Neural Correlates of Automatic Beliefs About Gender and Race

Kristine M. Knutson, Linda Mah, Charlotte F. Manly, and Jordan Grafman*

*Cognitive Neuroscience Section, National Institute of Neurological Disorders and Stroke,
National Institutes of Health, Bethesda, Maryland*

Abstract: Functional MRI was used to identify the brain areas underlying automatic beliefs about gender and race, and suppression of those attitudes. Participants ($n = 20$; 7 females) were scanned at 3 tesla while performing the Implicit Association Test (IAT), an indirect measure of race and gender bias. We hypothesized that ventromedial prefrontal cortex areas (PFC) would mediate gender and racial stereotypic attitudes, and suppression of these beliefs would recruit dorsolateral prefrontal cortex (DLPFC) and the anterior cingulate cortex (ACC). Performance data on the IAT revealed gender and racial biases. Racial bias was correlated with an explicit measure of racism. Results showed activation of anteromedial PFC and rostral ACC while participants implicitly made associations consistent with gender and racial biases. In contrast, associations incongruent with stereotypes recruited DLPFC. Implicit gender bias was correlated with amygdala activation during stereotypic conditions. Results suggest there are dissociable roles for anteromedial and dorsolateral PFC circuits in the activation and inhibition of stereotypic attitudes. *Hum Brain Mapp* 28:915–930, 2007. © 2006 Wiley-Liss, Inc.

Key words: social cognition; emotion; behavior; attitudes; automatic processing; stereotypes; orbitofrontal cortex; prefrontal cortex; gender differences

INTRODUCTION

Clinical and experimental studies of patients with prefrontal cortex (PFC) lesions implicate a role for this region in social cognition and behavior, specifically, the orbitofrontal cortex (OFC) [Bechara et al., 1994; Damasio et al., 1994; Harlow, 1848; Hornak et al., 1996; Mah et al., 2005; Milne and Grafman, 2001; Rolls et al., 1994; Stone et al., 1998]. Recent functional imaging studies of healthy volunteers using social cognitive paradigms corroborate these clinical data [e.g., Mitchell et al., 2002a; Moll et al., 2002]. In addition, the amygdala has been implicated in stereotyping people of other races [Hart et al., 2000; Phelps et al., 2000; Wheeler and Fiske, 2005]. In this study, we directly examined the neural correlates of gender and racial stereotypic attitudes using the Implicit Association Test (IAT) performed during functional MRI [Greenwald et al., 1998].

The IAT implicitly measures social attitudes by examining the differential association of two object categories

A portion of this work has been presented at the Cognitive Neuroscience Society Meeting, New York, NY, March 29th–April 1st, 2003, and at the American Neuropsychiatric Association Meeting, Bal Harbour, Florida, February 21–24, 2004.

Lindah Mah is currently at Mood and Anxiety Disorders Program, Molecular Imaging Branch, National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland 20892-0135.

Contract grant sponsor: National Institute of Neurological Disorders and Stroke, NIH.

*Correspondence to: Jordan Grafman, Ph.D., Cognitive Neuroscience Section, NINDS/NIH, Building 10, Room 5C205, 10 Center Drive, MSC 1440, Bethesda, Maryland 20892-1440.

E-mail: grafmanj@ninds.nih.gov

Received for publication 31 January 2006; Revision 12 May 2006; Accepted 26 June 2006

DOI: 10.1002/hbm.20320

Published online 28 November 2006 in Wiley InterScience (www.interscience.wiley.com).

(e.g., male versus female) with two attribute categories (e.g., strong versus weak). Using the example of an IAT to examine gender biases, participants classify names as to their gender (e.g., “Robert” is a male name; “Mary” is a female name), and categorize words based on an attribute; in this case, whether the word has a connotation of strength or weakness (e.g., “dominant” is a strong word; “fragile” is a weak word). Name and word trials are then combined into blocks representing either the “congruent” or “incongruent” condition, based on stimulus-response mapping. In the congruent condition, participants are instructed to map stereotypically compatible concepts onto a single response key (e.g., press the left key for a female name or a weak word), while in the incongruent condition, they map incompatible concepts onto the same key (e.g., press the right key for a male name or a weak word). Because performance is typically slower in the incongruent condition, the difference in response latency provides a measure of attitudinal bias (“IAT effect”). This effect is robust across a broad range of stereotypic beliefs, and predicts actual behavior [Greenwald and Nosek, 2001]. For example, an early IAT study used Koreans and Japanese names as concepts, and pleasant and unpleasant words as attitudes. For those participants of Korean ethnicity, response times were faster when Korean (as opposed to Japanese) names and pleasant words were associated with the same response key; for those of Japanese ethnicity, response times were faster when Japanese names and pleasant words were paired with the same response key [Greenwald et al., 1998]. Greenwald et al. explained that ethnically Korean participants found it more difficult to perform the Japanese + pleasant than the Korean + pleasant combination and vice-versa due to ethnic bias. This bias was stronger the more the participants were socially involved with their own ethnic group.

Because it often does not correlate with explicit self-report of attitudes, the IAT has been interpreted as a measure of implicit or unconscious beliefs or attitudes [Baron and Banaji, 2006; Greenwald et al., 1998; Hofmann et al., 2005]. Studies using the IAT also support the notion that the activation and use of stereotypes is an automatic process that is, however, context and exposure sensitive [Dasgupta and Greenwald, 2001; Mitchell et al., 2003b; Payne, 2005; Rudman et al., 2001a] and moderated by executive control [Payne, 2005].

A previous study in our laboratory showed that patients with lesions in the ventromedial PFC (VMPFC) exhibit a diminished IAT effect for stereotypic associations about gender, as compared with healthy volunteers or patients with dorsolateral prefrontal cortex (DLPFC) damage [Milne and Grafman, 2001]. One interpretation of this finding is that VMPFC lesions in humans result in degraded access to previously acquired gender attitudes, and that learned associations about our social world are represented within the VMPFC. We followed up these clinical findings using functional magnetic resonance imaging (fMRI) of 20 healthy volunteers while they performed two IAT tasks evaluating stereotypes about gender and race.

For this experiment, blocks of the congruent condition alternated with blocks of the incongruent condition. In addition, a lower-level control task was included in which participants classified the stimuli as being either names or words, independent of their attribute (e.g., whether words were strong or weak). Thus, the IAT tasks consisted of three conditions: congruent (CONG; compatible associations mapped onto the same response key; e.g., male names, strong words), incongruent (INCONG; incompatible associations mapped onto the same key; e.g., female names, strong words), and a baseline task (BASE; classify stimuli as names or words). The gender IAT included 80 each of (1) male names, (2) female names, (3) strong words, and (4) weak words. The racial IAT included 80 each of (1) names considered typical of Caucasian-American males (“white” names), (2) names considered typical of African-American males (“black” names), (3) pleasant words, and (4) unpleasant words. Following the scanning session, participants completed a series of questionnaires to obtain explicit self-report measures of attitudes about gender and race. Reaction times, D scores, and error data for the gender and racial IATs were analyzed separately, as well as together to form scores for an “overall IAT”. The D score divides the difference between the INCONG and CONG response times by the standard deviation of the individual’s response times. This removes the effect of the individual’s latency variability from the measure [Greenwald et al., 2003]. The data obtained during the runs were analyzed with SPM2 using random-effect analyses ($n = 20$ participants, $P < 0.02$ uncorrected with $k = 20$).

Previous functional imaging and lesion data guided our hypotheses regarding the role of specific brain areas in mediating these automatic beliefs. The VMPFC/medial OFC (M-OFC) region is implicated in the execution of over-learned, automatic processes [Koechlin et al., 2002], including processing of emotional and social information [Mah et al., 2005; Mitchell et al., 2002b; Raichle et al., 1994]. We predicted that the VMPFC would be activated while participants made associations consistent with stereotypic attitudes about race and gender. In contrast, we hypothesized that structures involved in cognitive and inhibitory control, error detection, and response reversal would be recruited while participants suppressed these well-learned associations. These regions included the lateral OFC, anterior cingulate cortex (ACC), and DLPFC [Carter et al., 1998; Chee et al., 2000; MacDonald et al., 2000; Rolls, 1996, 2004].

MATERIALS AND METHODS

Participants

Participants were right-handed, native English speakers, aged 19–35 years old, with no history of neuropsychiatric abnormalities and no neurological deficits on examination. Prior to participation in the study, all participants gave informed consent to a protocol that had been approved by

the Institutional Review Board. The experiment conformed to the Code of Ethics of the World Medical Association (Declaration of Helsinki) and all NIH regulatory standards. All participants had normal or corrected-to-normal visual acuity. A total of 36 participants were recruited into the study. Ten participants were excluded because their error rate on the Implicit Association Test (IAT) tasks was unacceptable ($>10\%$). Another six were excluded due to excessive head motion (greater than 3 mm translation or 3° rotation in any plane). For clarity, we present behavioral and imaging data for the remaining 20 participants only. This sample included 13 males and 7 females, all of whom were Caucasian except for one Asian-American female. The racial and gender makeup of our participants was relevant to our results. Participants had an average of 17.75 years of education (SD 4.81), and a mean age of 31.45 years (SD 6.75).

Behavioral Procedures

IAT task

The IAT was adapted for functional magnetic resonance imaging (fMRI) using a block design. Stimuli and trials were added to increase statistical power. All participants performed a practice task as well as classification of names and words immediately prior to performing the IAT in the scanner. For the gender IAT, first participants classified names as representing either males or females. They then classified words as either being strong or weak. Similarly, for the racial IAT, participants classified names as being either typically white or black names, and then words as being pleasant or unpleasant. Participants were then presented with a practice task similar to the task given in the scanner (see below), with a different set of stimuli from the same sources (see the Stimuli section below for details).

The task design is depicted in Figure 1. Each trial consisted of a stimulus presented for 1,400 ms followed by a blank screen for 100 ms, for an intertrial interval of 1,500 ms. Runs included four 30-s blocks of each condition, with 20 trials per block. Conditions were congruent (CONG), incongruent (INCONG), and baseline (BASE). The order of conditions was counterbalanced within participants. Participants were presented with alternating runs of gender and racial IATs, and completed three runs of each. Half of the participants began with a gender IAT run (order A), while the other half began with a racial IAT run (order B), to counterbalance order effects. The total duration of the task was approximately 45 min. Participants were instructed to respond as quickly and as accurately as possible.

Stimuli

Stimuli were names and words that were visually presented using SuperLab Pro 1.74 (www.cedrus.com/) and a Macintosh G4 computer. Name and word lists for the gender and racial IAT were mutually exclusive. The gender IAT included 80 each of (1) male names, (2) female names,

(3) strong words, and (4) weak words. Stimuli were selected from a previous gender IAT study [Rudman et al., 2001b] as well as from other sources. Additional names were obtained from U.S. census data (www.census.gov/genealogy/names) and supplementary strong and weak words were compiled using internet searches on thesaurus and dictionary websites. Male and female names were matched for number of letters and frequency in the U.S. population. Word stimuli were generated using an initial pool of 264 words that were classified by healthy participants ($n = 32$) in a pilot study as being either strong, weak, or neither. Words with at least 70% agreement were included in the final selection. Strong and weak words were matched for length, frequency, imageability, and concreteness using norms from the MRC linguistic database [Coltheart, 1981; www.psy.uwa.edu.au/MRCDatabase/uwa_mrc.htm] and imageability norms [Bird et al., 2001] for words not located in the MRC linguistic database.

The racial IAT included 80 each of (1) names considered typical of Caucasian-American males ("white" names, e.g., "Vince", "Bobbie"), (2) names considered typical of African-American males ("black" names, e.g., "Rashad", "Tyrell"), (3) pleasant words, and (4) unpleasant words. White and black names were chosen from those used in previous racial IAT studies [Dasgupta and Greenwald, 2001; Greenwald et al., 1998], and other sources included internet lists and books of African-American baby names. Participants rated an initial pool of 385 names as being typically associated with Caucasian- or African-American males in a pilot study. Names with at least 70% agreement were included. Black and white names were matched for number of letters; however, white names were represented more frequently in the U.S. population according to census data. Since all stimuli were presented across each condition in the racial IAT, this difference was unlikely to confound the interpretation of results. Pleasant and unpleasant words for the racial IAT were derived from a previous study [Greenwald et al., 1998] and from a normed list of words [Bellezza et al., 1986]. The initial pool of 270 words was rated as being either pleasant, unpleasant or neither by healthy volunteers in a pilot study. Pleasant and unpleasant words were matched for word length, frequency, imageability, and concreteness using the same norms as with the gender IAT.

Explicit measures

Following previous gender IAT studies [Milne and Grafman, 2001; Rudman et al., 2001b], participants were administered the Attitudes Toward Women Scale [AWS; Spence and Helmreich, 1972], the Ambivalent Sexism Inventory [ASI; Glick and Fiske, 1996], and a gender potency semantic differential scale (GSD). The AWS consists of items that assess traditional gender-role beliefs (e.g., "Swearing and obscenity are more repulsive in the speech of a woman than a man"). The ASI assesses both "benevolent" sexism ("Women should be cherished and protected

Figure 1.

Implicit association test. Example timeline of tasks in the gender IAT and racial IAT. Stimulus categories (gender, race, attributes) on the same side of the screen (left or right) are mapped onto the corresponding response key. Stimuli are either male or females names (gender IAT) or black or white male names (racial

IAT) alternating with pleasant or unpleasant words. Congruent and incongruent conditions differ only by reversal of stimulus-response contingencies. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

by men") and "hostile" sexism ("Women seek to gain power by getting control over men"). Finally, the gender semantic differential scale consisted of a series of bipolar adjectives (e.g., dominant–meek), which were derived from the word stimuli used for the gender IAT in this study. Participants were instructed to indicate where on the scale their conceptualization of men and women (as independent constructs) would lie.

Following previous racial IAT studies [Greenwald et al., 1998], participants were also administered the Modern Racism Scale [MRS; McConahay et al., 1981], and the Diversity and Discrimination Scales (DIS, DIV) [Wittenbrink et al., 1997]. These questionnaires include statements such as "It is easy to understand the anger of black people in America" (MRS), and "There is a real danger that too much emphasis on cultural diversity will tear the United States apart" (DIV). In addition, they completed the Should–Would Discrepancy Questionnaire (SWDQ) [Monteith and Voils, 1998], and a racial semantic differential scale (RSD). The SWDQ was designed to distinguish between beliefs about Blacks (e.g., "I should not feel uncomfortable in the company of Black people"), and self-report of actual behavior towards Blacks (e.g., "I would feel uncomfortable if I was assigned a Black roommate"). The RSD consisted of a series of bipolar adjectives (e.g.,

beautiful–ugly), which were derived from the word stimuli used for the racial IAT in this study. Similar to the GSD, participants were instructed to indicate where on the scale they perceived black and white individuals.

fMRI procedures

Imaging data were collected on a 3.0 tesla GE scanner using a standard GE head coil. Functional MR images consisted of a set of 22 interleaved T2*-weighted gradient-echo axial slices acquired using an echo-planar imaging (EPI) sequence (TE = 40 ms, TR = 3,000 ms, flip angle = 90, FOV = 24) encompassing the cerebrum and most of the cerebellum for all participants. Voxel resolution was $3.75 \times 3.75 \times 6 \text{ mm}^3$. The first 4 volumes per run were collected during blank stimulus presentation and were discarded to remove T1 saturation effects. A high-resolution ($0.9375 \times 0.9375 \times 1.5 \text{ mm}^3$) 124-slice T1-weighted fast spoiled-grass (SPGR) anatomical image was acquired prior to collecting functional MR data.

Stimuli were back-projected onto a screen at the foot of the scanner gantry, which the participant viewed using a mirrored prism mounted on the head coil above the eyes. A response box in the right hand allowed the participant to respond by pressing the left or right button. Head

motion was restricted using foam pads placed around the participant's head.

Statistical Analysis

Performance data

Both response latencies and error rates were recorded. Data for the gender and racial IATs were analyzed separately, as well as together to form scores for an "overall IAT". Median response latencies and mean error rates for each condition were calculated. Paired *t* tests were used to determine the presence of a significant IAT effect (increased response latency for the incongruent condition relative to the congruent condition).

D scores were computed using the improved algorithm [Greenwald et al., 2003], and mean D scores were analyzed for all tasks for racial and gender IAT data separately, and for the first run of racial and gender IAT data separately. One-sample *t* tests were used to determine the presence of a significant IAT effect of D scores.

In addition, to evaluate the possibility of counterbalancing order effects, a repeated measures analysis of variance with one between-group factor (counterbalancing order) and three within-subjects factors (IAT type, run, condition) was performed. The gender IAT data was also analyzed using a two-sample *t* test to determine whether there were gender differences in performance. Finally, Pearson's rho was used to determine the strength of association between explicit measures (gender and racial attitude questionnaires) and the IAT effect as measured by reaction time differences and D scores for both gender and racial tasks.

Imaging data

Images were preprocessed and analyzed in SPM2 (Wellcome Department of Cognitive Neurology) implemented in MATLAB (Mathworks, Natick, MA). Functional volumes were motion-corrected, realigned using a rigid-body model with 4th degree B-spline interpolation and unwrapped, and the anatomical volume was coregistered to the mean functional image [Friston et al., 1995a]. Volumes were then normalized to Talairach space based on the Montreal Neurological Institute (MNI) template, using an affine model plus moderate regularization with a $7 \times 8 \times 7$ set of nonlinear basis functions [Ashburner and Friston, 1999; Friston et al., 1995a]. Functional volumes were written as $4 \times 4 \times 4$ mm³ voxels. The anatomical volume was written as $2 \times 2 \times 2$ mm³ voxels. Functional images were then smoothed using an 8-mm Gaussian kernel at full width, half maximum [Friston et al., 1995b].

Analysis was carried out using a general linear model and boxcar waveform convolved with a hemodynamic response function. A high pass filter was applied to remove low frequency drifts in signal changes due to physiological noise. Data were globally scaled at the individual participant level of analysis to allow comparison of images from different individuals at the group level of

analysis. Specific effects for each voxel were tested by applying linear contrasts, representing comparisons of interest, to the parameter estimates for each condition, so that one contrast image for each participant was created. A second-level random-effects analysis was carried out by entering the appropriate contrast image from each participant into a one-sample *t* test performed across all participants. Random-effects analyses treat participants as a random effect, allowing inferences to be made regarding the general population, rather than the particular study sample [Friston et al., 1999].

Imaging data were analyzed using whole brain voxel-wise comparisons. Main effects of condition (CONG, INCONG, BASE) were determined by collapsing the imaging data across IAT type (i.e., combining gender and racial IAT data). The following contrasts were performed: (1) CONG > INCONG, (2) INCONG > CONG, (3) CONG > BASE, and (4) INCONG > BASE. The two IAT types (gender and race data) were also analyzed separately.

The relationship between the IAT response time bias and activation of brain regions in the contrast CONG > INCONG was examined for gender and racial IAT data separately and in combination. For each individual participant, the behavioral measure was regressed onto the magnitude of activation in the corresponding CONG > INCONG contrast image. The IAT effect was expected to correlate with activation of hypothesized regions (ventromedial prefrontal cortex areas (PFC)/medial OFC) in the CONG > INCONG contrast. Additionally, a regression analysis was performed using the D scores in place of RT differences.

Brodmann's areas were determined using the Talairach Daemon (ric.uthscsa.edu/projects/talairachdaemon.html) implemented in MEDx, which identifies brain regions and Brodmann areas using the stereotactic coordinate system outlined in the Talairach atlas [Talairach and Tournoux, 1988].

Statistical inferences were based on the theory of random Gaussian fields [Friston et al., 1995a]. All reported results are derived from second-level random-effects analyses. All reported PFC activations had probability values at the uncorrected $P < 0.02$ threshold with a cluster size threshold of 20; this corresponds to a per voxel false-positive $P < 0.000001$ [Forman et al., 1995]. This method of dealing with multiple comparisons has been used by our group [Knutson et al., 2004, 2006; Wood et al., 2003] as well as reported elsewhere [Konishi et al., 1998; Poldrack et al., 1999; Wagner et al., 2001]. It is based on the assumption that areas of true neural activity will activate areas of contiguous voxels, and the fact that there is a low probability of activation of a group of contiguous voxels due to chance alone. Due to a lack of a prior hypothesis for posterior portions of the brain, whole-brain activations were considered statistically significant if their probability value survived the false discovery rate (FDR) correction [Benjamini and Yekutieli, 2001; Yekutieli and Benjamini, 1999] for multiple comparisons at a threshold of $P < 0.05$.

Figure 2.

IAT performance by condition. (a) Gender IAT performance. (b) Racial IAT performance. Response latencies are significantly greater for the incongruent vs. congruent condition (gender: $P = 0.004$; racial: $P < 0.001$), indicating significant IAT effects. In both conditions, participants were quickest when performing the baseline condition.

RESULTS

Behavioral Data

Implicit association test

Mean response latencies for each of the three conditions (INCONG, CONG, BASE) for gender and racial IAT tasks separately are presented in Figure 2a,b. As expected, participants showed longer response times when incongruent stimuli were associated with the same response key, as compared with the congruent mapping. They were fastest when classifying stimuli as names or words in the baseline condition.

We calculated response time differences between the two conditions for each individual run of the gender and racial IAT tasks. These differences are shown in Figure 3a,b.

The overall difference (across runs) between incongruent and congruent conditions across both IAT tasks was 24.69 ms (main effect of condition) and was statistically significant ($t(19) = 4.72$, $P < 0.001$). The *first two runs* of the racial IAT showed significant IAT effects, with a trend towards significance on the last run ($t(19) = 7.11$, $P < 0.001$; $t(19) = 3.32$, $P = 0.004$; $t(19) = 1.88$, $P = 0.075$, for runs 1, 2, and 3 respectively). Response latencies for the gender IAT task for the incongruent condition were also significantly greater than those for the congruent conditions, with a mean difference of 11.66 ms across all three runs. Similar to the racial IAT task, the gender IAT effect was not present for the third run (differences were significant only for the second run, with a trend toward significance for the first run) ($t(19) = 2.05$, $P = 0.054$; $t(19) = 3.01$, $P = 0.007$ for runs 1 and 2 respectively). There was clearly an expo-

sure effect such that the IAT response bias was no longer present by the third run for each IAT task.

We also compared response latencies for each of the three conditions (incongruent, congruent, baseline) using a repeated-measures ANOVA with one between-subjects factor (order (2 levels)), and three within-subjects factors (IAT type, condition, run ($2 \times 3 \times 3$)). There was a significant interaction between IAT task and condition ($F(2,17) = 13.00$, $P < 0.001$), as participants showed a larger IAT effect on the racial task compared with the gender task, indicating that participants had more difficulty inhibiting stereotypic associations about race compared to automatic beliefs about gender. Importantly, the main effect of condition was highly significant ($F(2,17) = 119.32$, $P < 0.001$), and there was no main effect of order, ($F(1,18) = 0.094$, $P = 0.76$).

One sample t tests showed that D scores for the first run and for all runs for both racial and gender IAT data reached significance (see Fig. 4; $P < 0.001$ for the racial data; $P < 0.02$ for the gender data). D scores of the combined racial and gender IAT data were also significant (all runs; $P < 0.001$). Next, we compared D scores for race and gender using a repeated measures ANOVA with one between-subjects factor (order (2 levels)), and one within-subjects factor of IAT type (2 levels). There was a significant main effect of IAT type ($F(1,18) = 21.77$, $P < 0.001$) which was attributable to significantly larger D scores for the racial IAT data than for the gender IAT data (means = 0.34 and 0.12 respectively; $t(19) = 4.37$, $P < 0.001$). Unlike the RT results, there was also a main effect of order ($F(1,18) = 10.08$, $P = 0.005$) with the mean D score for racial IAT data for participants who performed the race block first significantly larger than the racial mean D score

Figure 3.

Mean response latencies for each run per condition. (a) Mean response latencies for gender IAT data. (b) Mean response latencies for racial IAT data. Note the different scales for the two graphs.

for those participants who performed the gender block first (means = 0.50 and 0.19, respectively; $t(9) = 4.04$, $P = 0.003$), providing further evidence for the D measure's greater power over the RT measure [Greenwald et al., 2003].

Error rates for the racial data were low (5.6%) and paralleled the pattern of results for the RT data (BASE $M = 3.67\%$, $SD = 2.48$; CONG $M = 5.53\%$, $SD = 2.67$; INCONG $M = 7.52\%$, $SD = 4.82$; $F(2, 118) = 27.72$, $P < .001$). There was a significant interaction between run and condition ($F(4, 76) = 2.89$, $P = 0.028$), attributable to a higher error rate on INCONG relative to CONG trials for runs 1 and 3 ($t(19) = 2.91$, $P = 0.009$; $t(19) = 2.37$, $P = 0.026$) respec-

tively). There was no significant difference between CONG and INCONG error rates for run 2 ($t(19) = -0.119$, ns).

Similarly, error rates for the gender data were low (3.7%) and paralleled the pattern of results for RT data (BASE $M = 2.93\%$, $SD = 2.24$; CONG $M = 3.68\%$, $SD = 3.19$; INCONG $M = 4.53\%$, $SD = 3.35$; $F(2, 118) = 7.63$, $P = 0.001$). There was a significant interaction between run and condition ($F(4, 76) = 4.76$, $P = 0.002$), attributable to a higher error rate on INCONG relative to CONG trials for runs 1 and 2 ($t(19) = 2.09$, $P = 0.05$; $t(19) = 2.91$, $P = 0.009$) respectively), which reversed by the last run so that more errors were made in CONG relative to INCONG (trend; $t(19) = -2.01$, $P = 0.059$). See Figure 5 for mean errors per run on combined racial and gender IAT data.

RT and error rate analyses indicate that with practice, participants became faster at performing INCONG trials while producing roughly the same number of errors, and slower at performing CONG trials with an increasing number of errors. Also, performance differences were more pronounced for the racial data, consistent with another IAT study where the IAT effect for race was larger than for other constructs [Greenwald et al., 1998].

Sex differences

Males had longer RTs than females in the CONG and INCONG conditions in both race and gender IATs (male overall mean = 784.41 ms; female overall mean = 719.05 ms; $F(1,58) \geq 6.26$; $P \leq 0.015$). In addition, there was a greater IAT effect (RT difference) on the racial IAT in male participants, as compared with female participants ($t(1,18) = 2.49$, $P = 0.023$). There was a trend for a greater IAT

Figure 4.

Mean D scores. Gender and race mean D scores for the first run and for all three runs together.

Figure 5.
Mean errors of each condition (baseline, congruent, incongruent) for each run for combined race and gender IAT data.

effect on the gender IAT in male participants, as compared with female participants ($t(1,18) = 1.97$, $P = 0.065$). No significant sex differences were found using D scores.

Explicit Measures

Results of correlational analyses are presented in Table I. In brief, although all of the explicit measures of gender and several of those of racial attitudes were correlated with one another, there was only one association between the implicit IAT effect and scores on the self-report questionnaires. There was an association in the expected direction between the implicit measure of racial D scores and the explicit measure of Modern Racism Scale scores (as low MRS scores indicate greater racism; $r = -0.52$, $P = 0.02$). The two implicit measures (RT bias and D scores) were highly correlated (gender: $r = 0.80$, $P < 0.001$; race: $r = 0.77$, $P < 0.001$), and the racial and gender D scores approached significance in correlation with each other ($r = 0.44$, $P = 0.051$).

Imaging Data

Since we did not find a main effect of order using RTs, we collapsed imaging data across order. We analyzed first the data for the main effects of condition, independent of IAT task (gender/racial), and then analyzed the data for both tasks separately. Results for contrasts CONG > BASE and INCONG > BASE were not significant. Results for combined gender and racial IAT data are presented in Table II. Areas of activation using the contrasts CONG >

INCONG and INCONG > CONG for combined IAT data are illustrated in Figure 6a,b.

Linear Contrasts

CONG > INCONG

We applied this contrast to test our hypotheses regarding the VMPFC regions' involvement in the representation or expression of automatic beliefs about gender and race. This analysis showed involvement of medial PFC in representing stereotypic attitudes, with activation within the right medial frontal gyrus (BA 10, $P = 0.004$, extending to anterior cingulate), and in right superior frontal gyrus (BA 8 and 9, $P = 0.005$) in the analysis of combined gender and racial IAT imaging data (see Fig. 6a). No significant PFC activation was found for the gender IAT data. Right medial frontal gyrus (BA 10, $P = 0.004$) and right insula (BA 13; $P < 0.001$) activation were found applying the same analysis to the racial IAT data. Detailed results are reported in Table II.

Since males had a larger racial IAT effect (using RTs) than females, and had a trend toward a larger gender IAT effect, we analyzed the imaging data for sex effects to determine if differences were evident there also. Males showed greater activation in left superior frontal gyrus (BA 8; $P = 0.001$) and anterior cingulate (BA 32; $P = 0.001$) in the analysis of combined gender and racial IAT imaging data for the CONG versus INCONG contrast. These areas of activation were in similar locations as for the whole group analysis, except on the left rather than the right for the superior frontal gyrus activation. Females had no PFC activation greater than males for this contrast. Using the

TABLE I. Correlations between implicit and explicit measures for gender and race

	Gender IAT effect	ASI	ASW	GSD	Gender D score			
Gender IAT effect	1.000	-0.258	-0.126	0.303	0.796**			
ASI	-0.258	1.000	-0.771**	-0.667**	-0.181			
ASW	-0.126	-0.771**	1.000	0.474*	-0.016			
GSD	0.303	-0.667**	0.474*	1.000	0.079			
Gender D score	0.796**	-0.181	-0.016	0.079	1.000			
	Racial IAT effect	MRS	DIV	DIS	Should	Would	RSD	Racial D Score
Racial IAT effect	1.000	-0.344	-0.146	-0.083	0.196	0.415	-0.092	0.769**
MRS	-0.344	1.000	0.381	0.652**	-0.508*	-0.571**	0.116	-0.516*
DIV	-0.146	0.381	1.000	0.525*	-0.185	-0.148	0.074	-0.071
DIS	-0.083	0.652**	0.525*	1.000	-0.423	-0.511*	0.218	-0.247
Should	0.196	-0.508*	-0.185	-0.423	1.000	0.674**	-0.092	0.216
Would	0.415	-0.571**	-0.148	-0.511*	0.674**	1.000	-0.439	0.356
RSD	-0.092	0.116	0.074	0.218	-0.092	-0.439	1.000	-0.023
Racial D Score	0.769**	-0.516*	-0.071	-0.247	0.216	0.356	-0.023	1.000

Consistent with previous studies, the IAT effect is not correlated with explicit measures of attitudes about gender or race. IAT effect, difference in reaction time between incongruent and congruent conditions combined for all runs; ASI, Ambivalent Sexism Inventory; ASW, Attitudes Toward Women; GSD, Gender Semantic Differential (GSD score on men/GSD score on women); Gender D score, Greenwald's D measure for gender IAT tasks; MRS, Modern Racism Scale; DIV, Diversity Scale; DIS, Discrimination Scale; Should, scale from Should-Would Discrepancy Questionnaire; Would, scale from Should-Would Discrepancy Questionnaire; RSD, Racial Semantic Differential (RSD score on whites/RSD score on blacks); Racial D score, Greenwald's D measures for Racial IAT tasks.

**Correlation is significant at the 0.01 level (2-tailed).

*Correlation is significant at the 0.05 level (2-tailed).

gender IAT data, males had no activation greater than females; females had extensive areas of activation greater than males' in bilateral superior frontal gyrus (BA 10; $P < 0.007$), right middle frontal gyrus (BA 46; $P < 0.001$), and left inferior and medial frontal gyrus (BA 9, 10; $P \leq 0.003$). In contrast to the gender IAT data, using the racial IAT data, males had greater activation than females in medial PFC areas similar to but more extensive than the above, including anterior cingulate (BA 32; $P < 0.001$) which extended into bilateral superior frontal gyrus (right BA 10 and left BA 9), and bilateral middle frontal gyrus (BA 8; $P \leq 0.001$). Females had no activation greater than males for this contrast.

INCONG > CONG

This contrast was performed to test our hypotheses regarding the involvement of PFC regions while participants associated incompatible concepts and suppressed stereotypic attitudes. We predicted recruitment of dorsolateral PFC while participants performed the incongruent condition of the IAT. We found evidence of left dorsolateral PFC involvement using the comparison of INCONG > CONG for the combined IAT data (BA 8, 9, and 46, at $P = 0.001$; see Fig. 6b). As expected, no activation was detected in anteromedial PFC in contrast to the results from the CONG > INCONG comparison. Contrary to hypotheses, we did not find greater activation in ACC or lateral OFC for the incongruent condition relative to the congruent condition. No significant PFC activation was found for the

gender or racial IAT data when analyzed separately. Detailed results are reported in Table II.

Race versus gender IAT data

We applied an analysis of the racial IAT data compared to the gender IAT data for the CONG > INCONG comparison to investigate the neural regions involved in the representation or expression of automatic beliefs about race versus those about gender. Since we had no hypothesis about this contrast, we looked at the whole brain using $P = 0.02$ with a cluster size of 20. We found no significant activation for the gender IAT data greater than the racial IAT data, but did find greater activation in left fusiform gyrus (BA 20, $P = 0.003$) and inferior and middle temporal gyri (BA 21, $P = 0.002$) using the racial IAT data minus the gender IAT data for the CONG > INCONG contrast. See Table III.

Regression Analyses

Analysis of the RT differences in correlation with PFC activation (CONG > INCONG), which results in areas showing greater activation as the level of attitudinal bias increases, resulted in medial frontal gyrus activation ([peak voxel = 0, 63, 8], BA 10; $P < 0.001$), extending into anterior cingulate ([0, 43, 9], BA 32; $P = 0.001$) for the combined IAT data. However, there was no significant correlation between PFC activation and the reaction time bias for the gender IAT data alone. For the racial IAT data, there was a correlation between RT bias and extensive areas of atti-

TABLE II. PFC activation

Anatomical localization of the peak of the cluster	Cluster size	Talairach coordinates			<i>t</i> -score	<i>P</i> -value
		<i>x</i>	<i>y</i>	<i>z</i>		
COMBINED CONGRUENT-INCONGRUENT						
R medial frontal gyrus (BA 10) extending to superior frontal gyrus (BA 9) and R anterior cingulate (BA 10)	50	20	51	9	3.02	0.004
R superior frontal gyrus (BA 8) extending to BA 9	22	16	49	38	2.89	0.005
COMBINED INCONGRUENT-CONGRUENT						
L middle frontal gyrus (BA 8 extending into BA 9 and 46)	54	-51	18	43	3.56	0.001
GENDER CONGRUENT-INCONGRUENT						
No significant activation						
GENDER INCONGRUENT-CONGRUENT						
No significant activation						
RACE CONGRUENT-INCONGRUENT						
R superior and medial frontal gyrus (BA 10)	23	12	59	19	2.94	0.004
R superior temporal gyrus (BA 42) extending into R insula (BA 13)	161	63	-30	16	4.01	<0.001
RACE INCONGRUENT-CONGRUENT						
No significant activation						
MALES > FEMALES: COMBINED CONGRUENT-INCONGRUENT						
L superior frontal gyrus, BA 8	33	-20	49	42	3.71	0.001
Anterior cingulate, BA 32	28	0	36	13	3.58	0.001
FEMALES > MALES: COMBINED CONGRUENT-INCONGRUENT						
No significant activation						
MALES > FEMALES: GENDER IAT CONGRUENT-INCONGRUENT						
No significant activation						
FEMALES > MALES: GENDER IAT CONGRUENT-INCONGRUENT						
R middle frontal gyrus, BA 46, extending into middle and superior BA 10	216	40	32	24	4.02	<0.001
L inferior frontal gyrus, BA 9	531	-59	9	22	3.89	0.001
L medial/superior frontal gyrus, BA 10	43	-16	47	12	3.12	0.003
MALES > FEMALES: RACIAL IAT CONGRUENT-INCONGRUENT						
L anterior cingulate, BA 32, extending into superior frontal gyrus, R BA 10, and L BA 9	433	0	36	17	4.23	<0.001
R middle frontal gyrus, BA 8	76	28	33	46	4.16	<0.001
L superior/middle frontal gyrus, BA 9 and 8	121	-20	52	38	4.04	<0.001
FEMALES > MALES: RACIAL IAT CONGRUENT-INCONGRUENT						
No significant activation						

Anatomical localization of the PFC and anterior cingulate activation associated with the basic fMRI data analyses for the combined and separate gender and race IATs, and for males and females combined and separate. $P = 0.02$ uncorrected with an extent threshold = 20. R, right; L, left.

vation including anterior cingulate ([4, 40, 16], BA 32; $P < 0.001$), bilateral medial frontal gyrus ([0, 33, 31], left BA 9; $P < 0.001$), ([4, 51, 1], right BA 10; $P < 0.001$), right post-central gyrus ([51, -17, 45], BA 3; $P < 0.001$), right inferior and middle frontal gyrus ([48, 14, 40], BA 8); ([55, 9, 33], BA 9; $P < 0.017$), right superior temporal gyrus ([51, -35, 9], BA 41; $P = 0.001$), right insula ([44, -19, 1], BA 13; $P = 0.001$), and left mammillary body ([-4, -12, -13]; $P = 0.001$) (all but BA 8 reaching significance at an FDR-corrected p of 0.05).

Similarly to the correlation with RT differences, the D scores [Greenwald et al., 2003] correlated with activation (CONG > INCONG) in right medial ([4, 40, 20], BA 9 and

10; $P < 0.008$), left inferior frontal gyri ([-51, 27, 6], BA 45 and 46; $P < 0.012$), right insula ([44, -19, 5], BA 13; $P = 0.001$), and left putamen extending into the caudate and claustrum ([-20, 8, -4]; $P = 0.001$) for the combined IAT data, and in right amygdala ([24, -8, -10]; $P < 0.001$) for the gender IAT data alone. There was a large area of activation correlated with D scores for the racial IAT data, including medial frontal gyrus ([0, 45, 38], BA 8; $P < 0.001$) which extended into bilateral anterior and posterior cingulate gyri, and right insula ([40, -19, 5], BA 13; $P = 0.001$) extending into right superior temporal lobe ([59, -19, 1]; $P = 0.002$). These racial data correlations using D scores were significant after FDR correction for multiple

Figure 6.

Combined gender and racial IAT imaging data. (a). Results for the CONG > INCONG contrast from the group random effects analysis across 20 participants with a threshold of $P < .02$ (uncorrected) and cluster threshold of 20, superimposed on a single subject template image. Analyses indicated significantly activated voxels in right medial frontal gyrus (BA 10, [20, 51, 9], $t = 3.02$, $P = 0.004$), and right superior frontal gyrus (BA 8,

[16, 49, 38], $t = 2.89$, $P = 0.005$). (b). Results for the INCONG > CONG contrast from the group random effects analysis across 20 participants with a threshold of $P < 0.02$ (corrected) and cluster threshold of 20, superimposed on single subject template image. Activated voxels in left middle frontal gyrus (BA 8, [-51, 18, 43], $t = 3.56$, $P = 0.001$).

comparisons, and resulted in similar findings as the RT correlations.

DISCUSSION

We used a modified version of the IAT [Greenwald et al., 1998] to identify the neural substrates underlying automatic beliefs about gender and race, and the structures recruited when these stereotypic attitudes were suppressed. We hypothesized that expression of stereotypic beliefs would be represented within medial PFC regions, particularly the VMPFC/medial OFC and that inhibiting these associations would activate dorsal and lateral PFC structures including lateral OFC, DLPFC, and ACC. Performance data on the racial and gender IAT tasks were consistent with previous behavioral studies. Specifically, the predicted IAT effect was observed in our sample and was modified through context and exposure [Chee et al., 2000; Dasgupta and Greenwald, 2001; Greenwald et al., 1998; Mitchell et al., 2003a; Rudman et al., 2001a,b]. As the education level of our participants was high, it is likely that the IAT effect would have been even larger if the education level was lower, as education level is negatively correlated with measures of racism and sexism [Coenders and Scheepers, 2003; Flammer, 2001]. The

correlation between the implicit measure of racial data D scores and the explicit measure of racism reflected in the Modern Racism Scale scores [McConahay et al., 1981] supports the construct validity of the racial D scores in our study. Our fMRI results provide support for involvement of medial PFC regions, including anteromedial PFC and the rostral ACC, while participants made stereotypic associations about gender and race. In contrast, dorsolateral PFC structures were recruited when participants were required to inhibit these well-learned associations about gender and race.

We were able to detect activation in anteromedial PFC and structures associated with the medial PFC network (AMPFC, rostral ACC, insula), but could not evaluate the most ventral regions of medial PFC because of signal dropout due to susceptibility artifact [Merboldt et al., 2001; Ojemann, et al., 1997]. Further studies using MR parameters designed to optimize the signal in the ventromedial PFC [Chen et al., 2003; Deichmann et al., 2003] are necessary to evaluate this area's contribution to mediating stereotypic attitudes.

Functional imaging studies of other automatic processes have detected activation in medial PFC areas similar to the regions found in our study. These include execution of overlearned visuomotor associations [Koechlin et al., 2002],

TABLE III. Whole brain activation for gender versus race IAT analyses

Anatomical localization of the peak of the cluster	Cluster size	Talairach coordinates			<i>t</i> -score	<i>P</i> -value
		<i>x</i>	<i>y</i>	<i>z</i>		
GENDER > RACE: CONGRUENT-INCONGRUENT						
No significant activation						
RACE > GENDER: CONGRUENT-INCONGRUENT						
L inferior/middle temporal gyrus (BA 21)	29	−59	−8	−13	3.22	0.002
L fusiform gyrus (BA 20)	20	−44	−24	−12	3.14	0.003

Anatomical localization associated with the fMRI data analyses for gender versus race ($P = 0.02$ uncorrected, and extent threshold = 20). L, left.

semantic priming [Rossell et al., 2001], and social judgments [Mitchell et al., 2002b; Moll et al., 2002]. The medial PFC region is characterized by relatively high baseline rates of metabolic activity [Raichle et al., 2001] and connectivity to the amygdala, hypothalamus, and periaqueductal gray region [PAG; Ongur et al., 1998]. Accordingly, it has been suggested that the medial PFC including the insula and other associated structures may function as a visceromotor system that modulates visceral activity in response to affective stimuli, allowing rapid detection and assessment of potential danger in the environment [Gallese et al., 2004]. Similarly, representation of learned associations about social groups within medial PFC regions may allow for rapid access to beliefs toward social groups that are negatively evaluated or are perceived as threatening.

We also found activation within the rostral subdivision of the ACC while participants made stereotypic associations about gender and race. Contrary to hypotheses, we did not detect ACC activation when they suppressed these associations. This lack of ACC activation for incongruent tasks adds to prior work showing how activation of particular parts of the frontal lobes is task-specific [MacDonald et al., 2000]. Others have made a distinction between ventral-rostral and dorsal ACC, suggesting that the former is involved in affective processing and evaluating the salience of information, while the dorsal ACC is recruited during selective attention and response inhibition [Bush et al., 2000; Carter et al., 1995, 1998; MacDonald et al., 2000; van der Meer and Schmidt, 1992; Whalen, 1998; Yamasaki et al., 2002]. This is consistent with the ventral-rostral ACC activation found during the correlation analysis between stereotypic tasks and RT (combined IAT data), while dorsal ACC activation resulted from the correlation between incongruent tasks and D scores (gender IAT data). In addition, representation of stereotypic attitudes within rostral ACC may be attributable to their social significance and greater familiarity, rendering them more salient [Elliott et al., 2000].

Based on its reciprocal connectivity to the OFC [Ongur and Price, 2000], and evidence from clinical and imaging data [Adolphs, 1999; Phelps et al., 2000; Winston et al., 2002] (but see [Phelps et al., 2003]), we expected a role for the amygdala in the expression of a variety of stereotypic

beliefs. Consistent with our hypothesis, we found amygdala activation as a result of the correlation with D scores using the gender IAT data. This shows that amygdala activation while participants were endorsing a gender stereotype was significantly correlated with an objective measure of their gender bias. In other words, participants with greater implicit gender bias had greater amygdala activation during performance of gender stereotypic tasks. While the viewing of faces of another race has been shown to activate the amygdala [Hart et al., 2000; Lieberman et al., 2005], further research is needed before stating that performing gender stereotypic tasks activates the amygdala, as amygdala activation to gender stereotyping has not been previously reported in the literature to our knowledge, although Phillips et al. found right amygdala activation for neutral faces while participants were determining the sex of the face [Phillips et al., 2001], and Dubois et al. found left amygdala activation while participants were determining the sex of the unknown versus known faces [Dubois et al., 1999]. Gender discrimination/stereotyping tasks may be some of the functions of the amygdala, which in general include providing a quick, unconscious, affect- [Cunningham et al., 2003] and relevance-based [Sander et al., 2003] evaluation of the environment that prepares one for immediate action.

Activation in the insula, which connects to the amygdala [Stefanacci and Amaral, 2000], was found while participants made stereotypic associations using racial IAT data. In addition, in correlation with D scores (combined, gender, and racial data) and with RTs (racial IAT data), there was greater activation in the insula and caudate while participants made stereotypic associations. These areas are consistent with the functional neuroanatomical connectivity of the OFC [Cavada et al., 2000; Ongur et al., 2003; Ongur and Price, 2000], and have been implicated in practice effects [Raichle et al., 1994], and processing of reward and other salient stimuli [Gottfried et al., 2003; O'Doherty et al., 2001, 2003; Zink et al., 2003]. In summary, the expression of stereotypic beliefs appears to activate regions within the OFC network that have also been implicated in the execution of automatic processes, including representation of reward and socially-relevant stimuli. Alternatively, insular activation may be due partly to unpleasant feelings

such as disgust [Krolak-Salmon et al., 2003; Wicker et al., 2003] experienced when white participants are presented with stereotypically black male names.

In contrast, dorsolateral regions of PFC including BA 9 and 46 were activated when participants were required to suppress automatic associations about race and gender; these areas have strong reciprocal connections with dorsal ACC [Devinsky et al., 1995]. These findings are consistent with previous fMRI findings supporting their role in cognitive control, maintaining task set, inhibitory control, and behavioral change due to changing task requirements [Carter et al., 1998; Collette et al., 2001; Cools et al., 2004; D'Esposito et al., 1999; Garavan et al., 2002; Jonides et al., 1998]. Thus, in contrast to representation of stereotypic associations within a medial PFC network, suppression of automatic beliefs activated DLPFC.

We also considered that the anteromedial PFC activation observed in the comparison of CONG > INCONG may reflect greater deactivation of anteromedial PFC in the incongruent task, which requires greater attentional resources, rather than an increase in activity associated with the congruent condition [Gusnard et al., 2001; Raichle et al., 2001]. This interpretation would still lend support for our hypotheses. Less reduction in anteromedial PFC activity during the congruent condition would suggest that the processes involved in activating stereotypic associations overlap with similar processes engaged during resting baseline [Mitchell et al., 2002b]. As a result, these attitudes may be more readily activated and accessible.

The differences between CONG versus INCONG contrasts showed hemispheric differences, in that the right hemisphere was activated more for CONG than INCONG contrasts, while left DLPFC was activated for INCONG minus CONG contrasts. The left DLPFC activation for INCONG minus CONG contrasts is consistent with the results of other IAT studies, including one looking at flower and insect categories with pleasant and unpleasant attributes [Chee et al., 2000], and another by the present authors which investigated political attitudes for Democrats and Republicans and pleasant and unpleasant words [Knutson et al., 2006] and which showed more activation in left DLPFC than right during inhibitory tasks. Chee et al. attribute the left DLPFC activation during inhibitory tasks as a type of inhibition driven by the need to access semantics [Thompson-Schill et al., 1997] and manipulate that information but also to inhibit the implicit tendency to respond in a congruent way. There is a consensus that DLPFC plays an important role in the top-down control of behavior in situations requiring supervision partly due to an attention mechanism [MacDonald et al., 2000].

Regression analyses showed areas that are more activated when RT differences between CONG and INCONG tasks (or D scores) were larger, meaning there was a larger attitudinal bias. For the stereotypic comparisons, both D scores and RT differences were correlated with much larger areas of activation for the racial IAT data than the gender IAT data, and the basic analyses resulted in greater

activation for racial than for gender data. This may be partly perhaps because of the greater regularity in the racial RT differences and D scores, but also perhaps because race was more salient than gender to the participants of our study when making stereotypic decisions. A contributing factor may be that none of the participants were black while seven were female. The presence of females may have kept the IAT effect for gender smaller (there was a trend toward a smaller IAT effect in females than in males on the gender IAT), while the lack of black participants may have led to a stronger IAT effect for race due to the all white participants' attitude toward the outgroup [Olsson et al., 2005]. When racial and gender IAT data were compared directly, there was greater activation during stereotypic tasks for the racial IAT data (in left fusiform gyrus), but no greater activation for the gender IAT data. This activation also is consistent with a low frequency effect in visual orthographic processing for the names [Kronbichler et al., 2004], as the black names were less familiar than the white names used in this study, while the male and females names were of equal familiarity.

This is the first study known to the authors to report gender differences in PFC activation using the IAT. Why males had relatively greater activation in superior frontal gyrus and other PFC regions than females for racially-stereotypic conditions, while females had relatively greater activation in these regions for gender-stereotypic conditions remains to be explored. One possibility is that these frontal brain regions code associated, more complex, knowledge about each of the stereotypes, and that we find gender differences in activation due to gender-specific social preferences, experience, concerns, and knowledge acquisition.

CONCLUSION

In summary, the results of this study support the contribution of several PFC regions in the representation and modification of automatic beliefs about gender and race. Distinct medial and lateral networks within the PFC have been delineated based on cortico-cortical connectivity, as well as their connections to sensory, limbic, striato-thalamic and visceromotor structures in other parts of the brain [Ongur et al., 2003; Ongur and Price, 2000]. Our data suggest that there are dissociable roles for the medial and lateral PFC circuits in the expression versus the suppression of stereotypic attitudes. We demonstrate that anteromedial PFC structures may be involved in the expression of stereotypic associations, while the DLPFC in humans is recruited during inhibition of complex social associations, in the form of automatic beliefs about race and gender [Roberts and Wallis, 2000; Wood and Grafman, 2003].

ACKNOWLEDGMENT

We thank Martijn Jansma, Ph.D. for helpful comments on an earlier version of the manuscript.

REFERENCES

- Adolphs R (1999): Social cognition and the human brain. *Trends Cogn Sci* 3:469–479.
- Ashburner J, Friston KJ (1999): Nonlinear spatial normalization using basis functions. *Hum Brain Mapp* 7:254–266.
- Baron AS, Banaji MR (2006): The development of implicit attitudes. *Psychol Sci* 17:53–58.
- Bechara A, Damasio AR, Damasio H, Anderson SW (1994): Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* 50:7–15.
- Bellezza FS, Greenwald AG, Banaji MR (1986): Words high and low in pleasantness as rated by male and female college students. *Behav Res Methods Instrum Comput* 18:299–303.
- Benjamini Y, Yekutieli D (2001): The control of false discovery rate in multiple testing under dependency. *Ann Stat* 29:1165–1188.
- Bird H, Franklin S, Howard D (2001): Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behav Res Methods Instrum Comput* 33:73–79.
- Bush G, Luu P, Posner MI (2000): Cognitive and emotional influences in anterior cingulate cortex. *Trends Cogn Sci* 4:215–222.
- Carter CS, Mintun M, Cohen JD (1995): Interference and facilitation effects during selective attention: An H215O PET study of Stroop task performance. *Neuroimage* 2:264–272.
- Carter CS, Braver TS, Barch DM, Botvinick M, Noll D, Cohen JD (1998): Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science* 280:747–749.
- Cavada C, Company T, Tejedor J, Cruz-Rizzolo RJ, Reinoso-Suarez F (2000): The anatomical connections of the macaque monkey orbitofrontal cortex. A review. *Cereb Cortex* 10:220–242.
- Chee MWL, Sriram N, Soon CS, Lee KM (2000): Dorsolateral prefrontal cortex and the implicit association of concepts and attributes. *Neuroreport* 11:135–140.
- Chen NK, Dickey CC, Yoo SS, Guttman CR, Panych LP (2003): Selection of voxel size and slice orientation for fMRI in the presence of susceptibility field gradients: Application to imaging of the amygdala. *Neuroimage* 19:817–825.
- Coenders AU, Scheepers M (2003): The effect of education on nationalism and ethnic exclusionism: An international comparison. *Polit Psychol* 24:313–343.
- Collette F, van der Linden M, Delfiore G, Degueldere C, Luxen A, Salmon E (2001): The functional anatomy of inhibition processes investigated with the Hayling Task. *Neuroimage* 14:258–267.
- Coltheart M (1981): The MRC psycholinguistic database. *Q J Exp Psychol A* 33:497–505.
- Cools R, Clark L, Robbins TW (2004): Differential responses in human striatum and prefrontal cortex to changes in object and rule relevance. *J Neurosci* 24:1129–1135.
- Cunningham WA, Johnson MK, Gatenby JC, Gore JC, Banaji MR (2003): Neural components of social evaluation. *J Pers Soc Psychol* 85:639–649.
- Damasio H, Grabowski T, Frank R, Galaburda AM, Damasio AR (1994): The return of Phineas Gage: Clues about the brain from the skull of a famous patient. *Science* 264:1102–1105.
- Dasgupta N, Greenwald AG (2001): On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *J Pers Soc Psychol* 81:800–814.
- Deichmann R, Gottfried JA, Hutton C, Turner R (2003): Optimized EPI for fMRI studies of the orbitofrontal cortex. *Neuroimage* 19:430–441.
- D'Esposito M, Postle BR, Jonides J, Smith EE (1999): The neural substrate and temporal dynamics of interference effects in working memory as revealed by event-related functional MRI. *Proc Natl Acad Sci USA* 96:7514–7519.
- Devinsky O, Morrell M, Vogt BA (1995): Contributions of anterior cingulate cortex to behaviour. *Brain* 118 (Part 1):279–306.
- Dubois S, Rossion B, Schiltz C, Bodart JM, Michel C, Bruyer R, Crommelinck M (1999): Effect of familiarity on the processing of human faces. *Neuroimage* 9:278–289.
- Elliott R, Dolan RJ, Frith CD (2000): Dissociable functions in the medial and lateral orbitofrontal cortex: Evidence from human neuroimaging studies. *Cereb Cortex* 10:308–317.
- Flammer LJ (2001): The nature of prejudice: Dimensions and patterns of racism, sexism, classism, and heterosexism among social groups. Temple University. Dissertation Abstract. p 2534.
- Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC (1995): Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. *Magn Reson Med* 33:636–647.
- Friston KJ, Ashburner J, Frith CD, Poline JB, Heather JD, Frackowiak RSJ (1995a): Spatial registration and normalisation of images. *Hum Brain Mapp* 2:165–189.
- Friston KJ, Holmes AP, Poline J-B, Grasby PJ, Williams SCR, Frackowiak RSJ, Turner R (1995b): Analysis of fMRI time-series revisited. *Neuroimage* 2:45–53.
- Friston KJ, Holmes AP, Worsley KJ (1999): How many subjects constitute a study? *Neuroimage* 10:1–5.
- Gallese V, Keysers C, Rizzolatti G (2004): A unifying view of the basis of social cognition. *Trends Cogn Sci* 8:396–403.
- Garavan H, Ross TJ, Murphy K, Roche RA, Stein EA (2002): Dissociable executive functions in the dynamic control of behavior: Inhibition, error detection, and correction. *Neuroimage* 17:1820–1829.
- Glick P, Fiske ST (1996): The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *J Pers Soc Psychol* 70:491–512.
- Gottfried JA, O'Doherty J, Dolan RJ (2003): Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* 301:1104–1107.
- Greenwald AG, Nosek BA (2001): Health of the implicit association test at age 3. *Z Exp Psychol* 48:85–93.
- Greenwald AG, McGhee DE, Schwartz JLK (1998): Measuring individual differences in implicit cognition: The implicit association test. *J Pers Soc Psychol* 74:1464–1480.
- Greenwald AG, Nosek BA, Banaji MR (2003): Understanding and using the implicit association test. I. An improved scoring algorithm. *J Pers Soc Psychol* 85:197–216.
- Gusnard DA, Akbudak E, Shulman GL, Raichle ME (2001): Medial prefrontal cortex and self-referential mental activity: Relation to a default mode of brain function. *Proc Natl Acad Sci USA* 98:4259–4264.
- Harlow JM (1848): Passage of an iron rod through the head. *Boston Med Surg J* 39:389–393.
- Hart AJ, Whalen PJ, Shin LM, McInerney SC, Fischer H, Rauch SL (2000): Differential response in the human amygdala to racial outgroup vs ingroup face stimuli. *Neuroreport* 11:2351–2355.
- Hofmann W, Gawronski B, Gschwendner T, Le H, Schmitt M (2005): A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Pers Soc Psychol Bull* 31:1369–1385.
- Hornak J, Rolls ET, Wade D (1996): Face and voice expression identification in patients with emotional and behavioural changes following ventral frontal lobe damage. *Neuropsychologia* 34:247–261.

- Jonides J, Schumacher EH, Smith EE, Koeppel RA, Awh E, Reuter-Lorenz PA, Marshuetz C, Willis CR (1998): The role of parietal cortex in verbal working memory. *J Neurosci* 18:5026–5034.
- Knutson KM, Wood JN, Grafman J (2004): Brain activation in processing temporal sequence: An fMRI study. *Neuroimage* 23:1299–1307.
- Knutson KM, Wood JN, Spampinato MV, Grafman J (2006): Politics on the brain: An fMRI investigation. *Soc Neurosci* 1:25–40.
- Koechlin E, Danek A, Burnod Y, Grafman J (2002): Medial prefrontal and subcortical mechanisms underlying the acquisition of motor and cognitive action sequences. *Neuron* 35:371–381.
- Konishi S, Nakajima K, Uchida I, Sekihara K, Miyashita Y (1998): No-go dominant brain activity in human inferior prefrontal cortex revealed by functional magnetic resonance imaging. *Eur J Neurosci* 10:1209–1213.
- Krolak-Salmon P, Henaff M-A, Isnard J, Tallon-Baudry C, Guenot M, Vighetto A, Bertrand O, Mauguier F (2003): An attention modulated response to disgust in human ventral anterior insula. *Ann Neurol* 53:446–453.
- Kronbichler M, Hutzler F, Wimmer H, Mair A, Staffen W, Ladurner G (2004): The visual word form area and the frequency with which words are encountered: Evidence from a parametric fMRI study. *Neuroimage* 21:946–953.
- Lieberman MD, Hariri A, Eisenberger NI, Bookheimer SY (2005): An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. *Nat Neurosci* 8:720–722.
- MacDonald AW, Cohen JD, Stenger VA, Carter CS (2000): Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science* 288:1835–1838.
- Mah LW, Arnold MC, Grafman J (2005): Deficits in social knowledge following damage to ventromedial prefrontal cortex. *J Neuropsychiatry Clin Neurosci* 17:66–74.
- McConahay JB, Hardee BB, Batts V (1981): Has racism declined in America? It depends on who is asking and what is asked. *J Conflict Resolut* 25:563–579.
- Merboldt K, Fransson P, Bruhn H, Frahm J (2001): Functional MRI of the human amygdala? *Neuroimage* 14:253–257.
- Milne E, Grafman J (2001): Ventromedial prefrontal cortex lesions in humans eliminate implicit gender stereotyping. *J Neurosci* 21:RC150:(1–6).
- Mitchell DG, Colledge E, Leonard A, Blair RJ (2002a): Risky decisions and response reversal: Is there evidence of orbitofrontal cortex dysfunction in psychopathic individuals? *Neuropsychologia* 40:2013–2022.
- Mitchell JP, Heatherton TF, Macrae CN (2002b): Distinct neural systems subserve person and object knowledge. *Proc Natl Acad Sci USA* 99:15238–15243.
- Mitchell CJ, Anderson NE, Lovibond PF (2003a): Measuring evaluative conditioning using the implicit association test. *Learn Motiv* 34:203–217.
- Mitchell JP, Nosek BA, Banaji MR (2003b): Contextual variations in implicit evaluation. *J Exp Psychol Gen* 132:455–469.
- Moll J, de Oliveira-Souza R, Eslinger PJ, Bramati IE, Mourao-Miranda J, Andreiulo PA, Pessoa L (2002): The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. *J Neurosci* 22:2730–2736.
- Monteith MJ, Voils CI (1998): Proneness to prejudiced responses: Toward understanding the authenticity of self-reported discrepancies. *J Pers Soc Psychol* 75:901–916.
- O'Doherty J, Kringsbach ML, Rolls ET, Hornak J, Andrews C (2001): Abstract reward and punishment representations in the human orbitofrontal cortex. *Nat Neurosci* 4:95–102.
- O'Doherty J, Critchley HD, Deichmann R, Dolan RJ (2003): Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. *J Neurosci* 23:7931–7939.
- Ojemann JG, Akbudak E, Snyder AZ, McKinstry RC, Raichle ME, Conturo TE (1997): Anatomic localization and quantitative analysis of gradient refocused echo-planar fMRI susceptibility artifacts. *Neuroimage* 6:156–167.
- Olsson A, Ebert JP, Banaji MR, Phelps EA (2005): The role of social groups in the persistence of learned fear. *Science* 309:785–787.
- Ongur D, Price JL (2000): The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys and humans. *Cereb Cortex* 10:206–219.
- Ongur D, An X, Price JL (1998): Prefrontal cortical projections to the hypothalamus in macaque monkeys. *J Comp Neurol* 401:480–505.
- Ongur D, Ferry AT, Price JL (2003): Architectonic subdivision of the human orbital and medial prefrontal cortex. *J Comp Neurol* 460:425–449.
- Payne BK (2005): Conceptualizing control in social cognition: How executive function modulates the expression of automatic stereotyping. *J Pers Soc Psychol* 89:488–503.
- Phelps EA, O'Connor KJ, Cunningham WA, Funayama ES, Gatenby JC, Gore JC, Banaji MR (2000): Performance on indirect measures of race evaluation predicts amygdala activation. *J Cogn Neurosci* 12:729–738.
- Phelps EA, Cannistraci CJ, Cunningham WA (2003): Intact performance on an indirect measure of race bias following amygdala damage. *Neuropsychologia* 41:203–208.
- Phillips ML, Medford N, Young AW, Williams L, Williams SC, Bullmore ET, Gray JA, Brammer MJ (2001): Time courses of left and right amygdala responses to fearful facial expressions. *Hum Brain Mapp* 12:193–202.
- Poldrack RA, Wagner AD, Prull MW, Desmond JE, Glover GH, Gabrieli JDE (1999): Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. *Neuroimage* 10:15–35.
- Raichle ME, Fiez JA, Videen TO, MacLeod AM, Pardo JV, Fox PT, Petersen SE (1994): Practice-related changes in human brain functional anatomy during nonmotor learning. *Cereb Cortex* 4:8–26.
- Raichle ME, MacLeod AM, Snyder AZ, Powers WJ, Gusnard DA, Shulman GL (2001): A default mode of brain function. *Proc Natl Acad Sci USA* 98:676–682.
- Roberts AC, Wallis JD (2000): Inhibitory control and affective processing in the prefrontal cortex: Neuropsychological studies in the common marmoset. *Cereb Cortex* 10:252–262.
- Rolls ET (1996): The orbitofrontal cortex. *Philos Trans R Soc Lond B Biol Sci* 351:1433–1444.
- Rolls ET (2004): The functions of the orbitofrontal cortex. *Brain Cogn* 55:11–29.
- Rolls ET, Hornak J, Wade D, McGrath J (1994): Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. *J Neurol Neurosurg Psychiatry* 57:1518–1524.
- Rossell SL, Bullmore ET, Williams SCR, David AS (2001): Brain activation during automatic and controlled processing of semantic relations: A priming experiment using lexical-decision. *Neuropsychologia* 39:1167–1176.
- Rudman LA, Ashmore RD, Gary ML (2001a): “Unlearning” automatic biases: The malleability of implicit prejudice and stereotypes. *J Pers Soc Psychol* 81:856–868.
- Rudman LA, Greenwald AG, McGhee DE (2001b): Implicit self-concept and evaluative implicit gender stereotypes: Self and

- ingroup share desirable traits. *Pers Soc Psychol Bull* 27:1164–1178.
- Sander D, Grafman J, Zalla T (2003): The human amygdala: An evolved system for relevance detection. *Rev Neurosci* 14:303–316.
- Spence JT, Helmreich R (1972): The Attitudes toward Women Scale: An objective instrument to measure attitudes toward the rights and roles of women in contemporary society. *JSAS Cat Select Doc Psychol* 2:66–67.
- Stefanacci L, Amaral DG (2000): Topographic organization of cortical inputs to the lateral nucleus of the macaque monkey amygdala: A retrograde tracing study. *J Comp Neurol* 421:52–79.
- Stone VE, Baron-Cohen S, Knight RT (1998): Frontal contributions to theory of mind. *J Cogn Neurosci* 10:640–656.
- Talairach P, Tournoux J (1988): *Co-planar Stereotaxic Atlas of the Human Brain*. Stuttgart: Thieme.
- Thompson-Schill SL, D'Esposito M, Aguirre GK, Farah MJ (1997): Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proc Natl Acad Sci USA* 94:14792–14797.
- van der Meer E, Schmidt B (1992): Finale, kausale und temporale Inferenzen - Analyse ihres kognitiven Hintergrundes. *Z Psychol* 200:303–320.
- Wagner AD, Pare-Blagoev EJ, Clark J, Poldrack RA (2001): Recovering meaning: Left prefrontal cortex guides controlled semantic retrieval. *Neuron* 31:329–338.
- Whalen PJ (1998): Fear, vigilance, and ambiguity: Initial neuroimaging studies of the human amygdala. *Psychol Sci* 7:177–188.
- Wheeler ME, Fiske ST (2005): Controlling racial prejudice: Social-cognitive goals affect amygdala and stereotype activation. *Psychol Sci* 16:56–63.
- Wicker B, Keysers C, Plailly J, Royet JP, Gallese V, Rizzolatti G (2003): Both of us disgusted in my insula: The common neural basis of seeing and feeling disgust. *Neuron* 40:655–664.
- Winston JS, Strange BA, O'Doherty J, Dolan RJ (2002): Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nat Neurosci* 5:277–283.
- Wittenbrink B, Judd CM, Park B (1997): Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *J Pers Soc Psychol* 72:262–274.
- Wood JN, Grafman J (2003): Human prefrontal cortex function: Processing and representational perspectives. *Nat Rev Neurosci* 4:139–147.
- Wood JN, Romero SG, Makale M, Grafman J (2003): Category-specific representations of social and nonsocial knowledge in the human prefrontal cortex. *J Cogn Neurosci* 15:236–248.
- Yamasaki H, LaBar KS, McCarthy G (2002): Dissociable prefrontal brain systems for attention and emotion. *Proc Natl Acad Sci USA* 99:11447–11451.
- Yekutieli D, Benjamini Y (1999): Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J Statist Plann Inference* 82:171–196.
- Zink CF, Pagnoni G, Martin ME, Dhamala M, Berns GS (2003): Human striatal response to salient nonrewarding stimuli. *J Neurosci* 23:8092–8097.