased distress and anxiety associated with guilt and the motivation to take actions to reduce this distress (Baumeister *et al.*, 1994; Griswold, 2007). In social interaction, distressing events, such as being isolated by others or betrayed by one's romantic partner, cause substantial feeling of pain (MacDonald *et al.*, 2005) and reliably elicit activations in cingulate cortex and insula (Eisenberger, 2012). This argument is also consistent with our postscan rating that the feeling of guilt and distress are significantly more intense in the 'Self_Incorrect' than in the 'Both_Incorrect' condition (Table 2). Therefore, we interpret the activations in aMCC and AI as reflecting the distressing emotional state arising from causing harm to the partner, i.e. interpersonal guilt.

We did not observe activations of dACC and SMA as one previous study did (Chang *et al.*, 2011). Interestingly, however, the dACC and the SMA clusters observed in that study overlapped with the emotion regulation source obtained through the meta-analysis (Figure 2A), indicating that those brain regions may serve to suppress the selfish impulse of the trustee in a Trust Game context. This is in line with the

experience of certain affective states is translated to behaviors. We highlighted the important role of midbrain nucleus in this function, which fits with the theoretical framework concerning the function of the midbrain nucleus in mediating brain–body interaction. Finally, utilizing structural imaging technique, we showed that the aMCC GM volume contributes to the individual difference related to compensation behavior. These findings may shed light on the understanding of the neural basis of psychopathy.

SUPPLEMENTARY DATA

Supplementary data are available at SCAN online.

Conflict of Interest

None declared.

REFERENCES

Ashburner, J., Friston, K.J. (2000). Voxel-based morphometry – the methods. *Neuroimage*, *11*, 805–21.

Basile, B., Mancini, F., Macaluso, E., Caltagirone, C., Frackowiak, R.S.J., Bozzali, M. (2011). Deontological and altruistic guilt: evidence for distinct neurobiological substrates. *Human Brain Mapping*, *32*, 229–39.

Batson, C.D., Duncan, B.D., Ackerman, P., Buckley, T., Birch, K. (1981). Is empathic emotion a source of altruistic motivation? *Journal of Personality and Social Psychology*, *40*, 290–302.

Battigalli, P., Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, *97*, 170–6.

Baumeister, R., Stillwell, A.M., Heatherton, T.F. (1994). Guilt: an interpersonal approach. *Psychological Bulletin*, *115*, 243–67.

Benedict, R. (1946). *The Chrysanthemum and the Sword: Patterns of Japanese culture*. New York, NY: Houghton Mifflin.

Bernhardt, B.C., Singer, T. (2012). The neural basis of empathy. *Annual Review of Neuroscience*, *35*, 1–23.

Bernhardt, B.C., Klimecki, O.M., Leiberg, S., Singer, T. (in press). Structural covariance networks of the dorsal anterior insula predict females' individual differences in empathic responding. *Cerebral Cortex*. Advance online publication: doi:10.1093/cercor/bht072.

Berthoz, S., Grezes, J., Armony, J.L., Passingham, R.E., Dolan, R.J. (2006). Affective response to one's own moral violations. *Neuroimage*, *31*, 945–50.

Bingel, U., Quante, M., Knab, R., Bromm, B., Weiller, C., Büchel, C. (2002). Subcortical structures involved in pain processing: evidence from single-trial fMRI. *Pain*, *99*, 313–21.

Blair, J. (2006). The emergence of psychopathy: implications for the neuropsychological approach to developmental disorders. *Cognition*, *101*, 414–42.

Buhle, J.T., Kober, H., Ochsner, K.N., et al. (in press). Common representation of pain and negative emotion in the midbrain periaqueductal gray. *Social Cognitive and Affective Neuroscience*, *8*, 609–16.

Chang, L.J., Smith, A., Dufwenberg, M., Sanfey, A.G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron*, *70*, 560–72.

Craig, A.D. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Reviews Neuroscience*, *3*, 655–66.

Craig, A.D. (2009). How do you feel: now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, *10*, 59–70.

de Hooge, I.E., Nelissen, R.M.A., Breugelmans, S.M., Zeelenberg, M. (2011). What is moral about guilt? Acting "prosocially" at the disadvantage of others. *Journal of Personality and Social Psychology*, *100*, 462–73.

Delgado, M.R., Nystrom, L.E., Fissell, C., Noll, D.C., Fiez, J.A. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of Neurophysiology*, *84*, 3072–7.

Devinsky, O., Morrell, M.J., Vogt, B.A. (1995). Contributions of anterior cingulate cortex to behaviour. *Brain*, *118*, 279–306.

Eisenberger, N.I. (2012). The pain of social disconnection: examining the shared neural underpinnings of physical and social pain. *Nature Reviews Neuroscience*, *13*, 421–34.

Eisenberger, N.I., Lieberman, M.D., Williams, K.D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, *302*, 290–2.

Fehr, E., Schmidt, K.M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*, 817–68.

Fehr, E., Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137–40.

Fliessbach, K., Weber, B., Trautner, P., et al. (2007). Social comparison affects reward-related brain activity in the human ventral striatum. *Science*, *318*, 1035–8.

Griswold, C.L. (2007). *Forgiveness: A Philosophical Exploration*. New York, NY: Cambridge University Press.

Haidt, J. (2003). The moral emotions. In: Davidson, R.J., Scherer, K.R., Goldsmith, H.H., editors. *Handbook of Affective Sciences*. New York, NY: Oxford University Press, pp. 852–70.

Hoffman, M.L. (1982). Development of prosocial motivation: empathy and guilt. In: Eisenberg, N., editor. *The Development of Prosocial Behavior*. San Diego, CA: Academic Press, pp. 281–313.

Hui, K.K.S., Liu, J., Makris, N., et al. (2000). Acupuncture modulates the limbic system and subcortical gray structures of the human brain: evidence from fMRI studies in normal subjects. *Human Brain Mapping*, *9*, 13–25.

Inui, K., Tran, T.D., Hoshiyama, M., Kakigi, R. (2002). Preferential stimulation of Aδ fibers by intra-epidermal needle electrode in humans. *Pain*, *96*, 247–52.

Jhou, T.C., Fields, H.L., Baxter, M.G., Saper, C.B., Holland, P.C. (2009). The rostromedial tegmental nucleus (RMTg), a GABAergic afferent to midbrain dopamine neurons, encodes aversive stimuli and inhibits motor responses. *Neuron*, *61*, 786–800.

Kanai, R., Rees, G. (2011). The structural basis of inter-individual differences in human behaviour and cognition. *Nature Reviews Neuroscience*, *12*, 231–42.

Kiehl, K.A. (2006). A cognitive neuroscience perspective on psychopathy: Evidence for paralimbic system dysfunction. *Psychiatry Research*, *142*, 107–28.

Koban, L., Corradi-Dell'Acqua, C., Vuilleumier, P. (in press). Integration of error agency and representation of others' pain in the anterior insula. *Journal of Cognitive Neuroscience*. Advanced online publication: doi:10.1162/jocn_a_00324.

Krajbich, I., Adolphs, R., Tranel, D., Denburg, N.L., Camerer, C.F. (2009). Economic games quantify diminished sense of guilt in patients with damage to the prefrontal cortex. *Journal of Neuroscience*, *29*, 2188–92.

Lamm, C., Decety, J., Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *Neuroimage*, *54*, 2492–502.

Laviolette, S.R., Alexson, T.O., van der Kooy, D. (2002). Lesions of the tegmental pedunculopontine nucleus block the rewarding effects and reveal the aversive effects of nicotine in the ventral tegmental area. *The Journal of Neuroscience*, *22*, 8653–60.

Laviolette, S.R., van der Kooy, D. (2003). Blockade of mesolimbic dopamine transmission dramatically increases sensitivity to the rewarding effects of nicotine in the ventral tegmental area. *Molecular Psychiatry*, *8*, 50–9.

MacDonald, G., Kingbury, R., Shaw, S. (2005). Adding insult to injury: social pain theory and response to soical exclusion. In: Williams, K.D., Forgas, J.P., von Hippel, W., editors. *The Social Outcast: Ostracism, Social Exclusion, Rejection, and Bullying*. New

Wager, T.D., Scott, D.J., Zubieta, J-K. (2007). Placebo effects on human μ-opioidactivity during pain. *Proceedings of the National Academy of Sciences*, *104*, 11056–61.

Wager, T.D., Davidson, M.L., Hughes, B.L., Lindquist, M.A., Ochsner, K.N. (2008). Prefrontal-subcortical pathways mediating successful emotion regulation. *Neuron*, *59*, 1037–50.

Wager, T.D., van Ast, V.A., Hughes, B.L., Davidson, M.L., Lindquist, M.A., Ochsner, K.N. (2009a). Brain mediators of cardiovascular responses to social threat, Part II: prefrontal-subcortical pathways and relationship with anxiety. *Neuroimage*, *47*, 836–51.

Wager, T.D., Lindquist, M.A., Nichols, T.E., Kober, H., van Snellenberg, J.X. (2009b). Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *Neuroimage*, *45*, S210–21.

Wagner, U., N'Diaye, K., Ethofer, T., Vuilleumier, P. (2011). Guilt-specific processing in the prefrontal cortex. *Cerebral Cortex*, *21*, 2461–70.

Zahn, R., Moll, J., Paiva, M., et al. (2009). The neural basis of human social values: evidence from functional MRI. *Cerebral Cortex*, *19*, 276–83.