SOCIAL NEUROSCIENCE, 2014
Vol. 9, No. 2, 203–218, http://dx.doi.org/10.1080/17470919.2013.878750

Neural correlates of experienced moral emotion: An fMRI investigation of emotion in response to prejudice feedback

Melike M. Fourie[1,2], Kevin G. F. Thomas[1], David M. Amodio[3], Christopher M. R. Warton[4], and Ernesta M. Meintjes[4]

[1]Department of Psychology, University of Cape Town, Cape Town, South Africa
[2]Department of Psychology, University of the Free State, Bloemfontein, South Africa
[3]Department of Psychology and Center for Neural Science, New York University, New York, NY, USA
[4]Department of Human Biology, University of Cape Town, Cape Town, South Africa

Guilt, shame, and embarrassment are quintessential moral emotions with important regulatory functions for the individual and society. Moral emotions are, however, difficult to study with neuroimaging methods because their elicitation is more intricate than that of basic emotions. Here, using functional MRI (fMRI), we employed a novel social prejudice paradigm to examine specific brain regions associated with real-time moral emotion, focusing on guilt and related moral-negative emotions. The paradigm induced intense moral-negative emotion (primarily guilt) in 22 low-prejudice individuals through preprogrammed feedback indicating implicit prejudice against Black and disabled people. fMRI data indicated that this experience of moral-negative emotion was associated with increased activity in anterior paralimbic structures, including the anterior cingulate cortex (ACC) and anterior insula, in addition to areas associated with mentalizing, including the dorsomedial prefrontal cortex, posterior cingulate cortex, and precuneus. Of significance was prominent conflict-related activity in the supragenual ACC, which is consistent with theories proposing an association between acute guilt and behavioral inhibition. Finally, a significant negative association between self-reported guilt and neural activity in the pregenual ACC suggested a role of self-regulatory processes in response to moral-negative affect. These findings are consistent with the multifaceted self-regulatory functions of moral-negative emotions in social behavior.

Keywords: Anterior cingulate cortex; Behavioral inhibition; fMRI; Guilt; Moral emotions.

Moral emotions are powerful motivational forces that help us distinguish between right and wrong and to act adaptively in response to both moral transgressions and triumphs (Moll, Zahn, de Oliveira-Souza, Krueger, & Grafman, 2005). Guilt, shame, embarrassment, and pride are quintessential moral emotions that belong to the family of self-conscious emotions. Following a moral event, these emotions provide immediate feedback on behavior while promoting learning, through punishment or reinforcement, in ways that typically function to preserve social bonds (Tracy & Robins, 2004).

The neural underpinnings of moral emotions remain poorly understood, however, largely because of the challenges involved in eliciting genuine emotional responses in real time in a neuroimaging

Correspondence should be addressed to: Melike M. Fourie, Department of Psychology, University of Cape Town, Rondebosch, Cape Town 7701, South Africa. E-mail: marethem@gmail.com

This work was supported by the National Research Foundation (NRF) of South Africa, the Oppenheimer Memorial Trust, the AW Mellon Foundation, the University of Cape Town, and the University of the Free State.

context. The elicitation of authentic moral emotion also often requires deception and/or complex social interaction, which are difficult to simulate within the scanner. Furthermore, it is difficult, if not impossible, to elicit a "pure" moral emotion, because different emotions often co-occur (Izard, 1991). Following a social transgression, for example, a person may feel guilty about his wrongdoing, while at the same time feeling embarrassed/ashamed because the event was witnessed by others (Finger, Marsh, Kamel, Mitchell, & Blair, 2006). Researchers often try to overcome this issue by showing that the effects for one emotion emerge above and beyond those of other emotions (e.g., Amodio, Devine, & Harmon-Jones, 2007), yet the extent to which these emotions are separable at the level of neural mechanism is often difficult to determine.

To date, most neuroimaging studies investigating moral emotions have either (a) asked participants to relive a previous emotional episode (e.g., Shin et al., 2000; Wagner, N'Diaye, Ethofer, & Vuilleumier, 2011), which may differ phenomenologically from the original emotional encounter (Herrald & Tomaka, 2002); or (b) used paradigms where emotive sentences or vignettes are presented to participants in the scanner (e.g., Moll et al., 2007; Morey et al., 2012). Paradigms that use such descriptive scenarios, however, tend to focus on the interpretation of socially relevant stimuli, rather than on the elicitation of emotional responses that motivate social behavior. Moreover, participants are not placed within realistic emotion-evoking situations that are relevant to them personally (i.e., as the person who performed the embarrassing/shameful act). Rather, they are asked to *imagine* hypothetical events with themselves as the protagonist, which may not elicit any strong emotion.

Moral emotions to social transgressions: The case of prejudice

In the present neuroimaging study, we employed a social prejudice paradigm to elicit current, self-relevant moral-negative emotion. Our paradigm was based on the well-documented finding that individuals who renounce prejudice tend to show socially biased tendencies on measures that tap automatic or implicit mental associations (Amodio, Harmon-Jones, & Devine, 2003; Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002). In particular, discrepancies between personal standards and actual responses (of prejudice) typically give rise to feelings of guilt when the individual's nonprejudiced standards are well-internalized (Devine, Monteith, Zuwerink, & Elliot, 1991). Guilt evoked by our paradigm could thus be described as deontological rather than altruistic, because it resulted from the violation of inner moral values, rather than from interpersonal transgressions.

During our functional MRI (fMRI) scans, a carefully selected sample of low-prejudice individuals performed several modified implicit association tests (IATs; Greenwald, McGhee, & Schwartz, 1998) purported to assess their attitudes toward various social groups. Instead of providing participants with their actual IAT results, however, all participants received preprogrammed bogus feedback. Our main interest was in guilt-eliciting feedback that indicated the participant was socially prejudiced toward Black and disabled people, thus contradicting the participant's nonprejudiced internal standards. As a contrasting control condition, we also included pride-eliciting feedback that indicated egalitarian responses on the IAT (Moll, De Oliveira-Souza, & Zahn, 2008). Additionally, we included a neutral condition, containing only neutral feedback in response to IATs of no topical importance, as another control condition.

Physiological and neural correlates of guilt

Although we are interested in moral-negative emotion broadly, we anticipated, based on previous findings (Amodio et al., 2007; Devine et al., 1991; Fourie, Kilchenmann, Malcolm-Smith, & Thomas, 2012), that guilt elicited by our prejudice paradigm would be the strongest emotion driving neural activation. We therefore based our neural activation hypotheses primarily on theory related to guilt.

Guilt is an aversive feeling associated with the belief that one has transgressed a personally relevant standard, and should have thought, felt, or acted differently (Kubany & Watson, 2003). As mentioned above, guilt is also a self-conscious emotion. Such emotions are founded in social relationships, and are therefore associated intimately with others' evaluation of self (Leary, 2004). The ability to evaluate ourselves through the eyes of others, a hallmark of Theory of Mind (ToM), is thus essential in recognizing or experiencing self-conscious emotions (Leary, 2007). Indeed, several neuroimaging studies of self-conscious emotions have detected activation in putative ToM areas, including the dorsomedial prefrontal cortex (DMPFC), posterior superior temporal sulci (STS), temporal poles, posterior cingulate cortex (PCC), and precuneus (Basile et al., 2011; Kédia, Berthoz, Wessa, Hilton, & Martinot, 2008).

Various researchers have argued that the feeling of guilt functions as a punishment cue that serves to heighten self-focus and to inhibit unwanted behavior (Monteith, 1993; Monteith, Ashburn-Nardo, Voils, & Czopp, 2002). In this regard, clinical studies point to specific involvement of the orbitofrontal cortex (OFC) in signaling appropriate social behavior: Patients with OFC dysfunction have reduced sensitivity to social norms, display an abnormally diminished sense of guilt, and are impaired at altering their behavior in response to socially aversive cues (Beer, John, Scabini, & Knight, 2006; Blair & Cipolotti, 2000; Krajbich, Adolphs, Tranel, Denburg, & Camerer, 2009). Neuroimaging studies also suggest that the lateral OFC is sensitive to a wide range of punishing stimuli (Kringelbach & Rolls, 2004).

Finally, guilt has been described variously as an emotion that inhibits transgressive behavior or that facilitates prosocial behavior (de Hooge, Zeelenberg, & Breugelmans, 2007; Monteith, 1993). Amodio et al. (2007) integrated these accounts by arguing that guilt functions dynamically, starting with behavior inhibition and transforming into approach-oriented, conciliatory behavior when an opportunity for amendment appears. Consistent with this view, Fourie et al. (2011) found that individuals who experienced heightened levels of real-time guilt also scored high on the Carver and White (1994) Behavior Inhibition System (BIS) scale, which has been associated with a conflict-monitoring mechanism via the supragenual ACC (supraACC) (Amodio, Master, Yee, & Taylor, 2008). In turn, conflict monitoring is readily associated with the interruption of action (van Veen & Carter, 2002).

In sum, we hypothesized that guilt would be associated with heightened activation in areas implicated in the neural substrates of self-reflection and mentalizing (DMPFC, PCC, and precuneus), social response reversal (lateral OFC), BIS-related conflict monitoring (supraACC), and heightened physiological arousal (ACC and insula). We also anticipated that participants might engage in emotion-regulatory strategies, which may depend on prefrontal and ACC control systems (Drabant, McRae, Manuck, Hariri, & Gross, 2009).

MATERIALS AND METHODS

Participants

Pre-experimental screening procedure

Six months prior to scanning, we conducted a web-based survey (N = 445) open to female students seeking to obtain course credit. We recruited a female-only sample to reduce possible sex differences in emotion physiology and experience (Manstead, 1992). The survey contained measures aimed at identifying individuals low in social prejudice who would be sensitive to bogus IAT feedback indicating prejudice, and hence likely to experience intense guilt.

To be eligible for participation in the fMRI study, volunteers had to be White, right-handed, heterosexual, and non-Jewish, and they had to self-report a neutral/positive attitude toward religion. They also had to have nonprejudiced (positive) attitudes toward Black people, disabled people, homosexual people, and Jewish people, which we assessed using four separate rating thermometers (Herek, 2000). These criteria were put in place to prevent disagreement between participants' own sexual/religious orientations and the IAT feedback, and to thus ensure the validity of our emotion manipulations. After applying these eligibility criteria, 98 remained in the pool of potential fMRI participants.[1]

To optimize the effectiveness of our prejudice manipulation further, we selected, using the Internal and External Motivation to Respond Without Prejudice scales (IMS/EMS; Plant & Devine, 1998), individuals who were highly motivated to respond without prejudice. The IMS assesses personal reasons for trying to respond in a nonprejudiced manner toward Blacks, whereas the EMS provides an index of participants' sensitivity to external pressures to appear nonprejudiced. Individuals with high IMS as well as high EMS scores are thought to have egalitarian values integrated into their self-concepts, yet often respond in ways discrepant from their personal standards and experience guilt as a result of this personal failure (Plant & Devine, 1998). Hence, we retained in our fMRI sample those individuals whose IMS and EMS scores were significantly above the scales' midpoints ($ps < .05$; IMS: $M = 7.44$, $SD = 1.08$; EMS: $M = 5.22$, $SD = 1.70$).[2]

Finally, the survey assessed participants' sensitivity to punishment and reward, using the Carver and White Behavioral Inhibition System and Behavioral Activation System scales (BIS/BAS). Because individuals more sensitive to punishment experience greater guilt (Fourie et al., 2011, 2012), we excluded individuals with extremely low BIS scores (>2 *SD* below the survey sample mean; $M = 23.55$, $SD = 2.65$).

[1] The large number of excluded individuals reflects both the multicultural and prejudiced nature of the South African population.

[2] These means resemble the average IMS and EMS scores typically seen for US samples, however (Devine et al., 2002).

Final sample of fMRI participants

Using the procedures described above, we selected 25 low-prejudice individuals to complete the full fMRI testing procedure. These participants were without any previously diagnosed neurological or psychiatric disorders, and none was on medication. They were also screened for the presence of depressive symptoms using the Beck Depression Inventory-II (M = 6.64, SD = 5.02; Beck, Steer, & Brown, 1996).

Data from three participants were excluded before statistical analysis because of either OFC signal loss (n = 1) or because the participant did not believe the prejudice feedback manipulation (n = 2). The final sample for data analysis thus consisted of 22 participants (age: M = 19.32 years, SD = 1.11).

All study procedures were approved by the Human Research Ethics Committee of the University of Cape Town's Faculty of Health Sciences.

IAT paradigm

The IAT is a dual categorization task designed to measure the strengths of implicit associations between mental representations of objects (Greenwald et al., 1998). When performing an IAT, one has to make timed responses to two critical blocks of trials: *congruent blocks*, in which concepts that are strongly associated for most respondents are paired, and *incongruent blocks*, in which less strongly associated concepts are paired. For example, in a commonly used racial IAT, White participants are instructed to classify, using two keys on a computer keyboard, positive and negative words as well as faces of White and Black individuals. For congruent trials, positive words and White faces share the first key, while negative words and Black faces share the second key. For incongruent trials, negative words and White faces share the first key, while positive words and Black faces share the second.

Because most White people hold more positive attitudes toward Whites than Blacks, participants usually find it easier, and are therefore faster, at responding to trials in the congruent than in the incongruent block. The difference in response times between trials in the incongruent and congruent blocks represents a difference in participants' evaluative associations with White and Black faces; this is taken to reflect the participants' degree of implicit (race) bias (i.e., the IAT effect). It is important to note that here we used IATs only as part of the emotion manipulation, as it provided a plausible basis for the manipulated feedback participants were to receive. Therefore, we were not interested in participants' actual IAT effects. Instead, we were interested in neural activity in response to preprogrammed bogus feedback that differed depending on the specific condition.

Our fMRI protocol consisted of six different IATs: two each in Neutral, Egalitarian, and Prejudice feedback conditions (Fourie et al. (2012) describe a validation study of this protocol). While the Neutral feedback condition consisted of IATs on topics (e.g., facial hair) for which no publicly endorsed responses exist, the Egalitarian and Prejudice feedback conditions consisted of IATs on more sensitive social topics (e.g., race and sexuality). The stimuli employed in each IAT button-press task consisted of 16 words and 8 colour images that participants had to sort into categories as quickly as possible. Words (or attributes) were categorized as either "good" (e.g., *joy*, *love*, *peace*) or "bad" (e.g., *agony*, *terrible*, *awful*). Images were either pictures or symbols of people from each target social group (see Supplementary Material for details).

We selected *race* and *disability* as the two IAT topics in the Prejudice condition. Preprogrammed feedback following these IATs were tailored to elicit guilt, and thus contradicted participants' beliefs that they held egalitarian attitudes toward Black and physically/intellectually disabled people. Data from Nosek et al. (2007) indicate that race and disability are topics for which most people tend to show a strong bias on implicit measures that conflict with their explicit beliefs. Hence, we felt that participants in our study would be likely to believe feedback indicating their bias after completing IATs on those topics, and would feel guilty, above other emotional responses, about harboring such biases.

We selected *religion* and *sexuality* as the two IAT topics in the Egalitarian condition. Preprogrammed feedback following these IATs confirmed participants' beliefs that they held egalitarian attitudes toward gay and Jewish people, and contained praiseworthy statements to elicit pride and satisfaction. Nosek et al. (2007) reported a strong correspondence between participants' nonprejudiced attitudes toward homosexual and Jewish people on both implicit and explicit measures. Hence, we considered these topics well-suited for our Egalitarian condition.

Finally, we selected *facial hair* and *glasses* as the two IAT topics in the Neutral condition. Preprogrammed feedback following these IATs indicated no preference for people with or without facial hair, or for people with or without glasses. No significant affect was expected in this condition because the IAT topics did not involve socially sensitive issues.

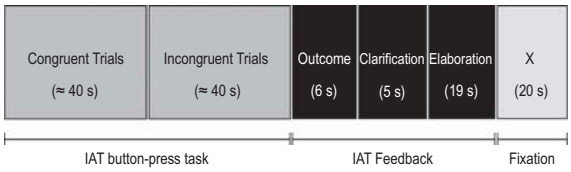

Figure 1. Time line for each IAT during fMRI scanning. Our analyses compared the 30-s IAT feedback intervals (black background) between the Neutral, Egalitarian, and Prejudice conditions. During these intervals, preprogrammed result sentences matched for length and complexity were displayed, so that any difference in brain activation between conditions was determined entirely by the emotional overlay evoked by the specific condition.

IATs employed in the fMRI paradigm consisted of only two critical blocks of trials where concepts are paired, either in congruent or incongruent fashion, followed by preprogrammed response feedback. Each IAT block included 20 trials, and we counterbalanced presentation of congruent and incongruent blocks across IATs. The design was self-paced. Targets thus remained on-screen until the correct response was given (a red error sign appeared following errors). A 350-ms fixation cross separated all trials.

Upon the completion of each IAT, the participant received a 30-s feedback presentation (Figure 1): (a) a sentence, appearing for 6 s, that indicated the result of the specific IAT (Outcome), (b) additional condition-specific feedback appearing for 5 s (Clarification) [e.g., *This is a regular/neutral response*; or *This is a low-prejudice/positive response!*; or *This is a high-prejudice/negative response!*], and, finally, (c) three formally structured sentences, appearing for 19 s, that presented extended feedback (Elaboration). The latter sentences elaborated on how the IAT result could be interpreted in terms of the participant's attitude and personality, and served to maintain/intensify the induced emotion. Feedback sentences were constructed to be similar in length, word structure, and readability across all conditions (see Supplementary Material).

Experimental procedure

On arrival, participants provided informed consent and received general instructions regarding the procedures. To heighten the realism of the emotion manipulation, they were told that the IATs tested their unconscious prejudice against various social groups, and that the purpose of the study was to identify brain activation associated with low- versus high-prejudice behavior. Each participant was informed that, based on her positive attitude toward all social groups assessed in the web-based survey, she was categorized as part of the study's low-prejudice group and was expected to respond accordingly in the scanner. Participants then completed practice versions of each of the six different IATs, each without performance feedback, to familiarize them with the dual-categorization task.

fMRI procedure

In the scanner, all IATs were pseudo-randomly arranged into three functional runs. Participants viewed stimuli through a mirror system mounted to the head coil. We used *E-Prime* software (Psychology Software Tools, Inc.) to display stimuli and to record participants' behavioral responses via a button box. Two variants of each of the six different IATs were presented, resulting in a total of 12 IATs.[3] Each run consisted of four different IATs, with at least one from each of the Neutral, Egalitarian, and Prejudice conditions, in counterbalanced order (but always beginning with a Neutral-condition IAT). Each IAT was interleaved with 20-s fixation periods; an initial 20-s fixation period served as the baseline. Participants were instructed to attend to the IAT feedback without any response.

Emotion ratings

After each functional run, participants reported their emotional response to the foregoing IAT feedback on eight separate items (arousal, anger, anxiety, pride, satisfaction, embarrassment, guilt, and shame), rated on 0 (*not at all*) to 9 (*very much*) visual analog scales. The set of ratings was followed by a 20-s rest period. We computed composite indices of moral-positive affect (mean of ratings for pride and satisfaction) and moral-negative affect (mean of ratings for guilt, embarrassment, and shame) from these ratings. We obtained baseline ratings of all items prior to the start of the IAT paradigm.

Post-scan manipulation check

After the scan, each participant made a forced-choice decision as to which emotion(s) she had experienced most intensely after each IAT. Each participant was given a choice of the emotion items listed above, plus "neutral", and then rated the intensity of

[3] IAT variants of the same topic contained different elaborated feedback sentences, and the order of presentation of congruent and incongruent trials in these variants was counterbalanced.

that emotion, from 1 (*not at all*) to 5 (*very much*). This measure provided a manipulation check to validate the elicitation of our target emotion of guilt. Participants were then probed for suspicion, using a funneled approach (Harmon-Jones, Amodio, & Zinner, 2007), before being debriefed fully.

Physiological measure

To obtain a physiological measure of arousal, we assessed peripheral pulse data continuously during the fMRI protocol (Vrana, Cuthbert, & Lang, 1989). We used a pulse oximeter (Siemens, Erlangen, Germany) with sensors placed over participants' left ring finger, and calculated heart rate from the pulse intervals. Data from only 16 participants were deemed reliable, however, due to inherent difficulties in collecting analyzable physiological recordings inside an MRI scanner. Nevertheless, this measure provided useful data for validating the manipulations.

fMRI image acquisition and analysis

We acquired MRI data on a 3T Allegra head-only system (Siemens, Erlangen, Germany). Sessions began with a high-resolution anatomical scan acquired with a T_1-weighted sequence (3D mprage, 160 slices, TR = 2300 ms, TE = 3.93 ms, flip angle = 9°, voxel size = 1 × 1 × 1 mm^3). Functional images covering the whole brain were then acquired with a T_2*-weighted echo-planar imaging (EPI) sequence using blood-oxygenation-level-dependent (BOLD) contrast (34 interleaved slices, slice thickness = 3 mm, gap = 0.9 mm, TR = 2000 ms, TE = 30 ms, flip angle = 90°, field of view = 200 × 200 mm, voxel size = 3.125 × 3.125 × 3 mm^3), while participants performed the task. The first four volumes of each functional run were discarded to allow for T_1 equilibration effects.

We performed all fMRI analyses using Brain Voyager QX, version 2.4 (Brain Innovation, Maastricht, Netherlands). Images were corrected for different slice acquisition times and linear trends, and temporally smoothed with a high-pass filter of 2 cycles/point. Images were motion-corrected relative to the first volume of the functional run with trilinear estimation and interpolation. We excluded two runs from subsequent analyses based on our motion criteria (>3 mm displacement or 3.0° rotation within a functional run). Each participant's functional data sets were then co-registered with her anatomical MRI and spatially normalized to Talaraich space, during which voxels are interpolated to 3 × 3 × 3 mm^3.

Whole-brain group analyses were first performed with a random effects analysis of variance using the general linear model (GLM) with predictors corresponding to known experimental blocks convolved by the standard hemodynamic response function. We defined separate predictors for each 30-s feedback period (i.e., feedback for the Neutral, Egalitarian, and Prejudice condition IATs), and for each corresponding fixation period (i.e., fixations following the Neutral, Egalitarian, and Prejudice conditions). An additional predictor corresponded to the IAT button-press task. The six motion correction parameters were added as predictors, but were not of theoretical interest. The main blocks of interest were those corresponding to IAT feedback during which emotion elicitation was expected.

We performed second-level analyses using single-factor repeated-measures ANOVA. The resulting *F*-map showed overall effects of condition in extensive brain regions at $p < .05$, voxel-wise corrected for multiple comparisons using the false discovery rate (FDR). To assess specific condition effects, we contrasted the Prejudice and Egalitarian condition feedback periods against the Neutral condition feedback period. We applied cluster-level thresholding using the Monte Carlo simulation tool implemented in Brain Voyager to compute the minimum number of voxels required for significance at a corrected $p < .05$ (Forman et al., 1995). This tool applies smoothing using a Gaussian kernel at the full-width-at-half-maximum (FWHM) of the functional voxel. We first applied cluster-level thresholding at an uncorrected $p < .0001$ to detect distinct prefrontal clusters that merged into one large cluster at a more relaxed uncorrected threshold. We inspected event-related averaging plots for each of these prefrontal clusters to determine whether the percent signal increase in activated areas extended over the entire 30-s feedback period. We then also applied cluster-level thresholding at an uncorrected $p < .005$ to increase the power of our analysis.

We performed region of interest (ROI) analyses for all clusters. We performed a random effects ANOVA on the average signal in each cluster for each participant using the GLM described above; this analysis generated beta values that reflect the mean percent signal change for each condition. We computed the mean percent signal change for the contrast of the Prejudice condition feedback against the Neutral condition feedback for each ROI for each participant. Finally, we used zero-order correlation analysis to examine the relation between ROIs and (a) subjective emotion reports, and (b) personality constructs.

RESULTS

Behavioral results

Response time

Our analyses of response time data from the IAT button-press tasks confirmed that participants showed implicit bias against Black and disabled people on Prejudice condition IATs, but that neural activation detected during the Prejudice condition could not be attributed to effects of task difficulty (see Supplementary Material).

Subjective emotion ratings

Table 1 shows the mean within-scan emotion ratings for the different conditions.[4] Paired *t*-tests showed that guilt, embarrassment, shame, and anger (but not anxiety) increased significantly over baseline in the Prejudice condition, *t*s(20) > 7.50, *p*s < .001, *r*s > .86, and that pride and satisfaction increased significantly over baseline in the Egalitarian condition, *t*s(20) > 3.80, *p*s = .001, *r*s > 65. In the Neutral condition, there were no significant increases over baseline for these emotions (*p*s > .30).

We examined changes in specific emotion ratings acquired during the IAT paradigm using a 3 (condition: Prejudice, Egalitarian, and Neutral) × 4 (emotion type: moral-negative affect, moral-positive affect, anger, and anxiety) repeated-measures ANOVA on emotion difference scores. Results indicated that the overall two-way interaction was significant, $F(2.34, 44.77)$[5] = 79.13, MSE = 2.69, $p < .001$: the observed increase in moral-negative affect, compared to changes in anxiety, anger, and moral-positive affect, was greater in the Prejudice condition than in the Neutral and Egalitarian conditions (*p*s ≤ .001, *r*s > .66). Moreover, increases in moral-negative affect were highest in the Prejudice condition, even when controlling for increases in basic negative emotions (i.e., anger and anxiety), $F(2, 36) = 14.04$, MSE = .92, $p < .001$. Similarly, in the Egalitarian condition compared to the other conditions, the increase in moral-positive affect outweighed the changes in any other emotion (*p*s ≤ .001, *r*s > .64). One-way repeated-measures ANOVAs within the Prejudice and Egalitarian conditions confirmed that, respectively, moral-negative and moral-positive affect were elicited more strongly than any other emotion (*p*s < .01, *r*s > .54).

TABLE 1
Emotion ratings at baseline and during the IAT paradigm (*N* = 21)

		Condition		
	Baseline	*Neutral*	*Egalitarian*	*Prejudice*
Arousal	4.63 (1.55)	3.39 (1.46)	3.76 (1.75)	5.96 (1.31)
Moral-positive affect	4.59 (1.23)	5.02 (1.24)	6.05 (1.33)	2.36 (0.84)
Pride	4.49 (1.50)	4.90 (1.52)	6.04 (1.39)	2.34 (0.86)
Satisfaction	4.68 (1.34)	5.15 (1.25)	6.06 (1.45)	2.38 (0.98)
Moral-negative affect	2.23 (1.55)	2.23 (1.30)	2.19 (1.29)	6.59 (1.37)
Guilt	2.14 (1.61)	2.15 (1.34)	2.20 (1.40)	6.73 (1.76)
Shame	2.06 (1.62)	2.22 (1.30)	2.26 (1.30)	6.43 (1.47)
Embarrassment	2.49 (1.72)	2.34 (1.32)	2.11 (1.28)	6.60 (1.29)
Anxiety	5.17 (1.50)	3.33 (1.66)	3.34 (1.46)	5.83 (1.66)
Anger	1.83 (1.43)	2.17 (1.45)	2.10 (1.29)	5.19 (1.60)

Notes: Data presented are means, with standard deviations in parentheses. Ratings could vary between 1 (*not at all*) and 9 (*very much*).

Manipulation check

Because within-scan ratings of guilt, shame, and embarrassment were highly correlated (*r*s > .65, *p*s < .01), we examined post-scan manipulation-check data to determine which moral-negative emotion participants experienced most strongly. Table 2 presents participants' ratings of their most salient emotional experiences following each IAT. These data indicate that in the Neutral condition participants primarily felt "neutral," in the Egalitarian condition participants experienced mostly pride and satisfaction, and in the Prejudice condition 96% of participants predominantly experienced guilt (mean intensity = 3.88, *SD* = .85). Although the prejudice manipulation thus elicited a constellation of relevant moral emotions, it primarily elicited our emotion of focal interest, namely guilt.

Arousal

Emotion conditions differed significantly in terms of subjective arousal: the Prejudice condition was associated with significantly higher arousal ratings than baseline ($p = .002$, $r = .63$), whereas the Neutral and Egalitarian conditions were associated with significantly lower arousal ratings than baseline (*p*s < .01, *r*s > .54). Consistent with these data, within-

[4] One participant did not complete within-scan emotion ratings correctly, and so her data were excluded from all subsequent analyses of emotion ratings.

[5] The degrees of freedom were corrected using Greenhouse–Geisser estimates of sphericity ($\varepsilon = .37$).

TABLE 2
Post-scan manipulation check: Percentage of participants reporting a particular emotion after each IAT (N = 22)

		Emotion							
Condition	*IAT topics*	*Neutral*	*Anxiety*	*Satisfied*	*Pride*	*Guilt*	*Embarrassed*	*Shame*	*Anger*
Neutral	Facial hair	68%	5%	27%					
	Glasses	54%	—	45%	—	—	—	—	—
	M	**61%**	2%	36%	—	—	—	—	—
Egalitarian	Sexuality	—	—	54%	72%				
	Religion	—	—	50%	63%	—	—	—	—
	M			**52%**	**68%**	—	—	—	—
Prejudice	Race	—	5%	—	—	100%	9%	—	14%
	Disability	—	9%	—	—	91%	27%	14%	9%
	M		7%			**96%**	18%	7%	11%

Notes: Predominant emotions in each condition are in boldface. *M* = mean.

scan ratings of moral-negative affect correlated significantly with overall HR reactivity during the Prejudice condition feedback, even when changes in basic negative affect (i.e., anger and anxiety) were covaried, r = .56, p < .05 (HR data are presented in the Supplementary Material). Additional analyses revealed that guilt ratings, specifically, correlated significantly with HR reactivity (p < .05), even when changes in all other negative emotions (i.e., anger, anxiety, shame, and embarrassment) were covaried, r = .57, p < .05. By contrast, the effects of shame and embarrassment did not remain significant when other emotions were covaried. Hence, participants' subjective response to the prejudice manipulation was corroborated by their physiological response, which further suggested the acute experience of a guilty emotion.

fMRI results

Contrast of the Prejudice condition against the Neutral condition

Our main goal was to identify brain regions recruited during the experience of guilt, and so the most critical contrast was that comparing activation during the Prejudice condition feedback against that during the Neutral condition feedback. After applying cluster-size thresholding at p < .0001 (uncorrected), this comparison revealed significantly higher activation in three left prefrontal clusters: (a) an area comprising the dorsomedial prefrontal cortex (DMPFC; *xyz* peak –4,28,30), extending to supraACC; (b) an area centerd on the left paracingulate region of the DMPFC (peak –4,43,18); and (c) an area directly anterior to the genu of the corpus callosum, and thus referred to as pregenual ACC (pACC; peak –7,34,15). Extraction of parameter estimates of activity (beta values reflecting the mean activity) from these clusters showed that all three areas were strongly selective in their responsiveness to the prejudice manipulation. Furthermore, event-related averaging plots for these clusters confirmed that the signal increase extended over the entire 30-s feedback periods. Figure 2 and Table 3 summarize these results.

We conducted an additional test of the guilt contrast at the more relaxed threshold of p < .005 (uncorrected). This test yielded significant activation in an extended area within the bilateral PFC and ACC, the left anterior insula and posteriolateral OFC, the PCC, precuneus, and right thalamus (Figure 3 and Table 3). In addition, two areas were significantly activated, but did not survive cluster-size correction: the posterior inferior insula (peak 32,–20,6; cluster size = 116 mm^3), and the hippocampus (peak 23,–23,–12; cluster size = 104 mm^3).

Other contrasts of interest

To identify brain regions recruited during the experience of pride, we contrasted activation during the Egalitarian condition feedback against that during the Neutral condition feedback. This contrast did not reveal any significant activation clusters, even at p < .005 (uncorrected). These results imply that affective responses to the Egalitarian and Neutral conditions were too similar to yield significant differences in neural activation.

The direct contrast between the Prejudice and Egalitarian conditions at p < .005 (uncorrected) revealed significant activation in the DMPFC and ACC, in the left anterior insula/posteriolateral OFC, and in the PCC. In addition, we observed significant activation in the right posterior STS, and in the left caudate nucleus (see Supplementary Table S4). No

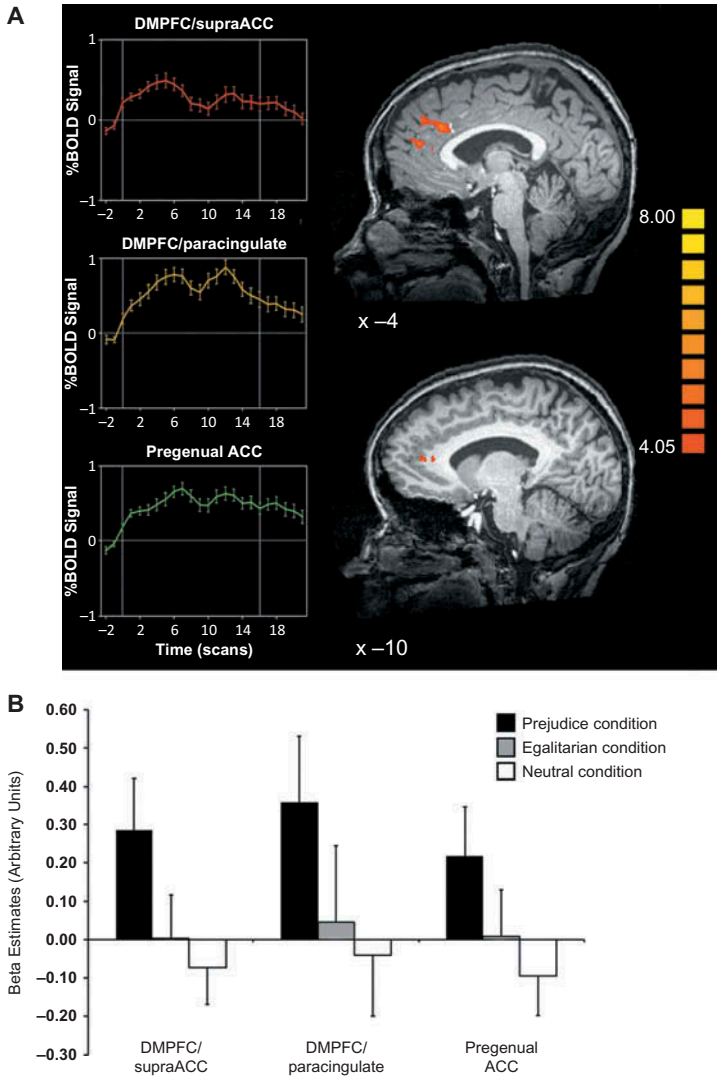

Figure 2. Guilt-specific processing in the prefrontal cortex (prejudice condition feedback > neutral condition feedback). (A) Activated clusters were in the DMPFC/supragenual ACC and DMPFC/paracingulate cortex (top panel), and pregenual ACC (bottom panel). Significant clusters survived Monte Carlo cluster-level thresholding at an uncorrected $p < .0001$ (clusters > 74 mm^3). Event-related plots are shown for each cluster of significant activation for the duration of the prejudice condition feedback period. (B) Parameter estimates of activation (betas) reflecting the average signal in each cluster for each condition. Error bars indicate 95% confidence intervals.

brain regions were more strongly activated during the Neutral or Egalitarian condition than during the Prejudice condition.

Finally, because our initial observations suggested that the effects of the prejudice manipulation may have persisted beyond the IAT feedback period, we also contrasted the Prejudice and Neutral condition fixation periods at $p < .005$ (uncorrected). Our results confirmed those initial observations: The comparison yielded significant activation in the DMPFC and posteriolateral OFC, as well as in several temporal lobe areas, including the bilateral anterior temporal lobes, middle STS, and TPJ (see Supplementary Table S5).

Correlational analysis

To explore further the role of the activated areas observed during the Prejudice condition, we computed Pearson's correlation coefficients between the mean percent signal changes for each ROI (see Table 3) and within-scan emotion ratings. Given the overlap among moral-negative emotions in participants' self-reports, we first examined associations between the composite measure of moral-negative emotion and neural activation, followed by analyses examining the more focal effects of specific moral-negative emotions. Results indicated that activity in

TABLE 3
Effects of guilty feeling (prejudice condition feedback > neutral condition feedback)

Region	BA	Hem.	Coordinates x	Coordinates y	Coordinates z	Cluster size (mm^3)	Max t
Cluster-size thresholding, $p < .0001$ (uncorrected; min cluster size 74 mm^3)							
DMPFC/supraACC	8/9, 24/32	L	–4	28	30	415	6.04
DMPFC/paracingulate cortex	9/32	L	–4	43	18	159	5.48
pACC	32	L	–7	34	15	197	4.96
Cluster-size thresholding, $p < .005$ (uncorrected; min cluster size 230 mm^3)							
DMPFC/paracingulate cortex/supraACC[a]	24/32/9	L/R	–4	28	30	10,008	6.04
Anterior insula/ posteriolateral OFC	13/47	L	–37	19	0	764	4.43
Posterior cingulate gyrus	23/31	L/R	–4	–20	33	1439	4.30
Precuneus	31/23	L/R	–1	–68	21	712	3.74
Mediodorsal thalamus	—	R	2	–23	3	316	4.82

Notes: Talaraich coordinates and *t*-score refer to the peak of each brain region. Reported clusters survived Monte Carlo cluster-level thresholding at an uncorrected $p < .0001$ and $p < .005$, respectively.
BA = Brodmann area; DMPFC = dorsomedial prefrontal cortex; supraACC = supragenual anterior cingulate cortex; pACC = pregenual anterior cingulate cortex; OFC = orbitofrontal cortex.
[a]The three local maxima detected at $p < .0001$ (uncorrected) are contained within this large prefrontal cluster.

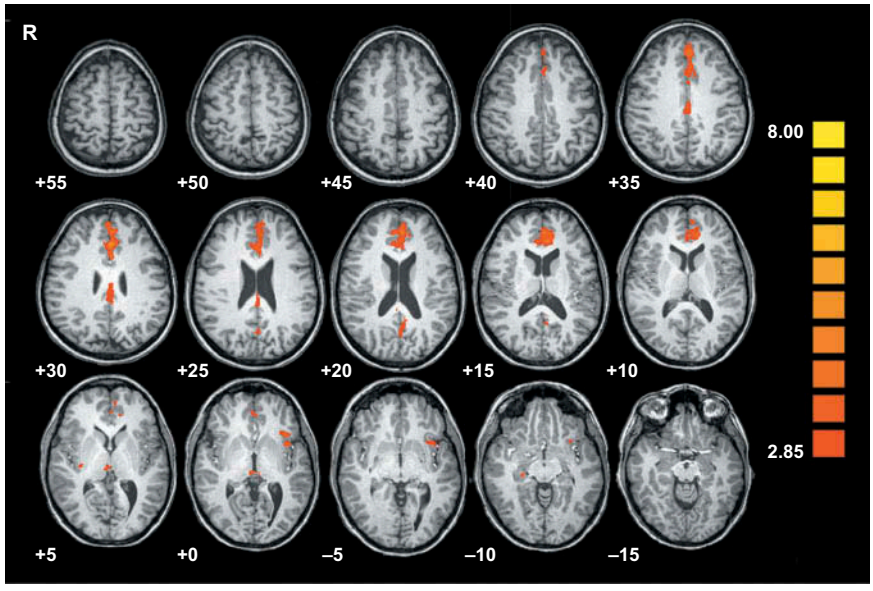

Figure 3. Guilt-specific processing in the whole brain (prejudice condition feedback > neutral condition feedback). Activated areas included the DMPFC/paracingulate cortex/supraACC, posterior cingulate gyrus, precuneus, left anterior insula and posteriolateral OFC, right thalamus, right posterior insula, and right hippocampus. All clusters, except the posterior insula and hippocampus, survived Monte Carlo cluster-level thresholding at an uncorrected $p < .005$ (clusters > 230 mm^3).

the pACC was specific to moral-negative affect: a significant negative linear relationship existed between activity in the pACC and moral-negative affect when changes in basic negative affect (i.e., anger and anxiety) were covaried, $r = -.61$, $p < .01$. Further analyses revealed that guilt ratings, specifically, correlated significantly with activity in the pACC ($p < .001$, Figure 4), even when changes in all other negative emotions (i.e., anger, anxiety, shame, and embarrassment) were covaried, $r = -.56$,

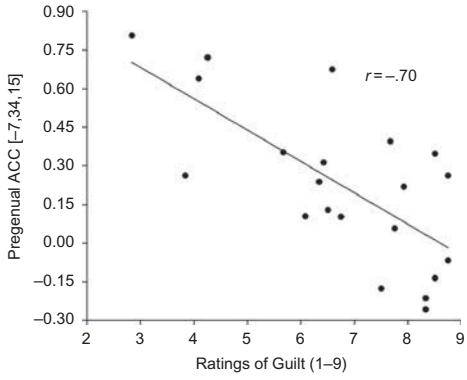

Figure 4. Mean activation level (parameter estimates) observed within the pregenual ACC during the Prejudice condition is negatively correlated with self-reports of guilt obtained during the scan. The line represents the linear best fit; *r* refers to the correlation coefficient. Coordinates refer to the peak activation.

$p < .05$. By contrast, the effects of shame and embarrassment did not remain significant when other emotions were covaried. These analyses suggested that brain activations associated with moral-negative emotion were driven primarily by the experience of guilt. No other regional activations correlated significantly with subjective guilt (or any other emotion).

To probe the motivational significance of the observed neural activation, we assessed the relation between self-reported behavioral motivation (i.e., scores on the BIS/BAS scales) and areas of significant activation during the Prejudice condition. BIS scores correlated positively with activity in the DMPFC/supraACC ($r = .48$, $p < .05$), and correlated negatively with activity in the pACC ($r = -.44$, $p < .05$). In contrast, higher BAS scores were associated with significantly greater activity in the pACC ($r = .68$, $p < .01$).

DISCUSSION

To advance our current understanding of the psychological function of moral-negative emotion, we investigated emotional responses to prejudice feedback using a novel and ecologically valid paradigm. Specifically, we employed a careful screening procedure to identify individuals likely to experience intense guilt, and then had them participate in an IAT-based fMRI paradigm with preprogrammed feedback. Analyses of subjective emotion reports and manipulation check data indicated that while the prejudice manipulation elicited a constellation of closely related moral-negative emotion, guilt was the most intensely felt emotion among these. Furthermore, the effects of this manipulation on guilt remained significant when adjusting for other (basic and moral) negative emotions.

Because we succeeded in measuring moral emotion—primarily guilt—when experienced as a salient and personally relevant affective state, this study makes important theoretical contributions to our understanding of the construct of guilt and the neural substrates underlying its experience. Consistent with our predictions, the obtained data reflected a distributed pattern of neural activation suggesting that guilt is a multifaceted construct, functioning not only to signal punishment and to interrupt ongoing behavior, but also to increase self-reflection and perspective-taking.

Neural correlates of experienced guilt

In this section, we discuss the ways in which the brain activation we observed confirms (or disconfirms) hypotheses about the nature of moral-negative emotion, with a focus on guilt, based on previous research and speculation.

Self-reflection and mentalizing

Psychological models of guilt suggest that the experience of this emotion requires self-reflection and self-evaluation (Leary, 2004). Likewise, in the prejudice literature, guilt following behavior inconsistent with one's internalized nonprejudiced values is believed to be accompanied by increased self-focus or retrospective reflection, which in turn forms part of a self-regulatory cycle to help the individual respond more appropriately in future (Monteith et al., 2002). In the present Prejudice condition we found significant activation in the paracingulate region of the DMPFC (BA 9/32), which may be interpreted, based on previous findings, as reflecting increased self-related processing.

The paracingulate region of the DMPFC has been identified as the primary prefrontal region subserving ToM (Walter et al., 2004), and it also appears to be intricately involved in processing related to the self. In a meta-analysis of self-related neuroimaging studies, Northoff et al. (2006) highlighted dorsal and ventral medial frontal cortex (MFC) as key nodes in a neural network subserving explicit self-association. The PCC and neighboring precuneus area are also important components of this network consisting of cortical midline structures. The PCC, for example, responds when participants engage in self-reflection (Johnson et al., 2002), or when they retrieve judgments related

to the self (Lou et al., 2004). Because of its dense connections with the hippocampus, which is implicated in autobiographical memory recall, Northoff et al. (2006) suggested that the PCC is important for integrating past and present self-referential stimuli in a temporal context. The co-activation of several cortical midline structures (DMPFC, PCC, and precuneus) following the current prejudice manipulation therefore suggests strongly that participants engaged in self-reflection and mentalizing.

Of the cortical midline structures, the paracingulate region of the DMPFC may support a unique social cognitive function. For instance, D'Argembeau et al. (2007) manipulated both self-referential processing and mentalizing and found that the left paracingulate region was a key area where these dimensions interacted. Similarly, Amodio and Frith (2006) suggested that the anterior rostral MFC may not be limited to reflections about our own subjective experience, but may be involved in concerns about our reputation, or about the image of ourselves in the minds of others. Anterior rostral MFC may therefore support reflected appraisals—representations of how we think others evaluate us, either real or imagined—a process considered integral to all self-conscious emotions (Leary, 2007).

Social response reversal

Arguably, the most important self-regulatory purpose of guilt is behavioral change. It functions to correct behavior that is not consistent with moral standards, while punishing misbehaviors (Tangney, Struewig, & Mashek, 2007). Moreover, the negative affective experience of guilt appears to drive reparatory behaviors so as to bring about restitution or appeasement (Kubany & Watson, 2003).

Finger et al. (2006) interpreted enhanced activity in the DMPFC (BA 8/9) and lateral OFC (BA 47) following moral and social transgressions as signaling the need for behavioral change, and, ultimately, the initiation of alternative motor responses. Several studies investigating neural responses to socially unacceptable or embarrassing scenarios have detected co-activation of those areas (e.g., Berthoz, Armony, Blair, & Dolan, 2002; Zahn et al., 2009). Hence, we interpret enhanced activity in the DMPFC and lateral OFC during our prejudice manipulation as related to the processing of specific social, contextual, and emotional cues to modify current behavior.

Amodio and Frith (2006) proposed that the posterior rostral MFC serves a self-regulatory function by monitoring the outcome of current actions while continually updating representations of, and evaluating the merit of, future actions. One could thus interpret significant DMPFC activation detected during our Prejudice condition as supporting enhanced cognitive processing during social transgressions, which appears to be important for initiating alternative behaviors and for evaluating possible future actions.

Lateral OFC, in turn, appears to be sensitive to cues of punishment and negative emotional reactions from others. These cues signal that our current behavior is socially unacceptable and should be curtailed or changed (Kringelbach & Rolls, 2004). The distinct connectivity and function of lateral OFC in diverse contexts lends further support to the importance of this area in facilitating appropriate social conduct. Notably, lateral compared to medial OFC receives more multimodal sensory-related inputs (Carmichael & Price, 1996), and appears to be more involved in changing responses under unexpected circumstances (Elliott, Dolan, & Frith, 2000). Wagner et al. (2011) demonstrated the involvement of the right lateral OFC specifically in relation to guilt, arguing that lateral OFC encodes negative affect particularly related to social contexts.

Behavioral inhibition

A current debate in psychology concerns guilt's behavioral implications: does it inhibit transgressive behavior (i.e., BIS) or promote prosocial behavior (i.e., BAS) (Amodio et al., 2007; Janoff-Bulman, Sheikh, & Hepp, 2009; Monteith et al., 2002)? The present study contributes, for the first time from a neuroimaging perspective, to that debate.

The DMPFC (BA 8/9) activation in the present Prejudice condition also extended to the supraACC (BA 24/32), an area consistently implicated in monitoring response tendencies for competition and in signaling the need for enhanced top-down control in conflict situations (Botvinick, Cohen, & Carter, 2004). In moral judgment tasks, for example, enhanced supraACC activity is associated with increased decision difficulty and with longer reaction times in response to complex moral dilemmas (Greene, Nystrom, Engell, Darley, & Cohen, 2004). Amodio et al. (2004) demonstrated that these conflict-detection processes are also sensitive to the automatic, but undesired, activation of racial stereotypes in low-prejudice individuals. In the context of our current prejudice paradigm, it is likely that participants experienced internal conflict (i.e., between their internalized moral standards and the IAT feedback suggesting they are prejudiced).

Of particular importance here, however, is the fact that conflict monitoring via the supraACC is

associated with interruption of action (van Veen & Carter, 2002; van Veen, Cohen, Botvinick, Stenger, & Carter, 2001). Likewise, the significant supraACC activity detected during our Prejudice condition could signify that guilt functions to interrupt ongoing behavior, before alternative motor responses are initiated.

Further support for guilt's immediate association with behavioral inhibition, rather than behavioral activation, stems from studies examining the neurocognitive correlates of the Behavioral Inhibition and Behavioral Activation Systems (BIS/BAS). Insofar as BIS sensitivity has been shown to correlate with conflict-related activity in the supraACC (Amodio et al., 2008; Beaver, Lawrence, Passamonti, & Calder, 2008), and BAS sensitivity is organized around the dopaminergic system, including the ventral striatum and ventromedial PFC (Beaver et al., 2008), the prominent supraACC (BA 24/32) activity detected in the current Prejudice condition points to acute guilt's association with behavioral inhibition. The positive association between participants' BIS scores and activity in the supraACC further supports this conclusion.

Taken together, our results lend support to the idea that the events triggered by guilt constitute an inhibitory, self-regulatory system, whereby self-focus is increased and reinforcement learning is promoted in order to respond more appropriately in future (Monteith et al., 2002). Guilt could thus be viewed as a form of automatic behavioral control to inhibit social transgressions.

Physiological arousal and affective feeling

We anticipated observing significant physiological arousal during the present prejudice manipulation, given that guilt is typically experienced as a strong aversive emotion. Because the ACC, anterior insula, and thalamus have been linked to autonomic output (Critchley, 2005), increased activation in these areas during the present Prejudice condition confirmed that participants experienced heightened physiological arousal during guilt.

Although various imaging studies on guilt have reported activation in the anterior insula, particularly on the left (e.g., Basile et al., 2011; Shin et al., 2000), this area is not necessarily specific to guilt. Rather, the anterior insula is conceived of as a unique cortical substrate specialized for evaluating internal body states, thereby instantiating *all* subjective feeling (Craig, 2009). In particular, the degree of anterior insula activation appears to be related to the degree of personal association with, as well as the perceived unpleasantness of, emotional stimuli (Akitsuki & Decety, 2009).

Affect regulation

An interesting finding of the present study was the significant pACC (BA 32) activity during the prejudice manipulation and, in particular, the inverse relation of activity in this area with self-reported guilt. Participants who recruited the pACC more effectively thus appeared to appraise the negative, guilt-inducing feedback as less salient. We believe that differences in pACC activity may point to individual differences in recruiting self-regulatory processes in response to the negative affect of guilt. Activity in this region may therefore not necessarily correspond to the set of neural correlates that underlie guilt across individuals.

The pACC forms part of rostral–ventral ACC, which has traditionally been associated with some form of emotion processing (Bush, Luu, & Posner, 2000). Increasing evidence points to the pACC's more specific involvement in emotion inhibition (regulation), however. For example, increased pACC activity has been detected when participants are instructed to inhibit affective responses to negative emotional stimuli (Ochsner et al., 2004; Shafritz, Collins, & Blumberg, 2006), or in the absence of explicit instruction to regulate emotion (Etkin, Egner, Peraza, Kandel, & Hirsch, 2006). Furthermore, rostral ACC is known to be activated by both placebo and opioid analgesia (Petrovic, Kalso, Petersson, & Ingvar, 2002).

A recent review proposes that the functional role of the ventral–rostral ACC involves regulation of affective processing through the suppression of limbic emotion regions, e.g., the amygdala (Etkin, Egner, & Kalisch, 2011). Specifically, the authors propose that ventral–rostral ACC might perform a generic negative emotion inhibitory function that can be recruited by other (PFC) regions when the need to suppress limbic reactivity arises. Interestingly, increased pACC activity in our sample was also associated with higher BAS scores. In this regard, previous research suggests that high BAS-sensitive individuals may be insensitive to cues of punishment, focusing their attention instead on cues of incentive (Patterson & Newman, 1993).

LIMITATIONS AND CONCLUSION

One limitation of this study is that our experimental paradigm was not successful in discriminating between responses to the Egalitarian and Neutral conditions. Although subjective reports were indicative of

significant pride and satisfaction following the egalitarian feedback, we did not detect any significant pride-specific neural activation. It is thus possible that the emotion elicited by the Egalitarian condition was merely cognitively pleasing. It should be noted, however, that our paradigm was designed primarily to elicit moral-negative affect (especially guilt); the Egalitarian condition served as another control condition to bolster the authenticity of the IAT feedback, but was not of theoretical interest to us.

A second limitation concerns the generalizability of the current data. The effectiveness of our paradigm relied heavily on participants fulfilling a set of carefully stipulated eligibility criteria (e.g., White, heterosexual, non-Jewish, low-prejudiced); we employed this selective sampling method to prioritize construct validity over external validity, given the challenges involved in experimentally eliciting strong guilt in past research. Furthermore, the stringent criteria used in our study corresponded to specific features of the experimental design and fMRI recoding environment —a necessity for conducting a well-controlled experiment (Amodio, 2010). The obvious limitation to this approach is that it may constrain the generalizability of our findings. That is to say, the psychological and neural mechanisms underlying guilt may be different in other samples of healthy individuals. Although such differences remain a possibility, we do not know of any evidence suggesting different mechanisms of guilt among different populations. It is important to acknowledge, however, that strong guilt may also be elicited without employing stringent selection criteria in other contexts (Wagner et al., 2011).

A third limitation, related to the one described above, is that we investigated guilt in a female-only sample. We did so to avoid confounds due to possible sex-by-emotion effects. More research is necessary to tease apart sex effects for emotion at the neural level, although at present it appears plausible to assume that any sex differences in the neural substrates of guilt are likely to be subtle (Wager, Phan, Liberzon, & Taylor, 2003).

Finally, although guilt was the emotion felt most strongly, our prejudice manipulation also evoked other negative emotions that may have impacted on neural activation responses associated with guilt. Moral emotions are complex and multifaceted, however, and we understand that it is not feasible to extract "pure guilt." Rather, the accompanying negative emotions may be viewed as part of the more general emotional profile of guilt, and may change from one situation to the next (Izard, 1991). For example, the emotional overlay of the present Prejudice condition may have been characterized by feelings of embarrassment and shame because participants were aware that their IAT performances were being monitored (Finger et al., 2006). Nevertheless, our statistical analyses were able to isolate the effects of guilt in order to show that the observed patterns of neural activity were specific to the experience of that moral emotion, as opposed to more general negative affect.

In conclusion, the present study contributed methodologically as well as theoretically to our understanding of moral-negative emotion. The results of our novel fMRI prejudice manipulation support the idea that guilt functions as a multifaceted construct consisting of several distinct sub-processes that together may serve to guide and direct moral behavior in complex ways. Because guilt is also associated with prosocial behavior, which was not investigated in the present study, it remains to be seen whether any of the distinct sub-processes associated with guilt are sufficient to motivate such behavior. We contend that such investigations are perhaps best carried out using ecologically valid paradigms that elicit social–moral emotions in real time.

Supplementary material

Supplemental data for this article can be accessed here [http://dx.doi.org/10.1080/17470919.2013.878750].

Original manuscript received 29 May 2013
Revised manuscript accepted 19 December 2013
First published online 22 January 2014

REFERENCES

Akitsuki, Y., & Decety, J. (2009). Social context and perceived agency affects empathy for pain: An event-related fMRI investigation. *NeuroImage, 47*, 722–734.

Amodio, D. M. (2010). Can neuroscience advance social psychological theory? Social neuroscience for the behavioral social psychologist. *Social Cognition, 28*, 695–716.

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2007). A dynamic model of guilt. Implications for motivation and self-regulation in the context of prejudice. *Psychological Science, 18*, 524–530.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*, 268–277.

Amodio, D. M., Harmon-Jones, E., & Devine, P. G. (2003). Individual differences in the activation and control of affective race bias as assessed by startle eyeblink response and self-report. *Journal of Personality and Social Psychology, 84*, 738–753.

Amodio, D. M., Harmon-Jones, E., Devine, P. G., Curtin, J. J., Hartley, S. L., & Covert, A. E. (2004). Neural signals for the detection of unintentional race bias. *Psychological Science, 15*, 88–93.

Amodio, D. M., Master, S. L., Yee, C. M., & Taylor, S. E. (2008). Neurocognitive components of the behavioral inhibition and activation systems: Implications for theories of self-regulation. *Psychophysiology, 45*, 11–19.

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*, 229–239.

Beaver, J. D., Lawrence, A. D., Passamonti, L., & Calder, A. J. (2008). Appetitive motivation predicts the neural response to facial signals of aggression. *The Journal of Neuroscience, 28*, 2719–2725.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the beck depression inventory-II*. San Antonio, TX: Psychological Corporation.

Beer, J. S., John, O. P., Scabini, D., & Knight, R. T. (2006). Orbitofrontal cortex and social behavior: Integrating self-monitoring and emotion-cognition interactions. *Journal of Cognitive Neuroscience, 18*, 871–879.

Berthoz, S., Armony, J. L., Blair, R. J. R., & Dolan, R. J. (2002). An fMRI study of intentional and unintentional (embarrassing) violations of social norms. *Brain, 125*, 1696–1708.

Blair, R. J., & Cipolotti, L. (2000). Impaired social response reversal. A case of "acquired sociopathy". *Brain, 123*, 1122–1141.

Botvinick, M., Cohen, J. D., & Carter, C. S. (2004). Conflict-monitoring and anterior cingulate cingulate cortex: An update. *Trends in Cognitive Sciences, 8*, 539–546.

Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences, 4*, 215–222.

Carmichael, S. T., & Price, J. L. (1996). Connectional networks within the orbital and medial prefrontal cortex of macaque monkeys. *Journal of Comparative Neurology, 371*, 179–207.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology, 67*, 319–333.

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

Critchley, H. D. (2005). Neural mechanisms of autonomic, affective, and cognitive integration. *The Journal of Comparative Neurology, 493*, 154–166.

D'Argembeau, A., Ruby, P., Collette, F., Degueldre, C., Balteau, E., Luxen, A., … Salmon, E. (2007). Distinct regions of the medial prefrontal cortex are associated with self-referential processing and perspective taking. *Journal of Cognitive Neuroscience, 19*, 935–944.

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

Devine, P. G., Monteith, M. J., Zuwerink, J. R., & Elliot, A. J. (1991). Prejudice with and without compunction. *Journal of Personality and Social Psychology, 60*, 817–830.

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology, 82*, 835–848.

Drabant, E. M., McRae, K., Manuck, S. B., Hariri, A. R., & Gross, J. J. (2009). Individual differences in typical reappraisal use predict amygdala and prefrontal responses. *Biological Psychiatry, 65*, 367–373.

Elliott, R., Dolan, R. J., & Frith, C. D. (2000). Dissociable functions in the medial and lateral orbitofrontal cortex: Evidence from human neuroimaging studies. *Cerebral Cortex, 10*, 308–317.

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

Etkin, A., Egner, T., Peraza, D. M., Kandel, E. R., & Hirsch, J. (2006). Resolving emotional conflict: A role for the rostral anterior cingulate cortex in modulating activity in the amygdala. *Neuron, 51*, 871–882.

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

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., & Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. *Magnetic Resonance in Medicine, 33*, 636–647.

Fourie, M. M., Kilchenmann, N., Malcolm-Smith, S., & Thomas, K. G. F. (2012). Real-time elicitation of moral emotions using a prejudice paradigm. *Frontiers in Emotion Science, 3*, 1–14.

Fourie, M. M., Rauch, H. G. L., Morgan, B. E., Ellis, G. F. R., Jordaan, E. R., & Thomas, K. G. F. (2011). Guilt and pride are heartfelt, but not equally so. *Psychophysiology, 48*, 888–899.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*, 389–400.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464–1480.

Harmon-Jones, E., Amodio, D. M., & Zinner, L. R. (2007). Social psychological methods in emotion elicitation. In J. A. Coan & J. J. B. Allen (Eds.), *Handbook of emotion elicitation and assessment* (pp. 91–105). New York, NY: Oxford University Press.

Herek, G. M. (2000). Sexual prejudice and gender: Do heterosexuals' attitudes toward lesbians and gay men differ? *Journal of Social Issues, 56*, 251–266.

Herrald, M. M., & Tomaka, J. (2002). Patterns of emotion-specific appraisal, coping, and cardiovascular reactivity during an ongoing emotional episode. *Journal of Personality and Social Psychology, 83*, 434–450.

Izard, C. E. (1991). *The psychology of emotions*. New York, NY: Plenum Press.

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology, 96*, 521–537.

Johnson, S. C., Baxter, L. C., Wilder, L. S., Pipe, J. G., Heiserman, J. E., & Prigatano, G. P. (2002). Neural correlates of self-reflection. *Brain, 125*, 1808–1814.

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*, 1788–1798.

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. *The Journal of Neuroscience, 29*, 2188–2192.

Kringelbach, M. L., & Rolls, E. T. (2004). The functional neuroanatomy of the human orbitofrontal cortex: Evidence from neuroimaging and neuropsychology. *Progress in Neurobiology, 72*, 341–372.

Kubany, E. S., & Watson, S. B. (2003). Guilt: Elaboration of a multidimensional model. *The Psychological Record, 53*, 51–90.

Leary, M. R. (2004). Digging deeper: The fundamental nature of "self-conscious" emotions. *Psychological Inquiry, 15*, 129–131.

Leary, M. R. (2007). Motivational and emotional aspects of the self. *Annual Review of Psychology, 58*, 22.01–22.28.

Lou, H. C., Luber, B., Crupain, M., Keenan, J. P., Nowak, M., Kjaer, T. W., … Lisanby, S. H. (2004). Parietal cortex and representation of the mental self. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 6827–6832.

Manstead, A. S. R. (1992). Gender differences in emotion. In A. Gale & M. W. Eysenck (Eds.), *Handbook of individual differences: Biological perspectives*. Chichester: Wiley.

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

Moll, J., De Oliveira-Souza, R., & Zahn, R. (2008). The neural basis of moral cognition. *Annals of the New York Academy of Sciences, 1124*, 161–180.

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). Opinion: The neural basis of human moral cognition. *Nature Reviews Neuroscience, 6*, 799–809.

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice-reduction efforts. *Journal of Personality and Social Psychology, 65*, 469–485.

Monteith, M. J., Ashburn-Nardo, L., Voils, C. I., & Czopp, A. M. (2002). Putting the brakes on prejudice: On the development and operation of cues for control. *Journal of Personality and Social Psychology, 83*, 1029–1050.

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*, 683–692.

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

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., … Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology, 18*, 36–53.

Ochsner, K. N., Ray, R. D., Cooper, J. C., Robertson, E. R., Chopra, S., Gabrieli, J. D., & Gross, J. J. (2004). For better or for worse: Neural systems supporting the cognitive down- and up-regulation of negative emotion. *NeuroImage, 23*, 483–499.

Patterson, C. M., & Newman, J. P. (1993). Reflectivity and learning from aversive events: Toward a psychological mechanism for the syndromes of disinhibition. *Psychological Review, 100*, 716–736.

Petrovic, P., Kalso, E., Petersson, K. M., & Ingvar, M. (2002). Placebo and opioid analgesia —Imaging a shared neuronal network. *Science, 295*, 1737–1740.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology, 75*, 811–832.

Shafritz, K. M., Collins, S. H., & Blumberg, H. P. (2006). The interaction of emotional and cognitive neural systems in emotionally guided response inhibition. *NeuroImage, 31*, 468–475.

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

Tangney, J. P., Struewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology, 58*, 345–372.

Tracy, J. L., & Robins, R. W. (2004). Putting the self into self-conscious emotions: A theoretical model. *Psychological Inquiry, 15*, 103–125.

van Veen, V., & Carter, C. S. (2002). The timing of action-monitoring processes in the anterior cingulate cortex. *Journal of Cognitive Neuroscience, 14*, 593.

van Veen, V., Cohen, J. D., Botvinick, M. M., Stenger, V. A., & Carter, C. S. (2001). Anterior cingulate cortex, conflict monitoring, and levels of processing. *NeuroImage, 14*, 1302–1308.

Vrana, S. R., Cuthbert, B. N., & Lang, P. J. (1989). Processing fearful and neutral sentences: Memory and heart rate change. *Cognition & Emotion, 3*, 179–195.

Wager, T. D., Phan, K. L., Liberzon, I., & Taylor, S. F. (2003). Valence, gender, and lateralization of functional brain anatomy in emotion: A meta-analysis of findings from neuroimaging. *NeuroImage, 19*, 513–531.

Wagner, U., N'Diaye, K., Ethofer, T., & Vuilleumier, P. (2011). Guilt-specific processing in the prefrontal cortex. *Cerebral Cortex, 21*, 2461–2470.

Walter, H., Adenzato, M., Ciaramidaro, A., Enrici, I., Pia, L., & Bara, B. G. (2004). Understanding intentions in social interaction: The role of the anterior paracingulate cortex. *Journal of Cognitive Neuroscience, 16*, 1854–1863.

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*, 276–283.