Neuropsychologia 50 (2012) 3600–3611

Contents lists available at SciVerse ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

Evaluative vs. trait representation in intergroup social judgments: Distinct roles of anterior temporal lobe and prefrontal cortex

Sam J. Gilbert [a], Jillian K. Swencionis [b], David M. Amodio [b,*]

[a] Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London WC1N 3AR, UK
[b] Department of Psychology, New York University, New York, NY, USA

ARTICLE INFO

Article history:
Received 13 June 2012
Received in revised form
20 August 2012
Accepted 3 September 2012
Available online 10 September 2012

Keywords:
FMRI
MVPA
Stereotyping
Bias
Orbitofrontal
Prefrontal
Temporal pole

ABSTRACT

When interacting with someone from another social group, one's responses may be influenced by both stereotypes and evaluations. Given behavioral results suggesting that stereotypes and evaluative associations operate independently, we used fMRI to test whether these biases are mediated by distinct brain systems. White participants viewed pairs of Black or White faces and judged them based on an evaluation (*who would you befriend*?) or a stereotype-relevant trait (*who is more likely to enjoy athletic activities*?). Multi-voxel pattern analysis revealed that a predominantly occipital network represented race in a context-invariant manner. However, lateral orbitofrontal cortex preferentially represented race during friendship judgments, whereas anterior medial prefrontal cortex preferentially represented race during trait judgments. Furthermore, representation of race in left temporal pole correlated with a behavioral measure of evaluative bias during friendship judgments and, independently, a measure of stereotyping during trait judgments. Whereas early sensory regions represent race in an apparently invariant manner, representations in higher-level regions are multi-componential and context-dependent.

1. Introduction

Social relationships are extremely complex, and a major goal of social cognitive neuroscience is to understand the mechanisms through which the human mind navigates the social world. When interacting with a person from a different social group, pre-existing beliefs about the group – referred to as stereotypes – influence our impressions (Allport, 1954). Social stereotypes may be learned through acculturation within one's society and may spring to mind automatically to influence impressions of outgroup members and behavior towards them (Darley & Gross, 1983; Devine, 1989). Although stereotypes may not accurately describe particular individuals, they can exert strong influence on how a perceiver approaches an interaction with a member of an outgroup (Bargh, Chen, & Burrows, 1996).

Social perceptions of outgroup members are also driven by evaluative processes (Fazio, Jackson, Dunton, & Williams, 1995). Like stereotypes, evaluative associations may influence judgments and behaviors without one's intention or awareness. For example, White Americans tend to associate Black people with negative concepts, compared with White people, in implicit behavioral responses (e.g. Fazio et al., 1995; Greenwald, McGhee, & Schwartz, 1998). These biases influence our behavior: people who show stronger implicit evaluative bias on behavioral tasks have been shown to respond in a less friendly manner toward a Black person during a real social interaction (Dovidio, Kawakami, & Gaertner, 2002; Fazio et al., 1995).

Recently, research has suggested that implicit stereotyping and evaluation processes may operate somewhat independently in behavior (Amodio & Devine, 2006), raising the possibility that these different facets of implicit bias reflect different underlying neural mechanisms. If these two forms of bias were found to reflect different underlying mechanisms, this finding might help explain why implicit evaluations and stereotypes appear to predict different forms of behavior. Furthermore, this finding would shed light on the mechanisms through which these two forms of social bias may be learned and unlearned, potentially informing interventions to reduce social prejudices. The present research was designed to test the hypothesis that racial evaluation and stereotypes reflect distinct neural processes.

1.1. Dissociation of stereotyping and evaluation in behavior

Although the concepts of evaluation (i.e., attitudes) and stereotyping have long been distinguished in social psychology research (Fiske, 1998), the notion that implicit stereotyping and evaluation processes might be rooted in different underlying

* Corresponding author. Tel.: +1 212 998 3875; fax: +1 212 995 4966.
E-mail address: david.amodio@nyu.edu (D.M. Amodio).

0028-3932/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neuropsychologia.2012.09.002

neurocognitive systems was proposed more recently by Amodio and Devine (2006). Amodio and Devine (2006) tested this distinction in a series of behavioral experiments. In each experiment, the authors assessed White American participants' stereotypic and evaluative associations with Black vs. White faces using two different Implicit Association Tests (IATs; Greenwald et al., 1998). One IAT, designed to measure evaluative associations, assessed the speed with which participants identified faces as Black vs. White and words as pleasant vs. unpleasant, when the response keys for these faces and words were paired in either a *congruent* (i.e., Black-unpleasant and White-pleasant) or *incongruent* (i.e. Black-pleasant and White-unpleasant) mapping. Used in hundreds of studies, this *evaluative IAT* has revealed a pervasive response bias among White American participants, such that they respond faster to Black faces paired with unpleasant words relative to pleasant words, in comparison with White faces (Nosek, Greenwald, & Banaji, 2007). This pattern is interpreted as indicating an implicit negative evaluation of Blacks, compared with Whites.

Amodio and Devine (2006) designed a second IAT to assess stereotypic associations, independently of evaluative associations. This *stereotyping IAT* was identical in task structure to the evaluative IAT, but it compared the speed with which participants categorized Black (vs. White) faces and athletic- vs. intelligence-related words in congruent vs. incongruent mappings. These word categories were chosen because athleticism and (un)intelligence are the two most common stereotypes of African Americans reported by White Americans (Devine & Elliot, 1995). Importantly, the words used in the stereotyping IAT were selected on the basis of pilot testing so that they were similar in valence. Because these target words were all moderately positive, this task could only be completed on the basis of semantic associations, and not evaluative associations.

In all three studies reported by Amodio and Devine (2006), participants exhibited significant racial bias on both the evaluative and stereotyping IATs. Yet scores on the two measures were uncorrelated (despite a combined sample size of 230), consistent with the idea that these two forms of bias reflect different underlying processes. More importantly, scores on evaluative and stereotyping IATs predicted different behavioral expressions of racial bias. Higher evaluative IAT scores uniquely predicted more negative feelings toward Black people and greater seating distance along a row of chairs from the belongings of their Black study partner. By contrast, higher stereotyping IAT scores uniquely predicted more stereotype-consistent trait impressions of a Black student and lower expectancies for the Black students' performance on a GRE-type academic test. Overall, these results demonstrate that implicit evaluative and stereotyping processes may operate independently at a behavioral level. However, it remains unclear how these behavioral effects relate to underlying neural mechanisms.

1.2. Cognitive neuroscience of implicit race bias

Several previous studies have examined the neural correlates of intergroup bias (for reviews, see Amodio, 2008; Eberhardt, 2005). The majority of these studies have investigated differences in brain activity associated with perception of Black versus White faces in White participants. Multiple brain regions have been reported to show such differences, such as amygdala, medial and lateral prefrontal cortex, hippocampus, and fusiform gyrus (Amodio, Harmon-Jones, & Devine, 2003; Golby, Gabrieli, Chiao, & Eberhardt, 2001; Lieberman, Hariri, Jarcho, Eisenberger, & Bookheimer, 2005; Wheeler & Fiske, 2005). The majority of these studies examined neural responses to pictures of White and Black individuals in passive viewing paradigms (Amodio et al., 2003; Cunningham et al., 2004; Phelps et al., 2000) or categorizing target faces (Lieberman et al., 2005; Wheeler & Fiske, 2005). Because these studies were interested in the emotional aspects of implicit prejudice, they focused primarily on amygdala activity in response to faces and, in most cases, observed greater amygdala activity while viewing Black than White faces. However, to date, research has not systematically examined the neural processes involved in conceptual representations of social ingroups vs. outgroups as they relate to evaluative and trait (i.e., stereotype) information.

1.3. The current research

The current research examined the neural processes involved in conceptual judgments of evaluative and semantic information. There were two broad aims. First, we investigated whether we could find evidence for distinct brain systems mediating evaluative versus semantic representations during social judgments. Whereas brain regions such as orbitofrontal cortex have been linked particularly to value-based assessment of stimuli (Grabenhorst & Rolls, 2008; Rushworth, Noonan, Boorman, Walton, & Behrens, 2011), other regions such as medial prefrontal cortex (Amodio & Frith, 2006; Krueger, Barbey, & Grafman, 2009) and temporal pole (Olson, Plotzker, & Ezzyat, 2007; Zahn et al. 2007) have been suggested to underlie conceptual social representations and their integration with decision making and emotion. However, the precise roles of these brain regions in social judgments are not well understood, despite recent evidence for separable functions (Gozzi, Raymont, Solomon, Koenigs, & Grafman, 2009). A second aim of the present study was to investigate whether neuroimaging results could be linked with behavioral IAT measures of evaluative and stereotyping bias. Insofar as the neuroimaging results can be linked with these behavioral indices, this provides evidence for their relevance to real-life behavior.

1.4. Multi-voxel pattern analysis and social cognitive neuroscience

The present study used the technique of multi-voxel pattern analysis (MVPA; Haynes & Rees, 2006; Norman, Polyn, Detre, & Haxby, 2006). Whereas previous studies have adopted standard univariate fMRI methodologies to investigate differential regional brain activity between perception of Black versus White faces, MVPA provides a finer-grained approach to distinguishing patterns of neural activity that is well suited for testing our hypotheses regarding neurocognitive representations. In studies using MVPA, fMRI data is typically investigated on a participant-by-participant basis, often using unsmoothed, unnormalized data. Two or more conditions are compared, and a pattern classifier is trained to distinguish voxel-by-voxel patterns of brain activity between those two conditions. Insofar as the classifier is able to distinguish these patterns, in a manner that generalizes to novel exemplars, this indicates that the brain region under investigation contains a representation that distinguishes these patterns. Thus, MVPA can be used to decode the representations contained within certain brain regions. This can apply to relatively low-level perceptual features, for example decoding the orientation of visually-presented lines by examining patterns of activity in primary visual cortex (Kamitani & Tong, 2005). It can also apply to higher-order brain regions, such as prefrontal cortex, and higher-level representations, such as the content of participants' delayed intentions (Gilbert, 2011; Gilbert, Armbruster, & Panagiotidi, 2012; Haynes et al. 2007). MVPA is an attractive technique to apply to social cognitive neuroscience, seeing as a major aim of this field is to investigate the nature of representations underlying social behavior. However, with few exceptions (e.g. Gilbert, Meuwese, Towgood, Frith, & Burgess, 2009; Natu, Raboy, & O'Toole, 2011; Ratner, Kaul, & Van Bavel, in press) this

approach has not yet been applied to the study of social processes.

To elicit the activation of either trait-based or evaluative representations of ingroup and outgroup members, participants viewed pairs of White or Black faces and either made a trait judgment related to an implicit stereotype (which person is more likely to enjoy athletic activities?) or an evaluative judgment (which person would you be more likely to befriend?). We then used MVPA in an attempt to decode whether participants were viewing White versus Black faces by looking at patterns of brain activity in these two judgment conditions. Insofar as race can be decoded by looking at brain activity across both conditions, this is consistent with a relatively invariant representation of race in the relevant brain region. However, if race can only be decoded from a particular brain region when participants are making one or the other type of judgment, this would suggest a preferential role of that brain region in maintaining or expressing either evaluative or stereotype-based representations, providing evidence of distinct brain systems mediating these two types of representation.

We also collected IAT measures of both evaluative bias and implicit stereotyping. In past research, these two measures were shown to index independent representations of intergroup social information, such that association strength scores on the two measures were uncorrelated with each other, and predicted unique behavioral outcomes (Amodio & Devine, 2006). In the present study, we tested whether scores on the behavioral measures of implicit racial evaluation and stereotyping would uniquely correlate with MVPA race decoding accuracy associated with friendship and trait judgments, respectively. This methodological design provides a highly specific test of our hypothesis while also establishing a meaningful connection between brain activity and behavior. Thus, we hypothesized that (a) race would be decoded in regions linked to visual processing independent of the type of conceptual associations activated for a particular judgment, but that (b) judgments based on evaluative and stereotype associations would recruit distinct patterns of activity in brain regions linked to evaluative processing and social cognition, and (c) these patterns would be uniquely associated with behavioral measures of implicit racial evaluation and stereotyping, respectively.

To test these hypotheses, it was critical to create an engaging and ecologically valid task that could be completed in the MRI scanner. To this end, we used an elaborate cover story. Participants were told that the study examined people's ability to infer information about others based solely on a picture of their face. Specifically, participants were told they would infer the types of activities a person might enjoy and whether a person is someone the participant might befriend. These two judgments were designed to rely on trait-related semantic versus affective processing, respectively. To bolster the cover story, participants completed questionnaires assessing their own preferences for various hobbies and interests and for the qualities that they value most in a potential friend. They were told that they would make judgments of other people who had completed the same set of questionnaires so that we could verify the accuracy of their inferences about each target person on these dimensions.

White American participants learned that, while in the scanner, they would see pairs of faces and would decide which of the two pictured individuals was more likely to (a) possess a particular trait or (b) be a friend, in a hypothetical circumstance (see Fig. 1). Participants were told that for the trait judgments, each participant would focus on just one particular trait. They were asked to select a piece of paper from a jar that indicated the activity they would judge. This choice was rigged so that every slip of paper indicated "athletic" as the trait. Athletic was used because it is a central African American stereotype that does not have strong evaluative associations, unlike many other stereotypes that hold negative value.

A critical feature of this design is that face pairs were always of the same race, such that participants always made trait or friendship judgments between two Black or two White faces. (Additionally, Asian faces were included in the stimulus set in order to aid the cover story, but were not analyzed.) This design precluded participants' concerns about showing explicit racial prejudice or the engagement of control in order to respond without prejudice. Thus, any patterns of activity distinguishing

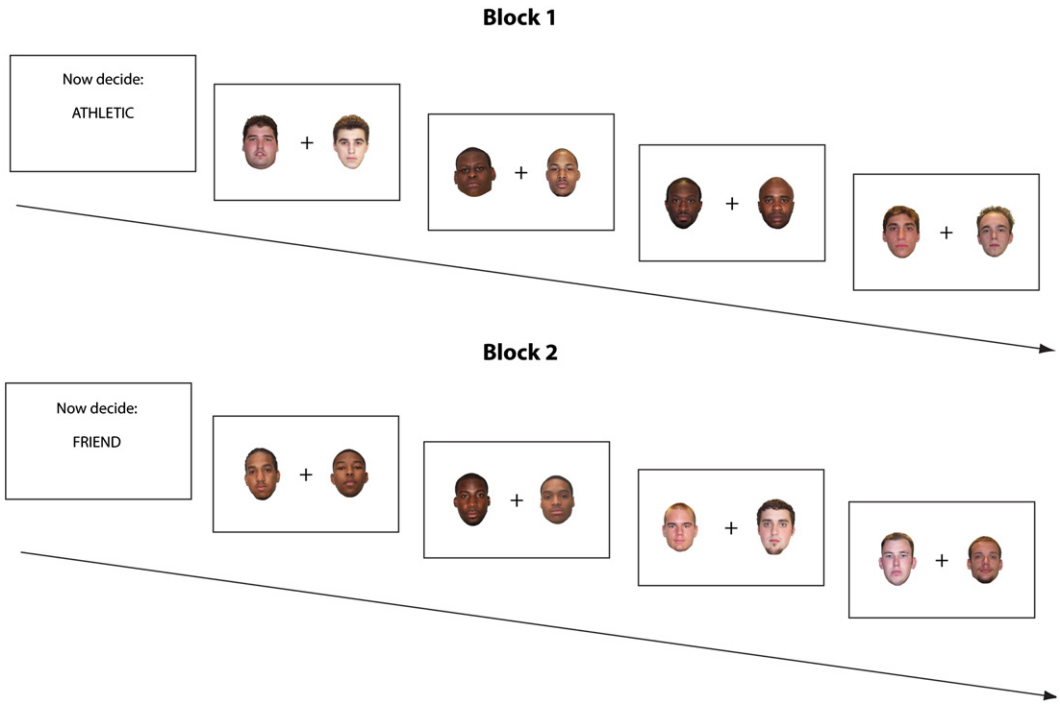

Fig. 1. Schematic illustration of the fMRI task. Face images in this illustration are adapted from the database of Minear and Park (2004).

judgments on Black versus White faces could not be attributed to explicit prejudice or self-regulatory efforts. The object of the experimental task was not to engender bias towards one or the other of the faces presented on a single trial, but rather to compare brain activity on separate trials in response to pairs of Black versus White faces. This comparison was made in two conditions designed to encourage participants to think about people either in terms of evaluations or trait judgments.

2. Methods

2.1. Participants and procedure

Twenty volunteers (14 male; age 18–22 years, mean 19) were recruited through flyers placed around the New York University campus. Participants completed the study individually in exchange for $40. All participants were right-handed, White, native-English speakers with no history of neurological illness. All experimental sessions were run by White experimenters. Upon arrival, participants provided informed consent and were screened for contraindications for MRI scanning. Participants were then told that the study consisted of two parts: a functional MRI scan and a set of computer tasks to be completed outside the scanner.

2.2. Scanner task

As described in Section 1, participants were told that their task was to judge pairs of faces according to two criteria: "Who is more likely to enjoy athletic activities?" and "Who would you be more likely to befriend?" Participants underwent two scans of approximately 11 min each. Each scan comprised 10 blocks, alternating between blocks of the two different types of judgments. At the start of each block, participants were presented with the cue "Now decide:", underneath which was presented an instruction (either "Friend" or "Athletic"). This instruction screen was presented for 3.5 s, followed by a blank screen for .5 s. There then followed 20 trials. On each trial a pair of faces were presented to the left and right of a fixation cross (see Fig. 1), for 2 s. There was then a blank screen for .5 s before the next pair of faces was presented. Participants could respond with a left or right keypress at any point within the 2.5 s of each trial. Within each block of 20 trials, 8 pairs of White faces, 8 pairs of Black faces, and 4 pairs of Asian faces were presented in randomized order. At the end of each block, participants were asked to report which judgment they had just been making, to ensure compliance with task instructions. They were presented with the cue "I was just deciding:" underneath which the words "Friend" and "Athletic" were presented on the left and right of the screen (in random positions). Participants had 4 s to respond to this question, after which a blank screen was presented for .5 s followed by the instruction for the next block. Following the completion of the task, a high-resolution structural image was collected. Note that the present design did not include any 'null' events. This is because every contrast of interest involved a differential effect between trials with White versus Black faces. In this case, inclusion of null events would have reduced the efficiency of our design. The most efficient design is to maximize the amount of time spent by participants engaged in the task, with the order of White and Black trials randomized (Henson, 2006).

2.3. Stimuli

Stimuli consisted of 80 photographs each of White and Black faces, and 40 photographs of Asian faces. Photographs were matched for perceived age (all young adults) and attractiveness based on ratings by two judges. All stimuli were cropped to present the face only (excluding the neck) and presented on a white background. Each photograph appeared a total of 4 times in the experiment, once for each crossing of judgment (athleticism/friendship) and position (left/right), every time paired with a different face.

2.4. Behavioral measures

Following the scanning session, participants were brought to another room to complete IAT measures of implicit stereotyping and evaluation, in an order that was counterbalanced across participants.

2.4.1. Evaluative IAT

The evaluative IAT assessed associations between images of Black vs. White faces and pleasant vs. unpleasant words, and consisted of seven blocks of trials. In Block 1 (20 trials), participants categorized a series of individual words as "pleasant" or "unpleasant" via left vs. right button press. The pleasant target words included *honor*, *lucky*, *diamond*, *loyal*, *freedom*, *rainbow*, *love*, *honest*, *peace*, and *heaven*. Unpleasant words included *evil*, *cancer*, *sickness*, *disaster*, *filth*, *vomit*, *bomb*, *rotten*, *abuse*, and *ugly*. These words were taken from Greenwald et al. (1998). Importantly, none of the target words were associated with stereotypes of African Americans, and thus responses could not be made on the basis of stereotypic associations. In Block 2 (20 trials), participants categorized a series of faces by race, indicating White or Black via left vs. right button press. In Blocks 3 (practice; 20 trials) and 4 (critical; 40 trials), the categories were combined, such that a series of words and faces appeared. Pleasant words and White faces were categorized with a left button-press, whereas unpleasant words and Black faces were categorized with a right button press. That is, these pairings were "compatible" with anti-Black/pro-White associations. In Block 5 (20 trials), participants categorized faces alone, but the mapping of response keys was reversed. In Blocks 6 (practice; 20 trials) and 7 (critical; 40 trials), judgments were again combined, but this time pleasant words and Black faces were assigned to one key, whereas unpleasant words and White faces were assigned to the other key. These pairings were "incompatible" with anti-Black/pro-White associations. Implicit evaluative bias on this task was indicated by longer response latencies on the incompatible blocks compared with compatible blocks. The order of compatible and incompatible blocks, and the assignment of responses to the right vs. left hand were independently counterbalanced.

2.4.2. Stereotyping IAT

The stereotyping IAT was designed to assess the associations of Black vs. White faces with two major dimensions of the African American stereotype: (un)intelligence and athleticism. Since the format of the IAT requires that words from the two categories map onto a single dimension, categories of "mental" and "physical" were used instead of intelligent and athletic, such that intelligence words were categorized as mental and athletic words were categorized as physical (Amodio & Devine, 2006). By using this categorical dimension, responses could only be completed based on semantic associations and not evaluative associations. Mental words included *math, brainy*, *aptitude*, *educated*, *scientist*, *smart*, *college*, *genius*, *book*, and *read*. Physical words included *athletic*, *boxing*, *basketball*, *run*, *agile*, *dance*, *jump*, *rhythmic*, *track*, and *football*. The procedure for the stereotyping IAT was identical to that of the evaluative IAT, except that the pleasant/unpleasant category labels and words were replaced with the mental/

physical labels and words. All face stimuli in the IAT task were novel, i.e. none had previously been presented in the fMRI task.

2.4.3. IAT scoring

Responses to the evaluative and stereotyping IATs were scored as the *D* statistic (Greenwald, Nosek, & Banaji, 2003). To exclude responses reflecting action slips or inattention, we included only correct responses with latencies between 300 and 2500 ms. *D* was quantified as the difference in mean response latency between incompatible and compatible blocks, divided by their pooled standard deviation. Higher *D* scores reflect stronger negative or stereotype-consistent associations.

2.5. fMRI acquisition and analysis

Brain images were acquired using a Siemens Allegra 3T head-only scanner in the Center for Brain Imaging at New York University. Functional runs comprised 319 EPI volumes, each consisting of 34 contiguous oblique-axial slices, acquired approximately parallel to the AC-PC line ($3 \times 3 \times 3$ mm^3 voxels; matrix: 80×64; $TR = 2000$ ms; $TE = 15$ ms). After the functional runs, a structural scan was acquired (T1-weighted MPRAGE: 256×256 matrix, $FOV = 256$ mm, 176 1 mm sagittal slices).

fMRI data were analyzed in SPM8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/) and custom-written code in Matlab 7.10. The volumes were realigned and corrected for different slice acquisition times (but not normalized or smoothed). Variance in the BOLD signal was decomposed with a set of regressors in a general linear model (Friston et al., 1995). Within each run, separate regressors coded for: (1) the instruction period preceding each block; (2) stimulus onset on trials to which participants failed to respond; (3) catch trials at the end of each block, requiring participants to report which task they had just been performing; (4) stimulus onset on athleticism judgment trials with pairs of Asian faces; and (5) stimulus onset on friendship judgment trials with Asian faces. These regressors were included to model variance related to events of no interest. Additionally, a separate pair of regressors coded for each of the 10 blocks of trials within each run (five athleticism judgments and five friendship judgments), with one regressor indexing the presentation of White faces in each block and one regressor indexing the presentation of Black faces. Thus there were 20 regressors of interest within each run, representing five blocks each of four conditions of interest: (a) Athleticism judgments, White faces; (b) Athleticism judgments, Black faces; (c) Friendship judgments, White faces; and (d) Friendship judgments, Black faces. Randomization of trial order allowed us to avoid excessive collinearity between these regressors of interest (mean unsigned correlation between regressors: $r = .14$). With the exception of the regressor representing instruction periods (modeled as a boxcar function, duration 3.5 s), all regressors were constructed as a delta function aligned to the onset of each event, convolved with a canonical hemodynamic response function. These regressors, together with regressors representing residual movement-related artifacts and the mean over scans, comprised the full model for each session. The data and model were high-pass filtered to a cutoff of 1/128 Hz.

Parameter estimates for each regressor were calculated from the least mean square fit of the model to the data. These parameter estimates were used as data for the multi-voxel pattern analysis (MVPA). A searchlight approach was used (Kriegeskorte, Goebel, & Bandettini, 2006), investigating decoding accuracy from a sphere of voxels centered on each voxel in the brain in turn. At each voxel, a spherical region of interest (ROI) was generated (radius: three voxels). Parameter estimates for each voxel within the ROI were extracted, separately for each regressor. This yielded a total of 40 vectors, each representing a single run of trials in one of the four conditions of interest. Each vector was normalized to mean 0, standard deviation 1, so that decoding accuracy was based on the distribution of activation over voxels rather than differences in mean level of activation. Care was taken to ensure that only voxels within the brain contributed to MVPA results. All first-level models, yielding the parameter estimates that were used as data for MVPA, employed the SPM default implicit masking option, excluding voxels with mean signal of less than 80% of the global intensity. The resulting masks were individually inspected to further exclude any remaining voxels that fell outside the brain (as indicated by the co-registered structural scan), to ensure that only within-brain voxels contributed to MVPA results. Note also that first-level models included movement parameters as nuisance regressors, in order to guard against the impact of task-related motion on parameter estimates.

Separate analyses were conducted for athleticism and friendship judgments. First, a linear support vector machine (SVM) was trained on the data from the first scanning run, attempting to discriminate vectors representing activity engendered by White face-pairs versus Black face-pairs (LIBSVM implementation, http://www.csie.ntu.edu.tw/~cjlin/libsvm/; regularization parameter *C*: 1). The SVM was then tested on data from the second scanning run, attempting to predict whether each vector represented White face-pair or Black face-pair trials. This process was repeated after flipping the training and testing sets, and the mean classification accuracy was recorded, after subtracting 50 so that zero indicated chance performance. There were two resulting maps, indicating at each voxel the White–Black classification accuracy during athleticism judgments and friendship judgments. These maps were normalized into Montreal Neurological Institute (MNI) space using 3 mm cubic voxels and 4th-degree B-spline interpolation, and smoothed with an isotropic 4-mm full-width half-maximum Gaussian kernel. This relatively small kernel size was used to avoid excessive smoothing, seeing as the searchlight analysis already imposes spatial smoothing on the data.

To assess results at the group level, a random effects analysis was conducted as follows. A pair of images from each participant was entered into a repeated-measures ANOVA (within-subject factor: athleticism vs. friendship judgment). Two covariates were additionally included in the ANOVA, representing *D* scores from the evaluative and stereotyping IAT measures. This allowed investigation both of brain regions in which decoding accuracy differs significantly from chance, and also brain regions in which the decoding accuracy was significantly associated with behavioral IAT measures. Results were assessed at an uncorrected threshold of $p < .005$, in conjunction with an extent threshold determined by SPM8 to yield a family-wise-error corrected whole-brain threshold of $p < .05$.

3. Results

Four participants were excluded due to excessive movement (three participants) or poor task compliance (falling asleep during the experiment; one participant), yielding a final sample of 16 participants.

3.1. Behavioral results: IAT

Both evaluative and stereotyping IAT D scores were significantly greater than zero, indicating strong preferences for White over Black people (evaluative IAT: $D = .53$, $t(15) = 8.1$, $p < .001$) and strong association of Blacks with athleticism and unintelligence, relative to Whites (stereotyping IAT: $D = .37$, $t(15) = 8.0$,

$p < .001$). Scores on these two IATs were uncorrelated ($r = .12$, $p = .66$), consistent with previous research suggesting that they reflect different underlying representations of evaluation and semantic content (Amodio & Devine, 2006).

3.2. Behavioral results: scanner task

Because task judgments were subjective, behavioral analyses focused on reaction times (RTs) and not accuracy. Mean RTs were as follows: Trait judgment, White: 1275 ms; Trait judgment, Black: 1290 ms; Friendship judgment, White: 1249 ms; Friendship judgment, Black: 1294 ms. These results were examined in a repeated-measures ANOVA with factors Judgment (trait/friendship) and Race (White/Black). There was no significant main effect of Judgment ($F(1,15) = 1.83$, $p = .2$) but the main effect of Race was significant ($F(1,15) = 8.0$, $p = .01$), qualified by a significant Judgment × Race interaction ($F(1,15) = 5.0$, $p = .04$). This interaction was driven by a significant difference in response time to White versus Black faces during friendship judgments ($t(15) = 4.0$, $p = .001$) but not athleticism judgments ($t(15) = 1.1$, $p = .29$). None of the neuroimaging results, reported below, were significantly correlated with RTs on the judgment task used in the scanner (i.e. variation between participants in neuroimaging data was not significantly correlated with variation in the relevant behavioral data in any analysis). Accuracy of catch trials, on which participants indicated the judgment they had just been making, was 97%. This indicates that participants judged faces on the correct criterion even though this criterion was described at the beginning of each block but not during individual trials.

Table 1
Regions from which it was possible to decode whether participants were viewing White versus Black faces. PFC = prefrontal cortex. BA = Brodmann Area.

Region	BA	Coordinate	Z_{max}	N voxels
Collapsed over Trait and Friendship judgments[1]				
Medial occipital cortex	17/18	15, −91, 10	3.72	46
		−12, −82, −5	4.11	148
Superior parietal cortex	7	−42, −79, 49	4.02	147
Lateral occipital cortex	19	−36, −79, −20	3.28	68
		30, −64, −20	3.56	64
Lateral temporal cortex	20/21	63, −40, −14	4.14	110
Inferior temporal cortex	20/36	24, −4, −38	3.71	42
Lateral frontal cortex	44	51, 17, 46	3.79	91
Medial frontal cortex	8	−6, 29, 55	4.01	77
Trait > Friendship[2]				
Medial rostral PFC	10	−3, 65, 22	3.73	79
Friendship > Trait[3]				
Lateral orbitofrontal cortex	11/47	−36, 50, −17	3.27	42

Minimum cluster size by whole-brain corrected extent threshold: [1]42; [2]44; [3]42.

3.3. Neuroimaging results

The first question we addressed in our analysis was whether race (White vs. Black) of face stimuli could be decoded with accuracy significantly above chance levels, collapsing over the two judgment types (Table 1 and Fig. 2). This analysis revealed significant effects in widespread regions of occipital cortex, superior parietal cortex, posterior frontal cortex and, most significantly, lateral temporal cortex. Thus, widespread, predominantly occipital brain regions exhibited patterns of activation that

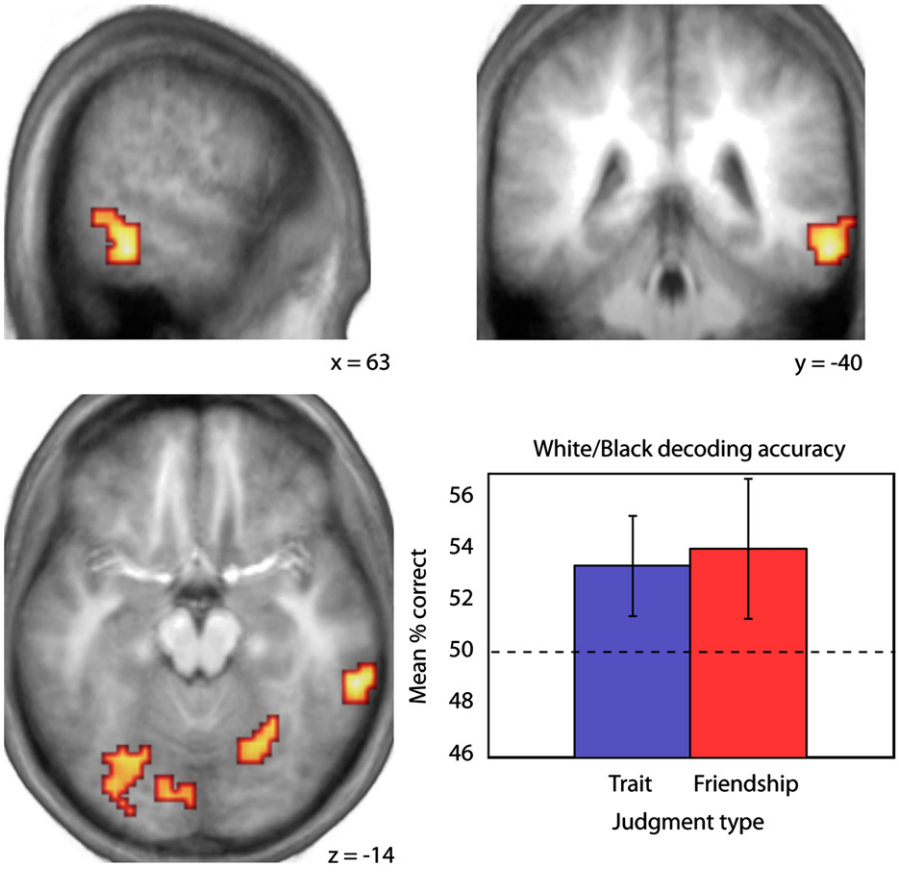

Fig. 2. Regions from which it was possible to decode whether participants were viewing White or Black faces, collapsed over the Trait and Friendship judgments. Results are plotted on the mean structural scan. Bar chart displays decoding accuracy from the two tasks in the peak region of lateral temporal cortex. Error bars represent 95% confidence intervals.

differed between trials on which White versus Black faces were presented.

The second question we addressed was whether we could accurately decode race from particular brain regions as a function of the evaluative vs. semantic judgment type. In these contrasts, decoding accuracy for one judgment type was subtracted from accuracy for the other judgment type, excluding voxels showing below-chance decoding accuracy in the nonpreferred condition ($p < .05$). This prevented the possibility that significant differences between the two judgment types could be driven by significantly below-chance decoding accuracy in the nonpreferred condition, which would be difficult to interpret. There was a single region of anterior medial PFC in which it was possible to decode during trait judgments whether a White or Black face pair was presented ($t(15)=4.63$, $p=.0003$) but not during friendship judgments ($t(15)=.81$, $p=.43$; peak co-ordinate for direct comparison between decoding accuracy in the two judgments: −3, 65, 22, $Z_{max}=3.73$, extent: 79 voxels). The reverse contrast revealed a single region of lateral orbitofrontal cortex, where it was possible to decode whether a White or Black face pair was presented during friendship judgments ($t(15)=4.1$, $p=.001$) but not trait judgments ($t(15)=.79$, $p=.44$; peak co-ordinate for direct comparison between decoding accuracy in the two judgments: −36, 50, −17, $Z_{max}=3.27$, extent: 42 voxels). Thus in both regions decoding accuracy was significantly above chance in the preferred condition, but not significantly different from chance in the nonpreferred condition. Mean decoding accuracy was extracted from each of these peak co-ordinates, and illustrated in Fig. 3.

3.4. Representation of implicit evaluation and stereotyping

Finally, we tested whether participants' implicit representations of value and stereotype knowledge associated with Black (vs. White) people, assessed behaviorally, were linked to these neuroimaging results. To this end, we examined relationships between decoding accuracy and IAT indices of evaluative bias and implicit stereotyping (Table 2 and Fig. 4). First we searched for brain regions where White–Black decoding accuracy during trait judgments (designed to encourage the expression of implicit stereotypes) was positively related with stereotyping IAT scores, using the same statistical threshold as the earlier analyses. This revealed significant effects in lateral temporal cortex, precentral sulcus, medial frontal cortex and, most significantly, left temporal pole. Next we searched for brain regions where White–Black decoding accuracy during friendship judgments (designed to encourage the

Table 2
Regions in which there was a significant relationship between White–Black decoding accuracy and post-scan behavioral IAT scores. The interaction effect indicates regions in which the correlation with stereotyping versus evaluative IAT differed as a function of trait judgments versus friendship judgments. BA = Brodmann Area.

Region	*BA*	*Coordinate*	Z_{max}	*N voxels*
Correlation between Trait decoding accuracy and stereotyping IAT[1]				
Temporal pole	38	−30, 26, −35	4.06	84
Occipito-temporal cortex	37	−42, −49, 1	3.89	56
Lateral temporal cortex	22/42	57, −31, 16	3.55	93
Superior frontal cortex	8	−15, 35, 61	3.90	72
Premotor cortex	6	48, −19, 67	4.02	184
Correlation between Friendship decoding accuracy and evaluative IAT[2]				
Temporal pole	38	−27, 26, −29	4.54	43
Medial occipital cortex	17	−3, −106, 13	4.86	65
Occipito-parietal cortex	7/19	18, −70, 28	3.45	48
Interaction[3]				
Temporal pole	38	−27, 26, −38	4.31	58

Minimum cluster size required by whole-brain corrected extent threshold: [1]56; [2]43; [3]42.

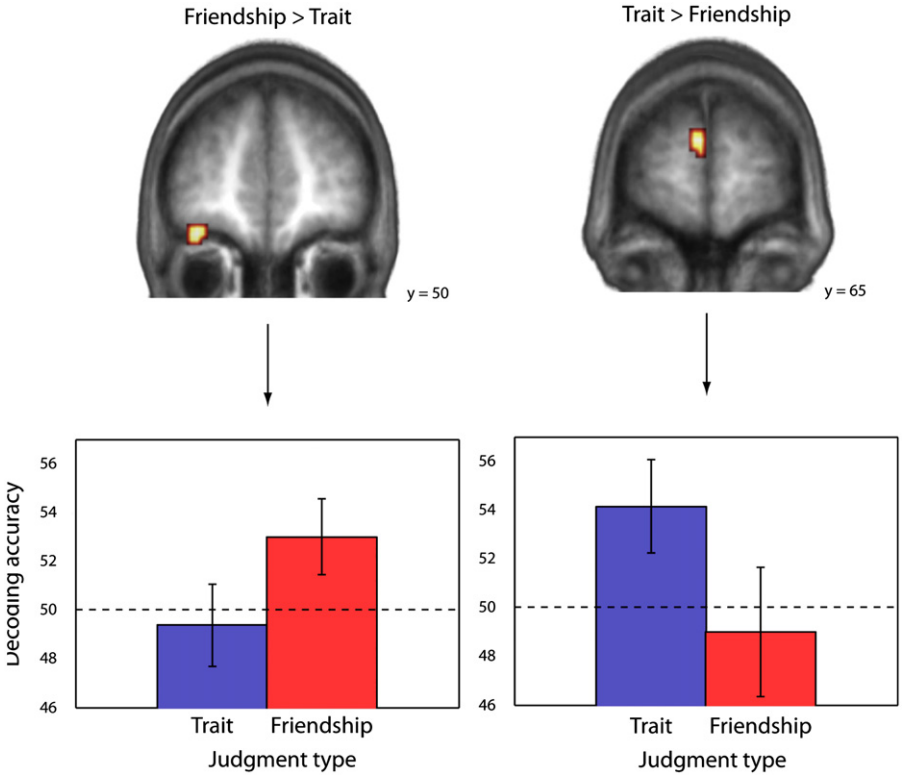

Fig. 3. Regions in which White/Black decoding accuracy was significantly different between Trait and Friendship judgments. Results are plotted on the mean structural scan. Bar chart displays decoding accuracy in the peak coordinates. Error bars represent 95% confidence intervals.

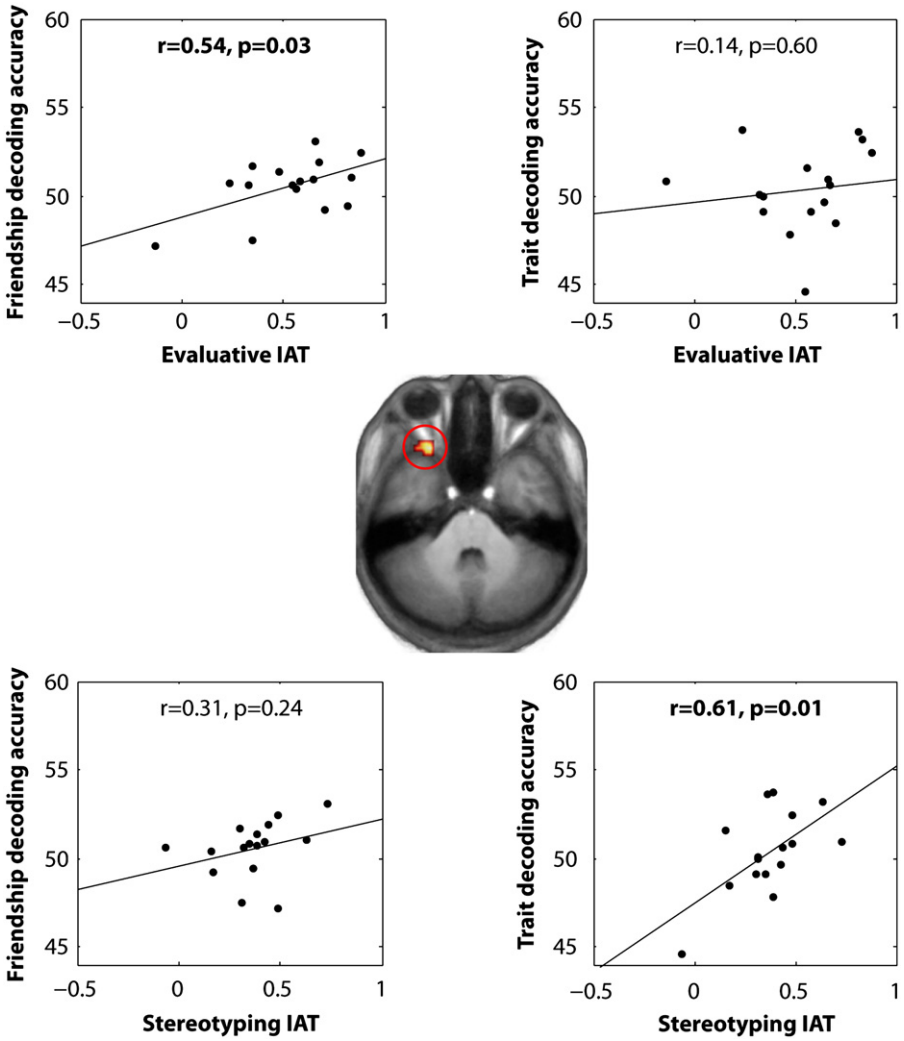

Fig. 4. Scatter plots indicating the relationship between White/Black decoding accuracy in left temporal pole (−27, 26, −38) for the two judgments (Friendship/Trait) and the two IAT scores (evaluative/stereotyping). An axial slice of the mean structural scan is also shown ($z=-38$), displaying results of the interaction analysis searching for brain regions with differential relationships between decoding accuracy and IAT scores for Friendship versus Trait judgments. The significant effect in temporal pole is highlighted. Note that the scatter plots are presented to illustrate this significant effect (which was subject to a whole-brain corrected threshold), rather than being based on an independently-defined region of interest.

expression of evaluative bias) was positively related with evaluative IAT scores. This revealed significant effects in medial occipital cortex, occipito-parietal cortex and, again, left temporal pole. Strikingly, the relationship between decoding accuracy in left temporal pole and IAT scores depended on the judgment context. Decoding accuracy in left temporal pole was associated with evaluative IAT scores during friendship judgments but with stereotyping IAT scores during athleticism judgments. This dissociation pattern is consistent with our finding that the two IAT measures were themselves uncorrelated (as in previous research, e.g., Amodio & Devine, 2006). To probe this finding further, we computed the interaction effect, i.e. searching for brain regions in which the difference in the relationship with evaluative versus stereotyping IAT was dependent on task (trait or friendship judgment). For this interaction analysis, regions were only considered significant when both of the individual expected correlations (i.e. trait decoding-stereotyping IAT; friendship-decoding-evaluative IAT) were significant ($p<.05$) at the peak voxel. This revealed a significant effect in left temporal pole, as predicted, but no other regions. The analogous analysis of the opposite interaction effect (i.e. regions with higher correlations for trait decoding-evaluative IAT and friendship decoding-stereotyping IAT) did not reveal any significant effects.

In order to visualize these effects, Fig. 4 displays scatter plots and correlation coefficients comparing (1) left temporal pole decoding accuracy during friendship judgments (using the peak coordinate from the interaction analysis); (2) decoding accuracy during trait judgments; (3) evaluative IAT scores; and (4) stereotyping IAT scores. These analyses revealed a highly specific dissociation pattern of results, whereby decoding accuracy for friendship judgments was predicted only by evaluative IAT scores, whereas decoding accuracy for trait judgments was predicted only by stereotyping IAT scores.

To investigate these results further, participants were divided into two groups on the basis of a median split of the evaluative or stereotyping IAT scores. Because participants' scores on both IATs were predominantly positive in value, these median splits represented high vs. low degrees of bias favoring Whites. Because these comparisons tested specific directional hypotheses one-tailed tests were conducted. Participants with relatively strong implicit preference for White people over Black people (i.e., high IAT scores) exhibited patterns of activity in left temporal pole that significantly distinguished White versus Black faces during friendship judgments (decoding accuracy: 51.10%; $t(7)=2.31$; $p=.03$). However, decoding accuracy among participants with relatively lesser

ingroup preference (low evaluative IAT scores) did not differ from chance (decoding accuracy: 49.99%; $t(7)=.02$, $p=.40$). Similarly, participants with relatively stronger stereotype associations exhibited a trend towards significant decoding of White versus Black from left temporal pole during trait judgments (decoding accuracy: 51.14%; $t(7)=1.69$, $p=.07$). However decoding accuracy in those with low stereotyping IAT scores was not significantly different from chance (decoding accuracy: 49.36%, $t(7)=.50$, $p=.32$). These results show that the interaction effect plotted in Fig. 4 is associated with significantly above-chance decoding in participants with above-average IAT scores, rather than significantly below-chance decoding in participants with below-average IAT scores.

In order to compare our MVPA results with standard univariate analysis techniques, we conducted analogous univariate analyses for each of the MVPA tests reported above (i.e. investigating differential mean signal change to White versus Black trials, rather than accuracy of decoding White versus Black trials). These tests did not produce any significant activations, consistent with the view that results from MVPA and univariate approaches need not mirror each other (see Gilbert et al., 2012, for further discussion).

4. Discussion

Knowledge about a person's social group, such as trait attributes or a global evaluation, can have a profound influence on how we perceive and act toward that person. An understanding of the neural structures that represent this knowledge is crucial to theories of how this knowledge is acquired, activated, and expressed. In this study, we used multi-voxel pattern analysis (MVPA) to investigate regional brain activity that distinguished presentation of White versus Black faces during two different judgment conditions encouraging the expression of evaluative and stereotyping race bias (friendship and trait judgments, respectively). Across both conditions, significant decoding was possible from an extensive, predominantly occipital network of brain regions. Furthermore, lateral orbitofrontal and anterior medial prefrontal regions showed preferential decoding effects during the friendship and trait judgments respectively. In addition, decoding accuracy in left temporal pole correlated with a behavioral measure of evaluative bias during the friendship task and with an independent behavioral measure of stereotyping bias during the athleticism task, even though these two behavioral measures did not correlate with each other.

These findings corroborate previous behavioral results showing that evaluative bias and implicit stereotyping towards members of an outgroup may be dissociable (Amodio & Devine, 2006), suggesting that they may operate independently of each other and be mediated by distinct brain systems. The present study provides neuroimaging evidence that some brain regions represent race in a relatively invariant manner, whereas others contain representations that are extremely sensitive to context.

4.1. A multi-componential representation of race and implicit associations

Perhaps unsurprisingly, occipital regions involved in relatively low-level visual processing tended to show invariant responses to race, i.e. similar decoding accuracy during the two tasks, uncorrelated with behavioral measures of evaluative bias or implicit stereotyping. These areas included early occipital cortical regions, and also higher-level visual processing regions such as fusiform gyrus and inferior temporal cortex, thought to play a critical role in representation of face characteristics (Ishai, 2008). Additionally, relatively invariant representations were seen in the posterior frontal lobes.

Other brain regions showed context-specific patterns of race-related activity. In lateral orbitofrontal and anterior medial PFC, regions were found that preferentially represented race during the friendship and trait judgment tasks, respectively. This suggests that these distinct brain regions may play a role in maintaining and/or expressing evaluative or semantic aspects of social information, rather than their depending on a single underlying cognitive mechanism. The association between patterns of activity in medial anterior PFC and implicit stereotyping is consistent with several lines of evidence from neuroimaging, neuropsychology, and transcranial magnetic stimulation (TMS). In an fMRI study, Knutson, Mah, Manly, and Grafman (2007) reported increased anterior medial PFC activity during stereotype-consistent versus stereotype-inconsistent conditions of a race and gender IAT. However, it is hard to rule out the possibility that this result reflected the absence of conflict (of any type) rather than being tied specifically to stereotyping (cf. Amodio, Devine, & Harmon-Jones, 2008; Amodio et al., 2004; Beer et al., 2008). More convincing evidence comes from Quadflieg et al. (2009), who found that medial anterior PFC activity was enhanced during expression of stereotypic associations between an activity (e.g. mowing the lawn) and gender, compared with stereotypic associations between an activity and a place (indoor versus outdoor). Further evidence linking medial anterior PFC and stereotyping comes from Saxe and Wexler (2005), who showed that medial anterior PFC activity tended to increase when a person was described as a foreigner before any specific mental content was ascribed. Such evidence has led some authors to conceptualize anterior medial PFC as a repository of social knowledge (Krueger et al., 2009) or, alternatively, a region important for the coordination of knowledge with complex social judgments, goals, and behaviors (Amodio & Frith, 2006; Contreras, Banaji, & Mitchell, in press).

Whereas medial anterior PFC activity has been linked with stereotype-consistent IAT responses (Knutson et al., 2007), lateral PFC activity has been linked with stereotype-inconsistent responses in various evaluative IAT tasks (Chee, Sriram, Soon, & Lee, 2000; Knutson et al., 2007; Luo et al., 2006)—a pattern that may reflect response-inhibition processes that are engaged on such trials (Aron, Robbins, & Poldrack, 2004). As noted above, however, it is hard to rule out the possibility that these lateral PFC activations relate to general conflict-resolution processes, rather than being specific to the activation of evaluative bias (Amodio, 2010; Amodio, Devine, & Harmon-Jones, 2007; Beer et al., 2008). Further evidence on the roles of lateral PFC and anterior medial PFC mediating implicit attitudes comes from the neuropsychological study of Gozzi et al. (2009). Whereas lesions to ventromedial PFC and anterior temporal lobe were associated with stronger bias in gender IAT performance, lesions to ventrolateral and orbitofrontal regions of PFC were associated with weaker bias (see also Milne & Grafman, 2001). Importantly, these patterns of results were not observed in a control IAT task. Seeing as the IAT used by Gozzi et al. (2009) paired male and female names with the concepts of strength versus weakness, which have both semantic and evaluative components, it is difficult to relate their findings directly to the present results. Nevertheless, the results of Gozzi et al. (2009) are consistent with the current finding of separable roles of medial anterior PFC and lateral orbitofrontal PFC in social judgment.

In the present study, lateral orbitofrontal cortex was linked with evaluative rather than stereotyping bias. This fits well with results from human neuroimaging (e.g. Sescousse, Redoute, & Dreher, 2010) and primate electrophysiology (e.g. Tremblay & Schultz, 1999) implicating orbitofrontal cortex in

the representation of reward value, particularly in the context of associative learning (Rushworth et al., 2011; Walton, Behrens, Buckley, Rudebeck, & Rushworth, 2010). Furthermore, face-selective cells in macaque orbitofrontal cortex have been reported, some of which represent face identity (Rolls, Critchley, Browning, & Inoue, 2006). The orbitofrontal cortex is thus well placed to integrate perceptual representations of identity with evaluative reward representations.

In addition to regions within PFC, left temporal pole stood out as a region exhibiting a highly context-sensitive effect. During friendship judgments, this region represented race insofar as participants exhibited high levels of evaluative bias on an independent behavioral measure. However, decoding accuracy in this brain region did not correlate with a behavioral measure of implicit stereotyping. During trait judgments, the pattern of results in left temporal pole reversed: decoding accuracy now correlated with a behavioral measure of implicit stereotyping but not evaluative bias. This pattern of results is consistent with recent theoretical accounts of anterior temporal lobe as playing a critical role in linking highly-processed perceptual inputs with social representations and emotional responses (Olson et al., 2007; Zahn et al., 2007). For example, Damasio, Tranel, and Damasio (1990) report that anterior temporal lobe lesions can give rise to a form of prosopagnosia whereby perception of faces is preserved while recognition is impaired. This deficit extends to non-face recognition cues such as voice and gait, suggesting a role in person-perception rather than simply face-recognition. We propose that engaging in the friendship and trait judgment tasks may have elicited left temporal pole representations of evaluation and stereotype content respectively—two primary and distinct forms of person knowledge. This interpretation is consistent with findings across previous neuroimaging studies of race in which patterns of brain activity vary depending on the behavioral task (Lieberman et al., 2005; Wheeler & Fiske, 2005). Furthermore, the effects in temporal pole may be particularly related to individual differences.

4.2. Implications for theories of implicit social cognition and intergroup bias

These results shed new light on the cognitive mechanisms underlying conceptual aspects of racial evaluations and stereotypes. Most importantly, the distinct patterns of activity for evaluation- and stereotype-based judgments observed in the temporal pole suggest that the content of racial evaluations and stereotypes may be represented in distinct semantic networks. This finding is consistent with previous behavioral research showing that the activation of evaluative associations may be dissociated from the activation of stereotype associations (e.g. Amodio & Devine, 2006). Our results corroborate the recent theoretical proposal that evaluative and stereotypic information may be learned, stored, and unlearned via different networks of information (Amodio & Ratner, 2011). Furthermore, a consideration of these distinctions is critical when designing interventions to change social attitudes or stereotypes. That is, an intervention to change evaluations might have little or no effect on the stereotype content of the group, and vice versa. Therefore, complementary interventions, which each target their respective associations, may be most effective in changing judgments and behavior toward a particular social group. This finding has application beyond the domain of intergroup relations; indeed, it may apply equally to efforts to change judgments and behavior toward any social or non-social object, such as a political candidate, consumer product, or one's own self-image.

It is notable that the present research focused on conceptual (or cognitive) aspects of implicit racial evaluation, whereas previous neuroimaging research has focused on affective components. Both are valid components of evaluation. Indeed, classic research in social psychology posits a tripartite model of attitudes, whereby an attitude comprises a combination of cognitive (i.e., conceptual), affective, and behavioral components (Breckler, 1984; McGuire, 1969). Contemporary measures of implicit racial associations, such as the IAT and others, are designed to pick up on conceptual aspects of either an evaluation (e.g., associations with the concepts of "good" or "bad") or stereotype content (Fazio et al., 1995; Greenwald et al., 1998), but they are less well suited for the measurement of affective processes (Amodio et al., 2003; Amodio & Mendoza, 2010). However, the extant neuroimaging research on implicit racial bias has focused almost exclusively on neural responses linked to affect, such as amygdala activity. Although the conceptual and affective components of implicit racial attitudes are likely related, new information about the conceptual processes, and their representation in the brain, will be particularly useful for understanding the forms of implicit evaluation assessed by behavioral tasks. Our findings also help to distinguish between implicit evaluation and implicit stereotyping as two different types of conceptual representations that contribute to intergroup judgments and behaviors. An important goal for future research will be to elucidate the relative contributions and interplay of these two aspects of implicit evaluation, as they likely relate to different mechanisms of learning, memory, and behavior. The present findings provide a foundation for examining this interplay in the brain and behavior. It is also important to investigate how far the present results generalize beyond the relatively homogenous college student sample studied here.

4.3. Using MVPA to probe neural representations of social cognition

In addition to providing evidence on the representation of race in the human brain, the present results more generally support the use of MVPA as a valuable technique for social cognitive neuroscience. A crucial aim of social cognition and social cognitive neuroscience research is to understand the underlying representations of social entities (e.g., individuals, groups) that influence social behavior. MVPA is particularly well suited to this aim, given its focus on representational content rather than the engagement of particular cognitive processes (Gilbert et al., 2012; Norman et al., 2006). For example, whereas univariate techniques are helpful for understanding the types of processes supported by particular brain regions (e.g. perception of line orientation in V1, or color perception in V4), MVPA may be used to decode the underlying representations within these regions (e.g. a bar oriented at 90°, from V1 activity; Kamitani & Tong, 2005). Of course, the underlying neural activity investigated by these two approaches may be identical; in that sense the distinction between process and representation may prove to be somewhat artificial. However, process-based and representation-based models currently provide distinct yet complementary approaches for linking underlying brain activity with higher-level cognitive theory, at different levels of description (Marr, 1982). In this sense, representation-based approaches, using MVPA, may constitute a valuable new technique for social cognitive neuroscience. Indeed, despite significant MVPA results in the present study, analogous univariate analyses failed to produce significant effects.

5. Conclusion

The present results suggest that the representation of race is multi-componential and potentially mutable. Our findings suggest that two distinct aspects of race bias—implicit stereotyping and

implicit evaluation—are mediated by distinct brain mechanisms. By learning more about the way in which different aspects of bias are represented within the brain, and potentially expressed via distinct brain pathways, this raises the hope of developing more sophisticated and effective interventions by which their unintended and harmful effects in society may be mitigated.

Acknowledgments

This research was supported by a grant from the National Science Foundation to D.M.A. (BCS 0847350). S.J.G. was supported by a Royal Society University Research Fellowship and a Royal Society International Travel Grant.

References

Allport, G. W. (1954). *The nature of prejudice*. Reading, MA: Addison-Wesley.
Amodio, D. M. (2008). The social neuroscience of intergroup relations. *European Review of Social Psychology*, *19*, 1–54.
Amodio, D. M. (2010). Coordinated roles of motivation and perception in the regulation of intergroup responses: Frontal cortical asymmetry effects on the P2 event-related potential and behavior. *Journal of Cognitive Neuroscience*, *22*, 2609–2617.
Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology*, *91*, 652–661.
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., Devine, P. G., & Harmon-Jones, E. (2008). Individual differences in the regulation of intergroup bias: The role of conflict monitoring and neural signals for control. *Journal of Personality and Social Psychology*, *94*, 60–74.
Amodio, D. M., & Frith, C. D. (2006). *Meeting of minds: The medial frontal cortex and social cognition*, *7*, 268–277*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., & Mendoza, S. A. (2010). Implicit intergroup bias: Cognitive, affective, and motivational underpinnings. In: B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition* (pp. 353–374). New York: Guilford.
Amodio, D. M., & Ratner, K. G. (2011). A memory systems model of implicit social cognition. *Current Directions in Psychological Science*, *20*, 143–148.
Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences*, *8*, 170–177.
Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype-activation on action. *Journal of Personality and Social Psychology*, *71*, 230–244.
Beer, J. S., Stallen, M., Lombardo, M. V., Gonsalkorale, K., Cunningham, W. A., & Sherman, J. W. (2008). The Quadruple Process model approach to examining the neural underpinnings of prejudice. *Neuroimage*, *43*, 775–783.
Breckler, S. J. (1984). Empirical validation of affect, behavior, and cognition as distinct components of attitude. *Journal of Personality and Social Psychology*, *47*, 1191–1205.
Chee, M. W., Sriram, N., Soon, C. S., & Lee, K. M. (2000). Dorsolateral prefrontal cortex and the implicit association of concepts and attributes. *Neuroreport*, *11*, 135–140.
Contreras, J.M., Banaji, M.R., & Mitchell, J.P. Dissociable neural correlates of stereotypes and other forms of semantic knowledge. *Social Cognitive and Affective Neuroscience*, in press.
Cunningham, W. A., Johnson, M. K., Raye, C. L., Chris Gatenby, J., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of black and white faces. *Psychological Science*, *15*, 806–813.
Damasio, A. R., Tranel, D., & Damasio, H. (1990). Face agnosia and the neural substrates of memory. *Annual Review of Neuroscience*, *13*, 89–109.
Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*, 20.
Devine, P. G. (1989). Prejudice and stereotypes: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*, 5–18.
Devine, P. G., & Elliot, A. J. (1995). Are racial stereotypes really fading? The Princeton trilogy revisited. *Personality and Social Psychology Bulletin*, *11*, 1139–1150.
Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*, 62–68.
Eberhardt, J. L. (2005). Imaging race. *American Psychologist*, *60*, 181–190.
Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline?. *Journal of Personality and Social Psychology*, *69*, 1013–1027.
Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In: D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology*, 2 (pp. 357–411). New York: McGraw-Hill.
Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. B., Frith, C. D., & Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, *2*, 189–210.
Gilbert, S. J. (2011). Decoding the content of delayed intentions. *Journal of Neuroscience*, *31*, 2888–2894.
Gilbert, S. J., Armbruster, D. J., & Panagiotidi, M. (2012). Similarity between brain activity at encoding and retrieval predicts successful realization of delayed intentions. *Journal of Cognitive Neuroscience*, *24*, 93–105.
Gilbert, S. J., Meuwese, J. D., Towgood, K. J., Frith, C. D., & Burgess, P. W. (2009). Abnormal functional specialization within medial prefrontal cortex in high-functioning autism: A multi-voxel similarity analysis. *Brain*, *132*, 869–878.
Golby, A. J., Gabrieli, J. D., Chiao, J. Y., & Eberhardt, J. L. (2001). Differential responses in the fusiform region to same-race and other-race faces. *Nature Neuroscience*, *4*, 845–850.
Gozzi, M., Raymont, V., Solomon, J., Koenigs, M., & Grafman, J. (2009). Dissociable effects of prefrontal and anterior temporal cortical lesions on stereotypical gender attitudes. *Neuropsychologia*, *47*, 2125–2132.
Grabenhorst, F., & Rolls, E. T. (2008). Selective attention to affective value alters how the brain processes taste stimuli. *European Journal of Neuroscience*, *27*, 723–729.
Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.
Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*, 197–216.
Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*, 523–534.
Haynes, J. D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Current Biology*, *17*, 323–328.
Henson, R. N. (2006). Efficient experimental design for fMRI. In: K. Friston, J. Ashburner, S. Kiebel, T. Nichols, & W. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (pp. 193–210). London: Elsevier.
Ishai, A. (2008). Let's face it: It's a cortical network. *Neuroimage*, *40*, 415–419.
Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Reviews Neuroscience*, *8*, 679–685.
Knutson, K. M., Mah, L., Manly, C. F., & Grafman, J. (2007). Neural correlates of automatic beliefs about gender and race. *Human Brain Mapping*, *28*, 915–930.
Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 3863–3868.
Krueger, F., Barbey, A. K., & Grafman, J. (2009). The medial prefrontal cortex mediates social event knowledge. *Trends in Cognitive Sciences*, *13*, 103–109.
Lieberman, M. D., Hariri, A., Jarcho, J. M., Eisenberger, N. I., & Bookheimer, S. Y. (2005). An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. *Nature Neuroscience*, *8*, 720–722.
Luo, Q., Nakic, M., Wheatley, T., Richell, R., Martin, A., & Blair, R. J. (2006). The neural basis of implicit moral attitude—An IAT study using event-related fMRI. *Neuroimage*, *30*, 1449–1457.
Marr, D. (1982). *Vision. A computation into the human representation and processing of visual information*. New York: W.H. Freeman.
McGuire, W. J. (1969). Attitude and attitude change. In: G. Lindzey, & E. Aronson (Eds.), *Handbook of social psychology* (pp. 136–314). Reading, MA: Addison-Wesley.
Milne, E., & Grafman, J. (2001). Ventromedial prefrontal cortex lesions in humans eliminate implicit gender stereotyping. *Journal of Neuroscience*, *21*, RC150.
Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, *36*, 630–633.
Natu, V., Raboy, D., & O'Toole, A. J. (2011). Neural correlates of own- and other-race face perception: Spatial and temporal response differences. *Neuroimage*, *54*, 2547–2555.
Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*, 424–430.
Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In: J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior* (pp. 265–292). Psychology Press.
Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007). The Enigmatic temporal pole: A review of findings on social and emotional processing. *Brain*, *130*, 1718–1731.
Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., et al. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience*, *12*, 729–738.
Quadflieg, S., Turk, D. J., Waiter, G. D., Mitchell, J. P., Jenkins, A. C., & Macrae, C. N. (2009). Exploring the neural correlates of social stereotyping. *Journal of Cognitive Neuroscience*, *21*, 1560–1570.
Ratner, K.G., Kaul, C., & Van Bavel, J.J. Is race erased? Decoding race from patterns of neural activity when skin color is not diagnostic of group boundaries. *Social Cognitive and Affective Neuroscience*, in press.

Rolls, E. T., Critchley, H. D., Browning, A. S., & Inoue, K. (2006). Face-selective and auditory neurons in the primate orbitofrontal cortex. *Experimental Brain Research*, *170*, 74–87.

Rushworth, M. F., Noonan, M. P., Boorman, E. D., Walton, M. E., & Behrens, T. E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron*, *70*, 1054–1069.

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia*, *43*, 1391–1399.

Sescousse, G., Redoute, J., & Dreher, J. C. (2010). The architecture of reward value coding in the human orbitofrontal cortex. *Journal of Neuroscience*, *30*, 13095–13104.

Tremblay, L., & Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature*, *398*, 704–708.

Walton, M. E., Behrens, T. E., Buckley, M. J., Rudebeck, P. H., & Rushworth, M. F. (2010). Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. *Neuron*, *65*, 927–939.

Wheeler, M. E., & Fiske, S. T. (2005). Controlling racial prejudice: Social-cognitive goals affect amygdala and stereotype activation. *Psychological Science*, *16*, 56–63.

Zahn, R., Moll, J., Krueger, F., Huey, E. D., Garrido, G., & Grafman, J. (2007). Social concepts are represented in the superior anterior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 6430–6435.