**Social Cognitive and Affective Neuroscience Advance Access published January 8, 2016**

# 'Imagined guilt' versus 'recollected guilt': Implications for fMRI

Neil Mclatchie[1], Roger Giner-Sorolla[2] and Stuart W.G. Derbyshire[3]

1. School of Psychology, Lancaster University, UK; 2. School of Psychology, University of Kent, UK; 3. Department of Psychology and A*STAR-NUS Clinical Imaging Research Centre, National University of Singapore, Singapore.

Correspondence:

Stuart W.G. Derbyshire

Department of Psychology

Block AS4, Level 2

National University of Singapore

Singapore 117570

Tel: (65) 6516 4115

Email: psydswg@nus.edu.sg

**Abstract**

Guilt is thought to maintain social harmony by motivating reparation (Haidt, 2003; Trivers, 1971). The present study compared two methodologies commonly used to identify the neural correlates of guilt. The first, imagined guilt, requires participants to read hypothetical scenarios and then imagine themselves as the protagonist. The second, recollected guilt, requires participants to reflect on times they personally experienced guilt. In the fMRI scanner, participants were presented with guilt/neutral memories and guilt/neutral hypothetical scenarios. Contrasts confirmed a priori predictions that guilt memories, relative to guilt scenarios, were associated with significantly greater activity in regions associated with affect (ACC, Caudate, Insula, OFC) and social cognition (TP, precuneus). Similarly, results indicated that guilt memories, relative to neutral memories, were also associated with greater activity in affective (ACC, amygdala, Insula, OFC) and social cognition (mPFC, TP, precuneus, TPJ) regions. There were no significant differences between guilt hypothetical scenarios and neutral hypothetical scenarios in either affective or social cognition regions. The importance of distinguishing between different guilt inductions inside the scanner are discussed. We offer explanations of our results and discuss ideas for future research.

Keywords: Guilt, Memories, Hypothetical Scenarios

## Introduction

A considerable body of research has demonstrated that guilt is elicited following a transgression against another individual or group and will influence subsequent moral decisions and moral behavior (Haidt, 2003; Trivers, 1971). Guilt can, for example, motivate individuals to act in a reparative (Ketelaar & Au, 2003; Nelissen, 2011; Tangney & Dearing, 2002) or generally prosocial manner (Regan, Williams, & Sparling, 1972).

The motivational component of guilt is generally considered to be part of the complex emotional experience that constitutes guilt (Baumeister, Vohs, DeWall, & Zhang, 2007; Izard, 2007). Guilt can be understood as an *emotion schema* (Izard, 2007), involving interactions of self-directed negative affect with self/other distinction, agency, counterfactual thinking, regret and future planning. It is this interaction of emotion with cognition that is believed to deliver the powerful motivation to act.

Findings from neuroimaging support the understanding of guilt as involving a complex interaction of affect and cognition (see Kédia, Berthoz, Wessa, Hilton, & Martinot, 2008 for review). Multiple studies have shown that feelings of guilt activate affect-related regions including the anterior cingulate cortex (ACC; Kédia et al., 2008; Shin et al., 2000), the orbital frontal cortex (OFC; Moll & de Oliveira-Souza, 2007; Morey et al., 2012; Zahn et al., 2009), the insula (Michl et al., 2014; Shin et al., 2000; Wagner, N'Diaye, Ethofer, & Vuilleumier, 2011), the amygdala (Berthoz, Grezes, Armony, Passingham, & Dolan, 2006; Kédia et al., 2008) and the basal ganglia (Kédia et al. 2008)

One or more of these regions have been activated during experiments involving: the perception of emotional stimuli, including facial expressions (e.g., amygdala, Hare, Tottenham, Davidson, Glover, & Casey, 2005; Hariri, Tessitore, Mattay, Fera, & Weinberger, 2002) and speech (e.g., basal ganglia, Paulmann, Pell, & Kotz, 2008; Pell & Leonard, 2003); changes in, and awareness of, physiological arousal (e.g., insula, Critchley, Wiens, Rotshtein,

Öhman, & Dolan, 2004; amygdala, Gläscher & Adolphs, 2003) ; motivation, including updating motivational states (e.g., ACC, Wager & Feldman-Barrett, 2004) and connecting motivational goals with visual information (e.g., basal ganglia, Kawagoe, Takikawa, & Hikosaka, 1998); reinforcing behaviours (e.g., the OFC, Bechara, Tranel, & Damasio, 2000; O'Doherty, Kringelbach, Rolls, Hornak, & Andrews, 2001; ACC, Bush et al., 2002; Etkin, Egner, & Kalisch, 2011) and memory encoding (e.g., amygdala, Buchanan, 2007; Canli, Zhao, Brewer, Gabrieli, & Cahill, 2000) and subsequent retrieval of emotional events (e.g., OFC, amygdala, (Maratos, Dolan, Morris, Henson, & Rugg, 2001)

Social cognition networks include the medial prefrontal cortex (mPFC; Basile et al., 2011; Finger, Marsh, Kamel, Mitchell, & Blair, 2006; Kédia et al., 2008; Morey et al., 2012; Takahashi et al., 2004), dorsolateral prefrontal cortex (DLPFC; Stone, Baron-Cohen, & Knight, 1998), temporo-parietal junction (TPJ; Finger et al., 2006; Kédia et al., 2008), the temporal poles (TP; Finger et al., 2006; Shin et al., 2000; Wagner, N'Diaye, Ethofer, & Vuilleumier, 2011), and the precuneus (Kédia et al., 2008; Moll & de Oliveira-Souza, 2007; Takahashi et al., 2004). Social cognition networks are broadly associated with processing social information, including: perception (e.g., mPFC, Harris & Fiske, 2007; TPJ, Pelphrey & Carter, 2008), attention (e.g., TPJ, Nummenmaa & Calder, 2009), and storage and retrieval (e.g., precuneus, Cavanna & Trimble, 2006; TP, Olson, McCoy, Klobusicky, & Ross, 2012). Social cognition networks, such as the TP (Olson, Plotzker, & Ezzyat, 2007), the mPFC (Dodell-Feder, Koster-Hale, Bedny, & Saxe, 2011; Gallagher et al., 2000) and the TPJ (Saxe, 2010), are activated during experiments where participants imagine what others might be thinking or feeling or where participants generally take the perspective of another. Specific regions in social cognition networks activated during episodes of guilt include the mPFC (Basile et al., 2011; Finger et al., 2006; Kédia et al., 2008; Morey et al., 2012; Takahashi et al., 2004), the temporal poles (Finger et al., 2006; Shin et al., 2000; Wagner et al., 2011), the

precuneus (Kédia et al., 2008; Moll & de Oliveira-Souza, 2007; Takahashi et al., 2004) and the TPJ (Finger et al., 2006; Kédia et al., 2008).

Thus, the neuroimaging literature is often interpreted as demonstrating guilt to involve negative affect combined with other-directed cognitions, or other variation of the *emotion schema* described by Izard (2007). This interpretation, however, cannot be fully justified because the limitations of neuroimaging have so far prevented direct assessment of guilt as an emotion schema.

A major limitation of neuroimaging research is the necessity for participants to remain stationary. Even small movements or rotations of the head render the images unusable. Consequently, the elaborate set-ups and manipulation tasks commonly employed by social psychologists to induce feelings of guilt (Nelissen & Zeelenberg, 2009; Regan et al., 1972) are not transferable for use with neuroimaging. To accommodate the limitations of neuroimaging, two methods of inducing guilt within the scanner have primarily been used: the *memory recollection task* and the *hypothetical scenario task*. Table S1 summarizes the different results when using these two methods to examine the neural correlates of guilt. Studies involving the memory recollection task typically ask participants to recall a time that they transgressed against another individual. The memory recollection task is commonly used in behavioral experiments to induce feelings of guilt (Bastian, Jetten, & Fasoli, 2011; Ketelaar & Au, 2003; Zhong & Liljenquist, 2006) but has been used much less frequently inside the scanner (for examples, see Shin et al., 2000; Wagner et al., 2011). Participants recalled personal events that induced feelings of guilt and these events were presented to the participants during scanning to capture the neural correlates of guilty feelings. These two studies demonstrated common activation in the insula and TP cortices.

Studies involving the hypothetical scenario task typically ask participants to imagine they are the main protagonist in a hypothetical scenario describing a guilt-inducing event.

Unlike recollection, hypothetical scenarios are not directly about the participant's personal behaviour, and there is less evidence that hypothetical scenarios induce feelings of guilt. A central component of guilt is the awareness of one's own responsibility for having committed a transgression against another person or group. When asked to consider hypothetical scenarios, the participant is not truly responsible for any transgression or harm. Consequently, the response to a hypothetical scenario is likely to involve anticipatory thoughts about guilt or the concept of guilt ('guilt thoughts') rather than a feeling of guilt ('guilt feelings'), which might be expected to generate considerably different neural activation. Only one study of guilt using hypothetical scenarios demonstrated activation of the insular cortex (Michl et al., 2014) and only one activated the TP cortex (Finger et al., 2006). Michl et al (2014) used hypothetical scenarios to elicit feelings of guilt and shame but noted as a limitation of their study that they could not guarantee the "success of imagination and generation of moral feelings" (p. 155).

Importantly, guilt thoughts lack the painful, self-directed negative affect that is central to guilt feelings. This difference may also impact their motivational consequences. While guilt feelings predict reparative or prosocial behaviours (Ketelaar & Au, 2003; Nelissen, 2012), guilt feelings can also motivate dysfunctional behaviour including self-punishment (Bastian et al., 2011) and anti-social behaviours (de Hooge, Nelissen, Breugelmans, & Zeelenberg, 2011). The self-directed negative affect of guilt feelings may motivate negative-state relief that conflicts with the motivation for prosocial behaviours (Miller, 2010 for a review).

In contrast, guilt thoughts have been shown consistently to motivate prosocial behaviors. Anticipating guilt has been associated with increased charity donations (Lindsey, 2005) and a decreased likelihood of cheating in exams (Malinowski & Smith, 1985). Other studies have shown that subtly making the concept of guilt accessible can promote reparative

and prosocial behaviours (Giner-Sorolla, 2001; Zemack-Rugar, Bettman, & Fitzsimons, 2007). The cognitive reflection associated with guilt thoughts may motivate actions to prevent guilt in the future and thus motivate prosocial behaviours (Baumeister et al., 2007). It seems likely, therefore, that neuroimaging studies of guilt feelings versus guilt thoughts will produce different activations as suggested by the summary in table S1. To date, no study has directly tested for differences between guilt feelings as induced by memory recollection and guilt thoughts as induced by hypothetical scenarios. Instead, studies have focused on different components of guilt, for example: deontological versus altruistic (Basile et al., 2011), presence versus absence of an audience (Finger et al., 2006), target of agency (Kédia et al. 2008), and whether the outcomes are self- or other-oriented (Morey et al. 2012).

A direct comparison of memory recollection and hypothetical scenarios will be obviously more definitive than the comparison in table S1 because a direct comparison will eliminate confounds such as differences in sample sizes and thresholding across studies. The average sample size of the eight studies in Table 1 equals 15.1, but only three of the studies in Table 1 had a sample size greater than this. There are also considerable differences in methods of analysis. Seven studies employed whole brain analysis with thresholds ranging from $p<0.05$ (Finger et al., 2006; Michl et al., 2004; Morey et al., 2012; Zahn et al., 2009) to $p<0.005$ (Takahashi et al., 2004) to $p<0.001$ (Wagner et al., 2011). Two studies used small volume corrections with a corrected p-value threshold < 0.001 (Shin et al., 2000; Kédia et al., 2008) and one study combined whole brain analysis ($p<0.001$) and small volume analysis ($p<0.05$) (Berthoz et al., 2006).

Given these differences, it is probably not surprising that table S1 does not indicate any single structure as significantly active across all studies. Nevertheless, it is notable that the two studies that used memory recollection reported significant activity in both affect-related and social cognition structures. This is precisely what would be predicted if guilt

feelings induced by memory recollection involve structures associated with affect and social cognition. Of the studies that employed the hypothetical scenario, two studies support the hypothesis that guilt thoughts should not result in increased activity of affect-related structures.

An additional issue is that studies comparing guilt with control scenarios have not taken care to ensure that control scenarios are equal in social content to guilt scenarios. This is an important confound. The times when the literature does show the activation of, in particular, social cognition structures, could thus merely be a function of the incidental social background involved in imagining someone else, not a key component of guilt as opposed to neutral but equally social situations. We therefore thought it was important to eliminate this confound by using neutral social situations as our control group to compare with guilt.

Our study assessed the distinction between guilt thoughts and guilt feelings within the same design by drawing on recalled or anticipated guilt-evoking situations, using fMRI. In additional to the improvement of including a social thought control condition, we took care to ensure that anticipated guilt thoughts were not incidentally drawing on actual memories of guilt feelings, by having the anticipated situations be intentionally chosen as dissimilar to existing experiences.

## Method

### Participants

Twenty-five right-handed students from the University of Birmingham (mean age=25.7, 4 males) took part in the study in exchange for £28 compensation. All participants provided written consent and were fully debriefed at the conclusion of the experiment. No participant had a history of neurological, psychiatric or other chronic clinical disorder.

### Pre-scanning Session

Participants were asked to provide a written description of ten specific memories: five instances of a time that they had caused harm or distress to another person, and five instances of an emotionally-neutral event. Participants typed their memories directly in to a Word document. They were asked to use 50-70 words for each description.

The memories were then serially presented to participants on a computer screen with their ten descriptions in a pseudo-random order (E-Prime). Participants rated the memories on a scale of how guilty each made them feel (0 not at all, 6 extremely guilty). Participants also rated the extent they felt that the behavior in each memory had violated a moral or social code (1 not at all, 5 completely broke a social or moral code).

Participants were then presented with 28 hypothetical scenarios (see Supplemental Material 2; S2). The scenarios were carefully matched so that all scenarios described a social event involving the protagonist and at least one other. This ensured that differences in neural activity were not the product of differences in the social content of guilt and neutral hypothetical scenarios (Finger et al., 2006). Sixteen of the scenarios described a situation in which the self violates a moral or social value (A), and 12 described an emotionally neutral event (B). For example:

A. Getting on to a packed train, you decide to sit in the priority seats, even though they are supposed to be given to more needy people than you, and there are elderly people standing up. After a few stops, you hear a bump. An elderly lady has fallen over. You realise you should have given your seat to her.

B. In order to get to university, you walk to your nearest bus stop. On the way, you bump in to a class mate who is also going to the bus stop. You have a conversation about the planned lessons and she tells you that she is going to town in the evening. After the bus journey, you both go to your lesson.

These hypothetical scenarios were presented to participants one at a time in a pseudo-random order, and participants were asked to rate the extent that the hypothetical protagonists described in each scenario had broken a moral or social code using the same scale they had used for their own memories. Participants also rated the extent that they could identify with the hypothetical scenario's main protagonist on a scale (1 - not at all, 5 - completely), and were told that being unable to identify with the protagonist could be the result of themselves not having committed the same act or one similar to it. Once completed, participants were debriefed, compensated and informed that they would be contacted regarding their participation in the fMRI stage of the experiment.

Each participant memory was matched with a hypothetical scenario according to their ratings of moral or social code violation. Exact matchings and ±1 matchings were considered acceptable. When there was more than one hypothetical scenario that matched with the memory, the hypothetical scenario with which the participant could least identify with was selected. The decision to measure the extent that participants could identify with the hypothetical memory is novel to the current research. While past research could have

benefitted from including a measure of identification, it was essential for the current study so as to control and minimize overlap between memories and scenarios.

Of the 25 participants who attended the first session, 20 were invited back for the scanning session based on successful memory-hypothetical pairings. That is, their ratings of the extent that their memories violated a social or moral code matched their ratings for the extent that a hypothetical scenario described a violation of a social or moral code. This resulted in five guilt memories and five hypothetical scenarios that participants believed involved an act that violated a moral or social code to the same extent, and five neutral memories and five neutral hypothetical scenarios, which participants did not feel described an event in which a moral or social code had been violated.

**fMRI Paradigm**

In the second session, participants were placed in an fMRI scanner. The study incorporated an epoch-based design with participants viewing and reflecting upon their memories for extended periods of time (>10s). While in the scanner, memories (5 guilty, 5 neutral) and hypothetical scenarios (5 guilty, 5 neutral) were presented on a screen positioned directly in front of the participant and viewed in a mirror placed above the participant. The experiment was comprised of three runs. In a single run, participants would view all 20 presentations (10 memories, 10 hypotheticals) in a pseudo-randomised order. Each presentation consisted of three stages: 'reading' (14s) during which they were asked to read what was presented to them, 'reflecting' (10s) during which they were asked to reflect on the presentation that they had just read, and then 'control' (10s), during which they were presented with a crosshair and instructed to empty their mind (see Figure 7). At the completion of the first two runs, participants were given a chance to get comfortable and relax before the next run started. Over the course of three runs, each memory was presented three times, for a total of 60

presentations. The experiment terminated following the third run. Participants were debriefed and received £25 compensation.

Figure 1 near here

**Data Acquisition**

Functional data was acquired using a Philips 3 T Achieva system to acquire BOLD contrast weighted echoplanar images (EPI) for the functional scans (repetition time TR=3000ms, echo time TE=2000ms, 48 sequentially acquired axial slices, 3mm thick with a 3x3mm in-plane resolution, FoV = 220mm). High-resolution structural images were acquired using the T1TFE technique.

*Pre-processing*

Pre-processing and analysis of the data was conducted using SPM8 (Wellcome Department of Imaging Neuroscience, London). The pre-processing followed the same methodology outlined elsewhere (Derbyshire & Osborn, 2009). The first four fMRI volumes were discarded to allow for T1 equilibrium effects. Functional images were first corrected for differences in slice timing by resampling all slices with respect to the middle slice. Movement between scans was corrected for by spatially realigning each scan with the first, and these were then reoriented in to the standardized anatomical space provided by the MNI template. To complete the pre-processing, each image was smoothed in the X, Y and Z dimensions using a Gaussian filter of 8mm (FWHM).

*fMRI Statistical Analysis*

Standard neuroimaging methods based upon the general linear model were used for single participant analysis, which provided contrasts for group analysis at the second level. A box-car model convolved with a hemodynamic delay function was fitted to each voxel generating a statistical image corresponding to condition. Specifically, employing an epoch-based design, individual images were generated by subtracting BOLD activation of: (i) neutral memories from brain activation during reflection of guilt memories, (ii) neutral hypothetical scenarios from brain activation during reflection of guilt hypothetical scenarios, and (iii) guilt hypothetical scenarios from brain activation of guilt memories. Each of the subtractions was also reversed, subtracting BOLD activity during (iv) guilt memories from neutral memories, (v) guilt hypothetical scenarios from neutral hypothetical scenarios, (vi) and guilt memories from guilt hypothetical scenarios. These individual contrasts were then entered into a second level model to provide a group level significance map.

Small-volume correction (SVC) was conducted across all contrasts for 10 predefined affective and social cognition regions. Specifically, a one-sample t-test was performed to assess group level bold activation using the contrast images generated at the individual level. Small volumes were predefined using the MRIcro atlas (www.mricro.com).

The Talairach and Tournoux (1988) atlas and the Talairach Daemon software (http://www.talairach.org/applet.html) were used to infer from the coordinates the region of activity. The coordinates were adjusted to allow for differences between the MNI and Talairach templates as outlined elsewhere (http://imaging.mrc-cbu.cam.ac.uk/imaging/MniTalairach).

*Thresholding*

Whole brain images were thresholded at $p_{uncorr}<0.001$, with an extent threshold of 23 contiguous voxels consistent with previous studies (Berthoz, Grèzes, Armony, Passingham,

& Dolan, 2006; Kédia et al., 2008; Shin et al., 2000; Wagner et al., 2011). In addition, a series of mask images were created for each of the five affective regions (OFC, amygdala, insula, basal ganglia, ACC) and each of the four social cognition structures (mPFC, precuneus, TP, TPJ) identified apriori. The multiple comparisons problem was addressed through Family-Wise Error corrections for each mask separately (a small volume correction). fMRI activations were considered statistically significant if they exceeded a corrected threshold of $p_{fwe}<0.05$. The coordinates of significant peak voxels and the size of the cluster were reported for each mask.

## Results

### Matching Procedure

There was no significant difference between the extent that guilt memories ($M$=3.47, $SD$=0.46) and guilt hypothetical scenarios ($M$=3.54, $SD$=0.42) were considered to have violated a moral or social code, $t(19)=1.02$, $p=0.32$. Similarly, there was no significant difference between the extent that neutral memories ($M$=0.80, $SD$=0.24) and neutral hypothetical scenarios ($M$=0.09, $SD$=0.26) were considered to have violated a moral or social code, $t(19)=0.57$, $p=0.58$.

When more than one hypothetical scenario was considered to have violated a moral or social code to the same degree as a memory, the hypothetical scenario that participants could least identify with was chosen. Results showed that participants could identify with the neutral hypothetical scenarios ($M$=2.59, $SD$=1.79) significantly more than they could identify with the guilt hypothetical scenarios ($M$=1.01, $SD$=0.67), $t(19)=3.76$, $p=0.001$.

### Emotional Ratings

Guilt feelings were significantly higher when participants reflected upon the guilt memories ($M$=4.31) than when reflecting on the neutral memories ($M$=0.13), $t(18)=27.07$, $p<0.001$. A mean rating of 4.31 corresponded to participants feeling "quite guilty" while reflecting on their memories.

### fMRI Data

The paradigm allowed for analysis of brain activity during the reading of each memory and hypothetical scenario, during reflection upon each memory and hypothetical scenario, and whilst observing the crosshair and being asked to empty their minds. Patterns of findings were similar for the reading and reflection periods of memories and hypothetical scenarios.

The current methodology most closely resembles that of Wagner et al. (2011). In line with their research, here we present the analyses of the 10s reflection period during which time participants reflected on the memory or hypothetical scenario they had just read.

**Guilt Memories vs. Neutral Memories**

When contrasted with neutral memories, guilt memories were associated with increased activity in both affective (OFC, ACC, Insula, Amygdala) and social cognition (mPFC, temporal poles, precuneus, TPJ) structures (see Figure 2, Table 1). Whole brain analysis revealed greater activation in the posterior cingulate and the inferior frontal cortices. There was no significant activity in the neutral memory condition when contrasted with the guilt memory condition.

**Guilt Hypothetical vs. Neutral Hypothetical**

When contrasted with the neutral hypothetical scenarios, guilt hypothetical scenarios were not associated with significant increased activity in any structure of the brain. This was the case following whole brain and regional analysis. Furthermore, there was also no significant activity for the reverse contrasts; neutral hypothetical scenarios were not associated with increased activity when contrasted with hypothetical guilt scenarios.

**Guilt Memories vs. Guilt Hypothetical Scenarios**

When neural activity during presentation of guilt memories was contrasted with neural activity during presentation of guilt hypotheticals, there was significantly more activity in both affective (OFC, ACC, caudate) and social cognition (mPFC, precuneus, superior temporal cortex) structures (Figure 3, Table 2). Additionally, other regions that were found to be significantly more active during the guilt memories compared to the guilt-laden

hypothetical scenarios were the thalamus, the posterior cingulate, the primary motor cortex, and the inferior parietal cortex. No regions were significantly more active during the guilt-laden hypothetical scenarios when contrasted with the guilt memories.

Figures 2 and 3 near here

Tables 1 and 2 near here

## Discussion

All participants provided personal accounts of recalled scenarios that involved neutral or guilt-related events (guilt memories). The experimenter generated hypothetical scenarios that also involved neutral or guilt-related events (guilt scenario). The guilt memories and guilt scenarios were successfully matched for the perceived extent to which they violated a moral or social code.

fMRI data revealed significant activation of affective (OFC, ACC, insula and amygdala) and social cognition (mPFC, temporal poles, precuneus and TPJ) structures after viewing guilt memories compared to neutral memories. There were no significant differences when comparing guilt scenarios and neutral scenarios. Direct comparison of activation following presentation of guilt memories with guilt scenarios confirmed greater activation of affective and social cognition structures after recalling guilt memories. These findings confirm the prediction that memories of personal events involving a moral violation will generate activity of both affect and social cognition structures.

In contrast to guilt memories, guilt scenarios did not activate either affect or social cognition structures significantly. Previous studies using hypothetical guilt scenarios have reported activation of social cognition structures (Basile et al., 2011; Berthoz, Grezes, et al., 2006; Finger et al., 2006; Kédia et al., 2008; Morey et al., 2012; Takahashi et al., 2004; Zahn et al., 2009). The absence of any significant activation in our study could be explained by the care we took to match both the guilt and neutral hypothetical scenarios so that each involved a social interaction. For example, one of our hypothetical guilt scenarios involved kicking a ball away from a group of children playing football while one of the neutral scenarios involved cutting a hedge for grandparents. Both scenarios involve social interaction and motor behavior and so might, therefore, activate similar brain regions. In contrast, past studies have offered hypothetical scenarios with varying levels of social content. For

example, Takahashi et al. (2004) presented participants with both solitary ("I change in to pajamas at night") and social ("I betrayed my friend") hypothetical scenarios. In the current study, even though the guilt scenario involves a moral violation, the participant has no personal involvement or responsibility and therefore additional areas associated with guilty feelings are not activated.

The lack of activity in response to guilt scenarios does not support the suggestion that merely considering guilt-related events generates residual guilty feelings (Baumeister, DeWall, Vohs, & Alquist, 2010). In the current study, however, subjective feelings of guilt during scenarios were not recorded concurrently with imaging. It is possible that feelings of guilt were impacted by the guilt-scenarios but this impact was insufficient to generate significant neural activity. Although future studies may address this possibility by recording guilt feelings in response to all scenarios and directly correlating guilt feelings with neural activity, they would do so only under a sort of Heisenberg uncertainty paradox, as engaging in such a repetitive focus on guilt in self-report may increase its anticipation and accessibility beyond what is usual. Collecting guilt measures after the scan might have provided some information but the large number of scenarios led us to not attempt any post-hoc measures. Future studies might consider experimenting with post-hoc measures to at least provide some insight regarding subjective experience in the scanner.

In contrast to the lack of activation after viewing guilt scenarios, viewing guilt memories resulted in activation of both affect and social cognition structures. This finding suggests that guilt memories successfully resulted in generating both affective responses and recall of the events involving other people and is consistent with the understanding of guilt as an emotion schema. Merely thinking about guilt when presented with a hypothetical guilt scenario did not generate activity in affect related regions, suggesting that guilt thoughts do not involve an automatic affective component (Winkielman, Berridge, & Wilbarger, 2005).

Moreover, in this study, reflecting on guilt scenarios also resulted in no additional social cognition activation, which suggests that the guilt thoughts were not markedly different than thoughts about other social situations.

Previous studies using guilt scenarios have demonstrated activation in affective or social cognition regions or both (see table S1). It is possible that these past studies inadvertently involved affect or cognitive triggers during the guilt scenarios that were effectively controlled in the current study. For example, in the current study, participants were presented with scenarios that they had previously rated very low in personal identification. This was to ensure that the distinction between memory and scenario was maintained. Ensuring that participants could not highly identify with the protagonist might also have avoided activating the affective and social cognition structures that were reported active in past studies. Indeed, it is possible that the low identification with the scenario led the participants to not engage strongly with the scenario and thus not generate any feelings of guilt at all. Low identity, however, does not mean low engagement and it is likely that the participants did engage with the scenarios. Participants were asked to imagine they were the protagonist and there is no reason to expect low identity to have prevented that imagined activity taking place; our participants did not report any difficulty in imagining what was happening and clearly were able to rate the scenarios for appropriateness, which indicates active engagement

Furthermore, it is unlikely that past research generated more identity with the protagonists for other hypothetical scenarios that have been used (e.g., attending dinner at a Japanese friend's house, not liking the food and spitting it out on a plate, Berthoz et al., 2006; forgetting to validate a friend's lottery ticket who had winning number, Kedia et al., 2008; or simply viewing an upset face; Basile et al., 2011). Moreover, given the similarity between the scenarios used in the current study and previous research, it is unlikely that low engagement

with the scenarios in the current study fully explains the differences in activation. Nevertheless, future studies might include a mix of high and low identification scenarios to address this concern. Here we chose to restrict our study to personal memories and low identification scenarios to ensure that the activation during guilt hypotheticals was not due to overlap with personal memories and to provide maximal scan data to assess our central hypothesis.

A further limitation is that it is not possible to know from the current study whether the guilt memories generated non-guilt related emotions such as frustration or despair that drove the affective activation. Similarly it is possible that the neutral scenarios also generated guilt related emotions that negated any activation from the guilt scenarios and/or involved a complexity that was not adequately controlled by the guilt scenarios. Future studies might address these points by including additional controls and recording additional subjective measures.

Future research could also provide a rigorous quantitative meta-analysis of previous research. Several methodologies exist that researchers could employ to provide an overview of neural activity during episodes of guilt. These methods, such as Activation Likelihood Estimation (ALE), work by pooling together the 3D coordinates of peak voxels reported as active across multiple studies and compare the observed number of peaks to a null hypothesis distribution (for a review of neuroimaging meta-analysis techniques, see Wager, Lindquist & Kaplan, 2007). Such a meta-analysis could provide better clarity than provided here in table S1 and address the issues pertaining to guilt raised by the current study.

Past imaging research has not explicitly distinguished guilt thoughts and guilt feeling. The results of the current study suggest researchers should be wary of drawing conclusions about emotions from studies based on hypothetical situations, as opposed to lived experience;

and should also study social emotions using neutral scenarios with equally social content, in order to separate incidental confounds from true neural correlates of emotion.

## References


Basile, B., Mancini, F., Macaluso, E., Caltagirone, C., Frackowiak, R. S. ., & Bozzali, M. (2011). Deontological and altruistic guilt: Evidence for distinct neurobiological substrates. *Human Brain Mapping*, *32*(2), 229–239.

Bastian, B., Jetten, J., & Fasoli, F. (2011). Cleansing the soul by hurting the flesh: The guilt-reducing effect of pain. *Psychological Science*, *22*(3), 334 – 335. http://doi.org/10.1177/0956797610397058

Baumeister, R. F., DeWall, C. N., Vohs, K. D., & Alquist, J. L. (2010). Does emotion cause behavior (apart from making people do stupid, destructive things). *Then a Miracle Occurs: Focusing on Behavior in Social Psychological Theory and Research*, 12–27.

Baumeister, R. F., Vohs, K. D., DeWall, C. N., & Zhang, L. (2007). How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review*, *11*(2), 167–203. http://doi.org/10.1177/1088868307301033

Bechara, A., Tranel, D., & Damasio, H. (2000). Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain*, *123*, 2189–2202.

Berthoz, S., Grezes, J., Armony, J. L., Passingham, R. E., & Dolan, R. J. (2006). Affective response to one's own moral violations. *Neuroimage*, *31*(2), 945–950.

Berthoz, S., Grèzes, J., Armony, J. L., Passingham, R. E., & Dolan, R. J. (2006). Affective response to one's own moral violations. *Neuroimage*, *31*(2), 945–950.

Buchanan, T. W. (2007). Retrieval of emotional memories. *Psychological Bulletin*, *133*(5), 761.

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*(1), 523–528.

Canli, T., Zhao, Z., Brewer, J., Gabrieli, J. D., & Cahill, L. (2000). Event-related activation in the human amygdala associates with later memory for individual emotional response. *The Journal of Neuroscience*, *20*, 1–5.

Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, *129*(3), 564–583.

Critchley, H. D., Wiens, S., Rotshtein, P., Öhman, A., & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, *7*(2), 189–195.

De Hooge, I. E., Nelissen, R., 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*(3), 462–473. http://doi.org/10.1037/a0021459

Derbyshire, S. W. G., & Osborn, J. (2009). Offset analgesia is mediated by activation in the region of the periaqueductal grey and rostral ventromedial medulla. *Neuroimage*, *47*(3), 1002–1006.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. *Neuroimage*, *55*(2), 705–712.

Etkin, A., Egner, T., & Kalisch, R. (2011). Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends in Cognitive Sciences*, *15*(2), 85–93.

Finger, E. C., Marsh, A. A., Kamel, N., Mitchell, D. G. ., & Blair, J. R. (2006). Caught in the act: the impact of audience on the neural response to morally and socially inappropriate behavior. *Neuroimage*, *33*(1), 414–421.

Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: an fMRI study of "theory of mind"in verbal and nonverbal tasks. *Neuropsychologia*, *38*(1), 11–21.

Giner-Sorolla, R. (2001). Guilty pleasures and grim necessities: affective attitudes in dilemmas of self-control. *Journal of Personality and Social Psychology*, *80*(2), 206 – 221.

Gläscher, J., & Adolphs, R. (2003). Processing of the arousal of subliminal and supraliminal emotional stimuli by the human amygdala. *The Journal of Neuroscience*, *23*(32), 10274–10282.

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

Hare, T. A., Tottenham, N., Davidson, M. C., Glover, G. H., & Casey, B. J. (2005). Contributions of amygdala and striatal activity in emotion regulation. *Biological Psychiatry*, *57*(6), 624–632.

Hariri, A. R., Tessitore, A., Mattay, V. S., Fera, F., & Weinberger, D. R. (2002). The Amygdala Response to Emotional Stimuli: A Comparison of Faces and Scenes. *NeuroImage*, *17*(1), 317–323. http://doi.org/10.1006/nimg.2002.1179

Harris, L. T., & Fiske, S. T. (2007). Social groups that elicit disgust are differentially processed in mPFC. *Social Cognitive and Affective Neuroscience*, *2*(1), 45–51.

Izard, C. E. (2007). Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on Psychological Science*, *2*(3), 260–280.

Kawagoe, R., Takikawa, Y., & Hikosaka, O. (1998). Expectation of reward modulates cognitive signals in the basal ganglia. *Nature Neuroscience*, *1*(5), 411–416.

Kédia, G., Berthoz, S., Wessa, M., Hilton, D., & Martinot, J.-L. (2008). An agent harms a victim: a functional magnetic resonance imaging study on specific moral emotions. *Journal of Cognitive Neuroscience*, *20*(10), 1788–1798.

Ketelaar, T., & Au, W. T. (2003). The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. *Cognition and Emotion*, *17*, 429–453.

Lindsey, L. L. (2005). Anticipated guilt as behavioral motivation. *Human Communication Research*, *31*(4), 453–481.

Malinowski, C. I., & Smith, C. P. (1985). Moral reasoning and moral conduct: An investigation prompted by Kohlberg's theory. *Journal of Personality and Social Psychology*, *49*(4), 1016 – 1027.

Maratos, E. J., Dolan, R. J., Morris, J. S., Henson, R. N. A., & Rugg, M. D. (2001). Neural activity associated with episodic memory for emotional context. *Neuropsychologia*, *39*(9), 910–920.

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*(2), 150–157.

Miller, C. (2010). Guilt and helping. *International Journal of Ethics*, *6*(2-3), 231–252.

Moll, J., & de Oliveira-Souza, R. (2007). Moral judgments, emotions and the utilitarian brain, 319–321.

Morey, R. A., McCarthy, G., Selgrade, E. S., Seth, S., Nasser, J. D., & LaBar, K. S. (2012). Neural systems for guilt from actions affecting self versus others. *Neuroimage*, *60*(1), 683–692.

Nelissen, R. (2011). Guilt induced self-punishment as a sign of remorse. *Social Psychological and Personality Science*.

Nelissen, R. M. A. (2012). Guilt-induced self-punishment as a sign of remorse. *Social Psychological and Personality Science*, *3*(2), 139–144.

Nelissen, R., & Zeelenberg, M. (2009). When guilt evokes self-punishment: Evidence for the existence of a "Dobby Effect." *Emotion*, *9*(1), 118–122.

Nummenmaa, L., & Calder, A. J. (2009). Neural mechanisms of social attention. *Trends in Cognitive Sciences*, *13*(3), 135–143.

O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience*, *4*(1), 95–102.

Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007). The enigmatic temporal pole: a review of findings on social and emotional processing. *Brain*, *130*(7), 1718–1731.

Paulmann, S., Pell, M. D., & Kotz, S. A. (2008). Functional contributions of the basal ganglia to emotional prosody: evidence from ERPs. *Brain Research*, *1217*, 171–178.

Pell, M. D., & Leonard, C. L. (2003). Processing emotional tone from speech in Parkinson's disease: A role for the basal ganglia. *Cognitive, Affective, & Behavioral Neuroscience*, *3*(4), 275–288.

Pelphrey, K. A., & Carter, E. J. (2008). Brain mechanisms for social perception. *Annals of the New York Academy of Sciences*, *1145*(1), 283–299.

Regan, D. T., Williams, M., & Sparling, S. (1972). Voluntary expiation of guilt: A field experiment. *Journal of Personality and Social Psychology*, *24*(1), 42–45. http://doi.org/10.1037/h0033553

Saxe, R. (2010). The right temporo-parietal junction: a specific brain region for thinking about thoughts. *Handbook of Theory of Mind*.

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*(1), 43–50.

Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience*, *10*(5), 640–656.

Takahashi, H., Yahata, N., Koeda, M., Matsudae, T., Asai, K., & Okubo, Y. (2004). Brain activation associated with evaluative processes of guilt and embarrassment: an fMRI study. *Neuroimage*, *23*(3), 967–974.

Talairach, J., & Tournoux, P. (1988). Co-planar stereotaxic atlas of the human brain. 3-Dimensional proportional system: an approach to cerebral imaging. Retrieved from http://www.citeulike.org/group/96/article/745727

Tangney, J. P., & Dearing, R. L. (2002). *Shame and guilt*. New York, NY, US: The Guilford Press.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, 35–57.

Wager, T. D., & Feldman-Barrett, L. (2004). From affect to control: Functional specialization of the insula in motivation and regulation. *Published Online at PsycExtra*. Retrieved from http://affective-science.org/pubs/2004/Wager_Edfest_submitted_copy.pdf

Wagner, U., N'Diaye, K., Ethofer, T., & Vuilleumier, P. (2011). Guilt-specific processing in the prefrontal cortex. *Cerebral Cortex*, bhr016.

Zahn, R., Moll, J., Paiva, M., Garrido, G., Krueger, F., Huey, E. D., & Grafman, J. (2009). The neural basis of human social values: evidence from functional MRI. *Cerebral Cortex*, *19*(2), 276–283.

Zemack-Rugar, Y., Bettman, J. R., & Fitzsimons, G. J. (2007). The effects of nonconsciously priming emotion concepts on behavior. *Journal of Personality and Social Psychology*, *93*(6), 927–939. http://doi.org/10.1037/0022-3514.93.6.927

Zhong, C.-B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, *313*(5792), 1451–1452.

## Figure Legends

**Figure 1.** Demonstrates the functional imaging block design. Presentations lasted 14s. Participants then reflected on the memory/scenario for 10s before being presented with a crosshair during which they were asked to clear their minds.

**Figure 2**. Results of the Guilt Memory/Neutral Memory Contrast. Effects are thresholded at $p< 0.001$, with a minimum cluster-size of 23 voxels. A. Left hemispheric activity of the TPJ (8). B. Right hemispheric activity of the OFC (1) and the Temporal Poles (6) C. Sagittal view of hemispheric activity of ACC (2), mPFC (5) and precuneus (7). D. Axial view of hemispheric activity of the amygdala (3) and insula (4).

**Figure 3.** Results of the Guilt Memory/Guilt Hypothetical Contrast. Effects are thresholded at $p<0.001$, with a minimum cluster-size of 23 voxels. A. Left hemisphere showing increased activity of the OFC (4) and the temporal pole (5). B. Sagittal view showing increased activity of ACC (1) and the precuneus (6). C. Axial view showing increased activity of the insula (2) and the caudate nucleus (3).

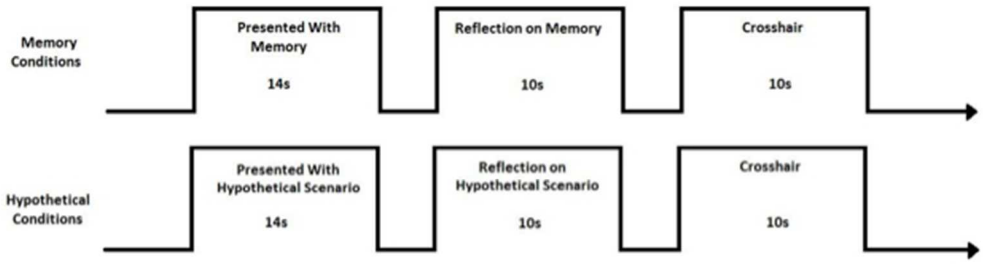
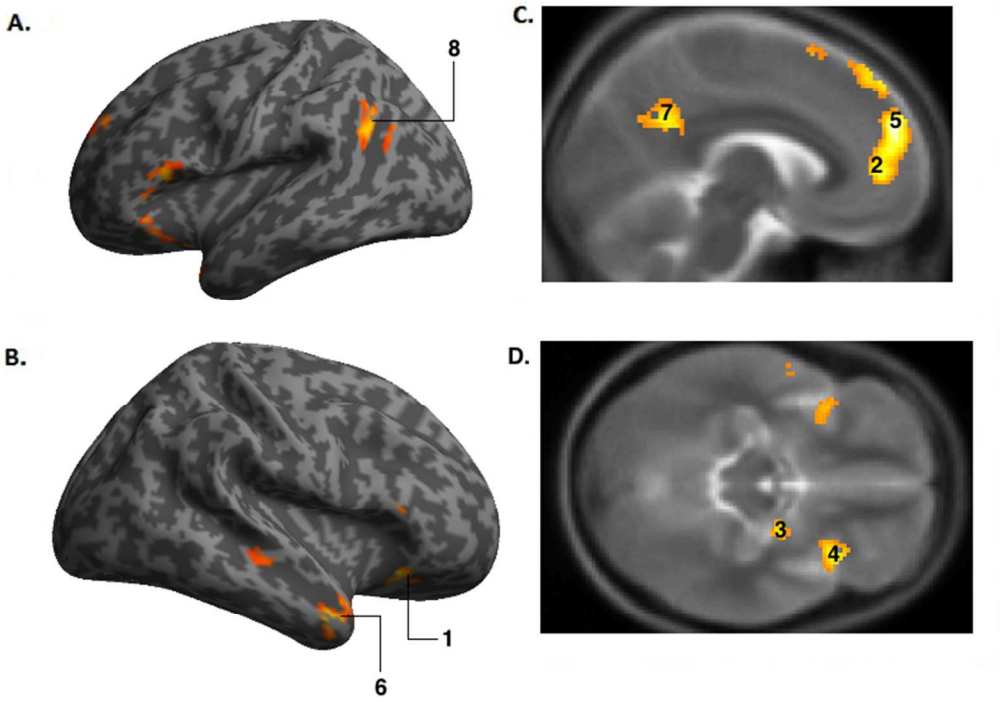



Figure 1. Demonstrates the functional imaging block design. Presentations lasted 14s. Participants then reflected on the memory/scenario for 10s before being presented with a crosshair during which they were asked to clear their minds.
221x67mm (72 x 72 DPI)



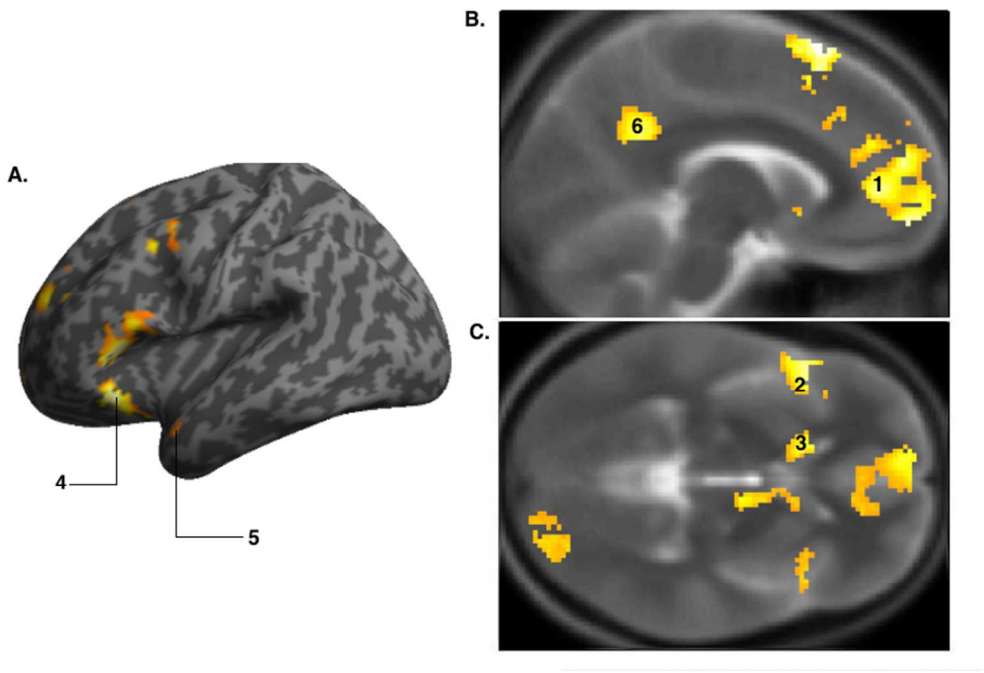

Results of the Guilt Memory/Neutral Memory Contrast. Effects are thresholded at p< 0.001, with a minimum cluster-size of 23 voxels. A. Left hemispheric activity of the TPJ (8). B. Right hemispheric activity of the OFC (1) and the Temporal Poles (6) C. Sagittal view of hemispheric activity of ACC (2), mPFC (5) and precuneus (7). D. Axial view of hemispheric activity of the amygdala (3) and insular (4).
279x204mm (72 x 72 DPI)

**Table 1**

*The Name, Hemisphere, and Z-score of Regions Associated with Increased Activity Following Guilt Memories/Neutral Memories Contrasts.*

| | ***Guilt Memory Reflect vs. Neutral Memory Reflect*** | | | |
|---|---|---|---|---|
| Figure Label | Brain Areas (x,y,z)(BA) | Hemi-sphere | Z-score | Voxels |
| 1 | Orbital Frontal Cortex* | | | |
| | (36, 24, -16)(BA47) | R | 3.85 | 21 |
| 2 | Anterior Cingulate Cortex* | | | |
| | (-6, 50, 4)(BA32) | L | 4.44 | 157 |
| 3 | Amygdala* | | | |
| | (26, -4, -14) | R | 3.92 | 6 |
| 4 | Insular* | | | |
| | (36, 20, -14) (BA47) | R | 4.40 | 87 |
| 5 | Medial Prefrontal Cortex* | | | |
| | (-4, 56, 28)(BA9) | L | 5.24 | 1145 |
| | (8, 54, 24)(BA9) | R | 4.79 | 1145 |
| 6 | Temporal Poles* | | | |
| | (-28, 16, 18)(BA13) | L | 4.00 | 130 |
| | (50, 10, -28)(BA21) | R | 4.67 | 177 |
| 7 | Precuneus* | | | |
| | (-6, -52, 30)(BA31) | L | 4.22 | 504 |
| 8 | Temporo-Parietal Junction* | | | |
| | (-50, -60, 28) | L | 4.17 | 104 |

*Indicates ROI, Peak-level threshold $p_{corr}<0.05$, >23 contiguous voxels. Coordinates (x, y, z) are in MNI space (Montreal Neurological institute), BA approximate Brodmann Area.




Results of the Guilt Memory/Guilt Hypothetical Contrast. Effects are thresholded at $p<0.001$, with a minimum cluster-size of 23 voxels. A. Left hemisphere showing increased activity of the OFC (4) and the temporal pole (5). B. Sagittal view showing increased activity of ACC (1) and the precuneus (6). C. Axial view showing increased activity of the insular (2) and the caudate nucleus (3).
301x203mm (72 x 72 DPI)

**Table 2**

*The Name, Hemisphere, and Z-score of Regions Associated with Increased Activity Following Guilt Memories/Guilt Hypothetical Contrasts*

| ***Guilt Memory Reflect vs. Guilt Hypothetical Reflect*** | | | | |
|---|---|---|---|---|
| Figure Label | Brain Areas (x,y,z)(BA) | Hemi-sphere | Z-score | Voxels |
| 1 | Anterior Cingulate Cortex* | | | |
| | (-8, 46, 6) | L | 4.50 | 896 |
| | (8, 30, 22) | R | 3.90 | 896 |
| 2 | Insular* | | | |
| | (-44, 18, 2) | L | 4.59 | 212 |
| | (32, 18, -16) | R | 4.60 | 250 |
| 3 | Basal Ganglia, caudate* | | | |
| | (-14, 18, -6) | L | 4.68 | 94 |
| 4 | Orbital Frontal Cortex* | | | |
| | (-2, 66, -2)(BA10) | L | 4.01 | 7 |
| | (31, 20, 13)(BA13) | R | 3.94 | 7 |
| 5 | Temporal Pole* | | | |
| | (-40, 18, -16) | L | 4.66 | 310 |
| 6 | Precuneus* | | | |
| | (-8, -50, 32)(BA31) | L | 4.19 | 194 |

*Indicates ROI, *Indicates ROI, peak-threshold $p_{corr}$<0.05, >23 contiguous voxels. Coordinates (x, y, z) are in MNI space (Montreal Neurological institute), BA, approximate Brodmann Area.