Anterior Temporal Lobe Tracks the Formation of Prejudice

Hugo J. Spiers[1], Bradley C. Love[1,2], Mike E. Le Pelley[3], Charlotte E. Gibb[1], and Robin A. Murphy[4]

Abstract

■ Despite advances in understanding the brain structures involved in the expression of stereotypes and prejudice, little is known about the brain structures involved in their acquisition. Here, we combined fMRI, a task involving learning the valence of different social groups, and modeling of the learning process involved in the development of biases in thinking about social groups that support prejudice. Participants read descriptions of valenced behaviors performed by members of novel social groups, with majority groups being more frequently encountered during learning than minority groups. A model-based fMRI analysis revealed that the anterior temporal lobe tracked the trial-by-trial changes in the valence associated with each group encountered in the task. Descriptions of behavior by group members that deviated from the group average (i.e., prediction errors) were associated with activity in the left lateral PFC, dorsomedial PFC, and lateral anterior temporal cortex. Minority social groups were associated with slower acquisition rates and more activity in the ventral striatum and ACC/dorsomedial PFC compared with majority groups. These findings provide new insights into the brain regions that (a) support the acquisition of prejudice and (b) detect situations in which an individual's behavior deviates from the prejudicial attitude held toward their group. ■

INTRODUCTION

Stereotypes and prejudice contribute to the intergroup attitudes that underlie a great deal of our social cognition (Fiske, 1998). Stereotypes are cognitive representations of the properties and features of social groups: how members of group X are similar to one another and different from members of group Y. Stereotypes can convey purely descriptive meaning and thus differentiate between groups on descriptive dimensions (e.g., "Scottish people are fair-skinned"; "Greek people are tanned"), or they can involve valenced information (e.g., "Buddhists are peaceful"; "Neo-Nazis are violent"). Relations of groups with valenced information make contact with the concept of prejudice, which refers to attitudes of liking or disliking members of social groups based on their membership ("I do not like Neo-Nazis"). On the one hand, stereotyping and prejudice have the capacity to improve decision-making by simplifying a large body of complex knowledge and allowing us to make rapid judgments regarding group members (Kunda, 1999); on the other hand, they can negatively bias our choices because of overgeneralization (Macrae & Bodenhausen, 2000; Fiske, 1998; Gilbert & Hixon 1991).

Recently, there has been substantial interest in using fMRI to explore the brain regions and to help unpack the processing involved in social judgments that might support stereotypic and prejudicial beliefs (e.g., Amodio, 2014; Gilbert, Swencionis, & Amodio, 2012; Contreras, Banaji, & Mitchell, 2011; Quadflieg & Macrae, 2011; Quadflieg et al., 2009, 2011; Amodio & Lieberman, 2009; Mitchell, Ames, Jenkins, & Banaji, 2009). Such studies have tended to explore either the brain regions involved in the application of learned attitudes (Contreras et al., 2011; Quadflieg et al., 2009, 2011) or the implicit impact that stereotypes have on cognition and brain activation (Gilbert et al., 2012; Mitchell, Macrae, & Banaji, 2006; Wheeler & Fiske, 2005; Cunningham, Raye, & Johnson, 2004; Phelps et al., 2000). From this research, a number of brain regions have been implicated, which include the anterior temporal lobe, amygdala, insula, striatum, dorsal medial PFC, and lateral PFC (see Amodio, 2014, for a review). Such brain regions overlap to some degree with brain regions implicated in thinking about other people's mental states (Amodio & Frith, 2006; Frith & Frith, 2006), which may relate to a need to consider the intentions of others when learning about them.

Despite these advances, there has been little investigation of the brain regions involved in the formation of intergroup attitudes. Although such attitudes can be developed indirectly via cultural transmission (e.g., we may hold prejudicial beliefs regarding groups that we have

[1]University College London, [2]Alan Turing Institute, London, [3]University of New South Wales, [4]University of Oxford

Journal of Cognitive Neuroscience 29:3, pp. 530–544
doi:10.1162/jocn_a_01056

never actually experienced: Katz & Braly, 1933), it is equally clear that our stereotypic and prejudicial attitudes are also influenced by our own experience with group members (Olson & Fazio, 2006; Fazio & Olson, 2003). Acquiring and updating attitudes on the basis of experience involve processes of learning and memory involved in categorization, wherein people learn to associate social groups with certain types of behavior, traits, or valence based on their experience with group members, according to principles described by learning theories (e.g., Murphy, Schmeer, Vallée-Tourangeau, Mondragon, & Hilton, 2011; Le Pelley et al., 2010; Sherman et al., 2009; Van Rooy, Van Overwalle, Vanhoomissen, Labiouse, & French, 2003). Previous neuroimaging studies have successfully applied the concepts in learning theories to examine the brain regions involved in predicting another individual's behavior (Mende-Siedlecki, Cai, & Todorov, 2012; Suzuki et al., 2012; Cloutier, Gabrieli, O'Young, & Ambady, 2011; Harris & Fiske, 2010; Behrens, Hunt, Woolrich, & Rushworth, 2008). However, such studies have focused on learning the attributes of specific individuals (person perception) rather than the perception of social groups. Moreover, there has been little research exploring how brain regions process minority social groups in comparison with majority groups.

As a particular class of social groups, minorities tend to be perceived differently from groups constituting most of the population (Hilton & Von Hippel, 1996). Rather than assuming that this bias is a function of a specific culturally transmitted experience, Hamilton and Gifford (1976) suggested that minority group biases might emerge from unbiased learning processes. It is a common perception in the media that minority groups possess, on average, more negative traits than majority group members (Hilton & Von Hippel, 1996). More generally, minority groups are perceived as less representative of the overall population than are majority groups. That is, in the context of a population engaged in mostly positive behaviors, minority groups are typically perceived as less positive than majority groups; in the context of a population engaged in mostly negative behaviors, minorities are perceived as less negative than majorities (Hamilton & Gifford, 1976). Hamilton and Gifford (1976) developed an experimental procedure for evaluating the acquisition of minority group perceptions and an attentional theory for explaining this effect. In a first experiment, they presented participants with statements describing behaviors performed by members of two fictitious groups, Groups A and B (e.g., "Joe, a member of group A, helped the old man across the street"). Members of both Groups A and B were described as engaging in the same proportion of positive and negative behaviors and with the same greater proportion of positive behaviors (69% positive, 31% negative). Hence, Hamilton and Gifford (1976) argued that an unbiased observer would judge both groups as being equally likeable. However, they also manipulated whether the statements described a majority or minority group by presenting participants with twice as many statements about members of Group A as Group B. On a range of subsequent measures, participants perceived the minority group (B) as less desirable than the majority group (A). The same result was found in their Experiment 2 in which both minority and majority groups were engaged in more negative behaviors than positive behaviors. Under these conditions, participants judged the minority group as less negative. This effect was termed an "illusory correlation" because there was no objective correlation between group membership and behavioral valence for the two groups. Subsequent research has found this to be a robust effect (for a review, see Berndsen, Spears, van der Pligt, & McGarty, 2002).

Recent work (Kutzner & Fiedler, 2015; Murphy et al., 2011) has suggested that this type of correlational learning is captured by the basic assumptions derived from error correction learning algorithms such as that proposed by Rescorla and Wagner (1972). Selective learning, like that observed in the illusory correlation study (Hamilton & Gifford, 1976), can be accounted for by the incremental acquisition of associative links between group labels and the emotional valence signaled by the behavioral statements (Murphy et al., 2011). Under this approach, the illusory correlation effect emerges naturally as a consequence of differences in the amount of learning about the two groups. Consider the situation in which both majority and minority groups are paired with mostly positive behaviors. Because associative models are sensitive to the degree of contingency between events, this approach anticipates that observers will—at asymptote—form equally strong, positively valenced beliefs regarding both groups. However, because changes in associative strength are a function of experience, less experience of the minority group will ensure that learning about this group lags behind learning about the majority group. Consequently, at a given point in training (before asymptote), positive attitudes will tend to be stronger for the more frequently presented majority group than the less frequent minority group. Consistent with this account, this was exactly the pattern observed empirically by Murphy et al. (2011).

Here, we used an adapted version of the procedure described by Hamilton and Gifford (1976) and fMRI to determine the brain regions involved in the formation of intergroup prejudicial attitudes. By applying a simple associative model to the participants' data, we were able to derive estimates of the trial-by-trial parameters capturing the learned valence of the social groups encountered and the prediction errors associated with learning. Whereas Hamilton and Gifford (1976) had participants learn about two social groups, we had participants learn about four groups, which differed in minority/majority status and valence (largely negative or largely positive). This allowed us to examine the brain regions involved in learning attitudes about social groups as a function of both the group's overall valence (positive vs. negative) and its minority/majority status.

METHODS

Participants

Twenty-two right-handed, healthy volunteers (11 men, age range = 20–34 years, mean = 24.1 years, *SD* = 4.3 years) with normal or corrected-to-normal vision participated in this experiment. All participants reported that they were free from neurological and psychiatric disease and gave informed written consent in accordance with the local research ethics committee. During scanning, head movements were monitored by observing the difference in position of the head between each volume acquired on the MRI console. Four (one man, three women) participants moved considerably (approximately >8 mm in one plane) at repeated moments during scanning. We excluded these participants because of the poor quality of data generated by such motion.

Stimuli

Stimuli presented during scanning were the names of 12 fictitious groups and 180 sentences describing valenced behaviors (see Figure 1A), which had been piloted with a separate set of participants to generate ratings of valence, commonness, and imageability. Eight of the group names were allocated to our main experimental trials (Hezlatts, Pellums, Grallacks, Selners, Kitils, Vatges, Drayos, and Trithans); four were used in each scanning session. Four names were allocated to control condition trials (Raymers, Breggs, Cetners, and Feplers); two were used in each scanning session. The experimental group names were pseudorandomly allocated to each type of group across participants such that the names were as likely to be associated with a positive group or a negative group and a minority group or a majority group. The control group names were similarly likely to be associated with a majority or minority control group. Sentences describing behaviors were selected from an initial pool of 500 sentences developed by our team. An initial inspection of the sentences suggested that commonness and imageability would be highly variable across them and might correlate with valence. Thus, 15 undergraduates who did not take part in the fMRI study rated the 500 sentences on valence, commonness, and imageability using a scale of 1–7 (with 1 representing extremely negative, uncommon, or unimageable; 7 representing extremely positive, common, or imageable; and 4 representing a neutral midpoint). Across all sentences, valence was positively correlated with commonness ($r = .62$, $p < .01$) and imageability ($r = .30$, $p < .01$). Commonness

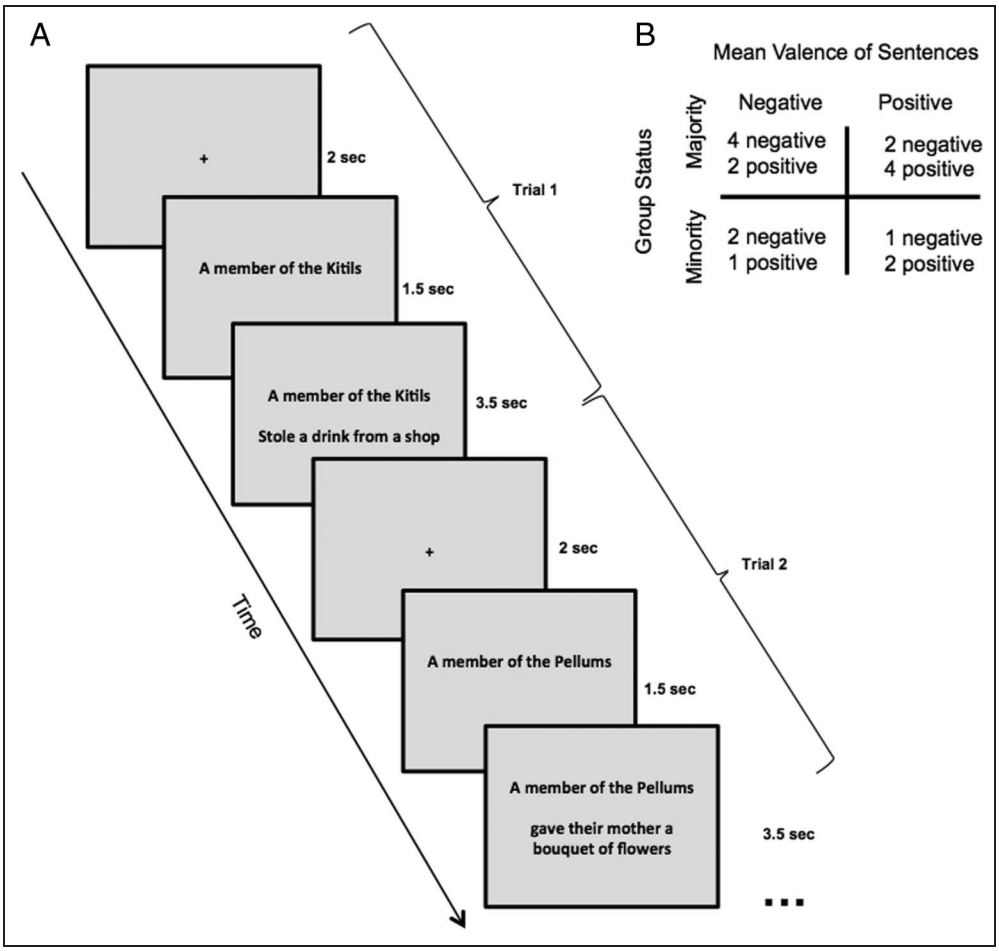

Figure 1. Experimental task. (A) Example of stimuli presented in 2 of the 18 trials composing each of the five experimental blocks. Each trial began with the presentation of a fixation cross, which was followed by the presentation of a group name and followed by a description of a behavior. Control block trials were identical to the main blocks with the exception that no behavior was presented below the group name and only two group names were used in the blocks. Each block of 18 trials was separated by the sequential presentation of each of the group names and a prompt to rate the valence of each group. There was no time constraint on the rating period. (B) Contingency table for the composition of trials in each block of 18 trials.

and imageability were also positively correlated ($r = .28$, $p < .01$).

A tertile split on valence ratings was used to divide sentences into positive (the top-rated third of sentences) and negative (the bottom-rated third of sentences) categories. A cutoff was imposed to exclude sentences with a mean rating lying within 0.5 points of the extremely valence pole. This was done to improve the likelihood that learning would be gradual (Murphy et al., 2011). Negative and positive sentences were ranked according to their standard deviation, and 90 sentences in each group with the smallest standard deviation were chosen as stimuli. The mean valence judgments for the negative and positive sentences that were chosen as stimuli were $M = 2.13$ ($SE = 0.06$) and $M = 6.09$ ($SE = 0.03$), respectively, where 1 = *most negative* and 7 = *most positive*. For example, the sentence "Consoled a friend whose grandfather had passed away" was rated as 6.53, whereas "Kicked a stray cat on the way home from work" was rated as 1.40. The sentences describing negative behaviors were judged as being less common behaviors than sentences describing positive behaviors (negative: $M = 3.96$, $SE = 0.10$; positive: $M = 5.11$, $SE = 0.66$; $t(178) = 9.38$, $p < .001$). Positive sentences were rated as having higher levels of imageability compared with negative sentences (positive: $M = 4.98$, $SE = 0.06$; negative: $M = 4.57$, $SE = 0.07$; $t(178) = 4.38$, $p < .001$). There was no significant difference in the lengths of the sentences (number of words) between negative and positive sentences (positive: $M = 8.3$, $SE = 1.9$; negative: $M = 8.9$, $SE = 2.08$; $t(178) = 1.76$, $p = .08$).

The learning task during fMRI consisted of two separate learning sessions, each consisting of five blocks of 18 trials. This approach is similar to that used by Murphy et al. (2011), which reported significant effects of learning over five blocks with two groups. Each trial consisted of the presentation of a group name followed by a description of a behavior. Each block of 18 trials contained six trials displaying the majority negative group name, six trials displaying the majority positive group name, three trials displaying the minority negative group name, and three trials displaying the minority positive group name (see Figure 1). For each positive group, there were two positive sentences for every one negative sentence, and vice versa for the negative group. Thus, similar to Hamilton and Gifford's (1976) design, the behaviors associated with the majority and minority groups were equally positive or negative but differed purely in terms of the frequency with which participants encountered them (see Figure 1B). Negative and positive sentences were randomly allocated to the groups across participants, and trial order was randomized within each block, with the constraints that (1) the mean valence was matched across minority and majority groups, (2) no more than two trials with the same group name could occur sequentially, and (3) at least one trial with each group occurred in the first six trials, the middle six trials, and the last six trials. This was done to minimize the influence of any recency bias in participants' valence ratings of the group names at the end of each block.

Experimental Design and Procedures

Participants were informed that they would read sentences describing the behavior of people in different social groups and would have to intermittently rate how positive or negative they perceived these social groups to be. They were not informed about differences in group frequency or proportion of negative/positive sentences. Before scanning, participants were familiarized with the trial and rating structure. A single practice trial was presented consisting of a group name and a sentence not used in the actual experiment. After this, participants were presented with the rating screen and allowed to practice making a rating.

Two scanning sessions were conducted. Each session involved learning about a new set of fictitious social groups. During each session, five learning blocks of 18 trials were presented. Each trial began with the presentation of one of the four fictitious group names used for that session; after a short delay (1.5 sec), a sentence describing a behavior appeared underneath (for 3.5 sec; see Figure 1A). The intertrial interval was 2 sec during which a fixation cross was displayed. At the end of the 2-sec fixation cross on the last trial of each block of trials, participants were asked to make an evaluative judgment about each of the groups by rating them on a scale from −10 (*strongly dislike*) to 10 (*strongly like*). Ratings were made by pressing buttons to increase or decrease the value presented on the screen, which was initially set to 0. The order of the group names presented in the rating period was randomized with the constraint that no group name could appear in the first, second, third, or last position more than twice across blocks. Rating was self-paced, and no feedback was given.

Learning blocks were interspersed with a control block containing nine trials. Six of these trials featured the majority control group name, and three featured the minority control group name. On each of these control trials, participants saw the group name and a fixation cross, but no sentence describing behavior was presented. After the nine trials of each control block, the names of the control groups were presented again in a similar fashion to the rating period for the experimental group names, but rather than rating them, participants were instructed to press a button when each group name appeared.

Modeling Learning from Behavioral Data

A simple associative model was fit to the behavioral data and used to provide trial-by-trial measures for model-based fMRI analyses. The model, based on the Rescorla–Wagner (RW) model (Rescorla & Wagner, 1972), can be viewed as a simple one-layer neural network (cf. Widrow

& Hoff, 1988) in which the input layer I consists of four units (one for each social group) and the output layer O consists of one unit for positive valence and one unit for negative valence. The activation of output unit O_j is calculated as:

$$O_j = \sum_i^m I_i w_{ij} \quad (1)$$

where m is the number of input units and w_{ij} is the learned association weight from input unit I_i to output unit O_j. Association weight w_{ij} is updated according to the following learning rule:

$$\Delta w_{ij} = n_i I_i \left(T_j - O_j\right) \quad (2)$$

where n_i is the learning rate associated with input unit I_i and T_j is the target (i.e., observed) value for O_j. Weights are adjusted by associative learning to minimize the discrepancy between predicted and observed outcomes. Different learning rate parameters allow for the possibility that different inputs vary in their saliency or associability. For example, social groups associated with negative behaviors may be more salient than groups associated with positive behaviors.

In the present task, all units were coded as 0 (*absent*) or 1 (*present*). For instance, when the second social group was presented, the input unit corresponding to this group would be set to 1 with the other input units set to 0. Likewise, during learning, when a negative behavior was observed, the output unit representing negative valence would be set to 1 with the unit representing positive valence set to 0.

fMRI Parameters and Acquisition

Participants were scanned at the Birkbeck-UCL Centre for Neuroimaging using a 1.5-T Siemens (Siemens Medical Systems, Erlangen, Germany) Avanto MRI scanner, with a 32-channel head coil. Functional scans were acquired using a gradient-echo EPI sequence (repetition time = 3000 msec, echo time = 48 msec, field of view = 205 × 205, matrix = 64 × 64). In each volume, thirty-six 3.2-mm-thick oblique axial slices were acquired. Anterior-to-posterior phase encoding and a tilt were applied to the sequence to improve signal in OFC and amygdala. After this, a high-resolution T1 structural scan was acquired (magnetization prepared rapid gradient echo, 176 slices, 1 × 1 × 1 mm resolution). Foam padding was used to minimize head motions, and ear plugs were used to dampen the noise of the scanner. Stimuli were projected centrally onto a screen at the front of the magnet, which participants viewed using a mirror mounted on the head coil (visual angle of the whole screen = 21° × 13°).

fMRI Preprocessing and Statistical Analysis

The first six functional volumes (dummy scans) of each of the two sessions conducted were discarded to permit T1 equilibrium. Initial inspection of the EPI images revealed that four participants moved excessively during their scans, and these were excluded before preprocessing (see Participants section). SPM (SPM8; www.fil.ion.ucl.ac.uk/spm/software/spm8) was used for spatial preprocessing and subsequent analyses. Images were spatially realigned to the first volume of the first session to correct for motion artifacts, normalized to a standard EPI template in Montreal Neurological Institute space with a resampled voxel size of 3 × 3 × 3 mm, and smoothed with an 8-mm FWHM Gaussian kernel filter.

After preprocessing, the smoothed, normalized functional imaging data were entered into three voxel-wise participant-specific general linear models (GLMs; i.e., the first-level design matrix). The first GLM was constructed to test for the effects of parameters derived from our RW model. The second GLM focused on the RW model parameters for the response for the negative groups because the participants' ratings for negative groups revealed clearer evidence of learning over the five blocks (see Results below). The third GLM was specified to examine the categorical effects of valence (negative vs. positive groups) and frequency (majority vs. minority groups).

The first GLM consisted of the following categorical regressors (modeled as box-car functions spanning the duration of each trial/period) for each session: (1) rating periods, (2) control block minority group trials, (3) control block majority group trials, and (4) all the trials of the groups in the main learning blocks. Two additional parametric modulators of the main experimental group trials were entered. These were the learned valence of each group (O_j) and the prediction error parameter ($T_j - O_j$) derived from our RW model of the behavioral data (see Figure 2). To examine the independent effects of learned valence and prediction error, we entered our parametric regressors without serial orthogonalization. These regressors convolved with the canonical hemodynamic response function. Furthermore, six participant-specific movement parameters (derived from the realignment phase of preprocessing) were also included as regressors of no interest in the model.

The second GLM explored parametric responses for the negative groups. This analysis was conducted after we had determined that the ratings of the negative groups showed more gradual learning over the blocks than did the ratings of the positive groups. This GLM was identical to the first GLM except for two alterations. The first alteration was that, in each session, the single regressor for all the main experimental trials was replaced by three regressors: one for the minority positive group, one for the majority positive group, and one for all of the negative group trials. The second alteration was that parametric modulation (learned affect and prediction

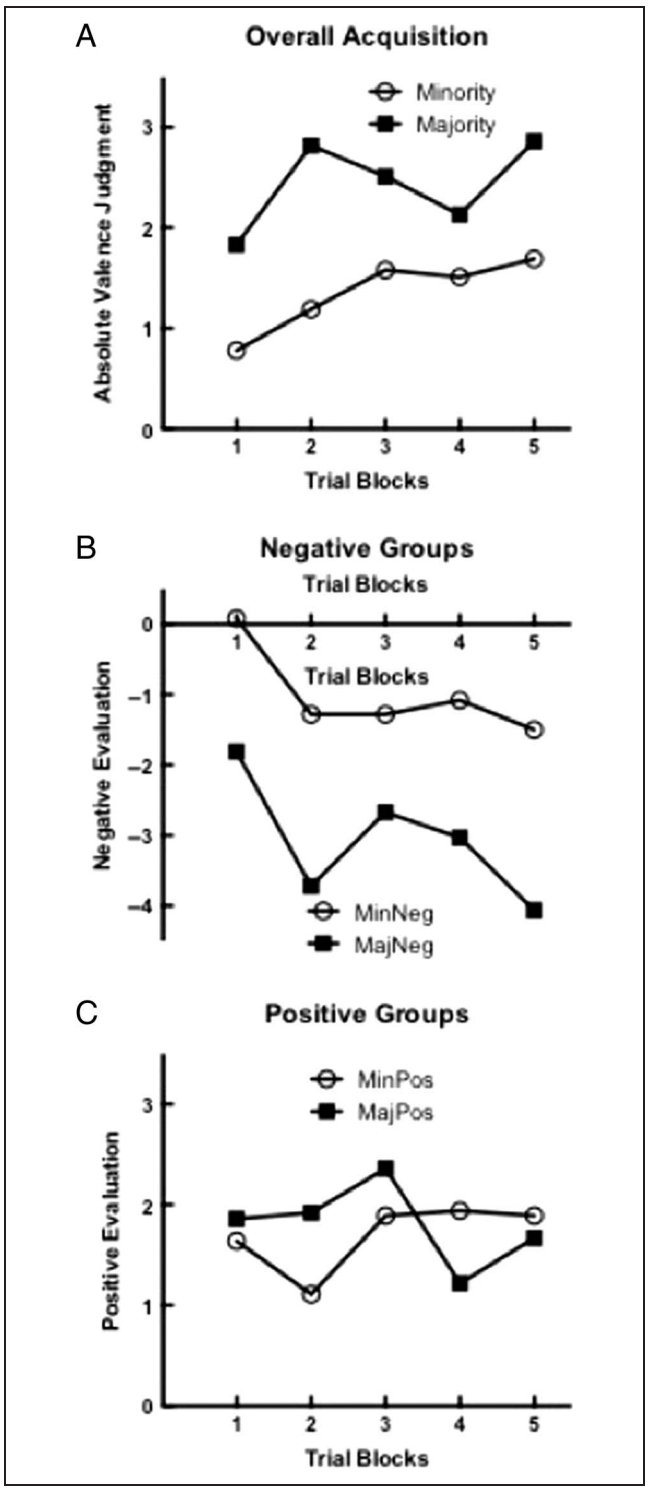

Figure 2. Mean valence ratings across learning sessions. In each plot, the mean valence of the groups is plotted against the trial block. This mean is derived from the average of the two learning sessions. Ratings could vary between 10 (*most positive*) and −10 (*most negative*). (A) Ratings collapsed over both the negative and positive groups. (B) Negative and positive groups separated.

error) was applied only to this negative group trial regressor.

The third GLM explored categorical effects of all groups and was identical to the second GLM except for two alterations. These were as follows: (1) the regressor for the negative trials was replaced by two regressors, one for the negative majority group trials and one for the negative minority group trials, and (2) no parametric modulation was applied.

For all models, a high-pass filter with a cutoff of 128 sec was used to remove low-frequency drifts. Temporal autocorrelation was modeled using an AR(1) process. At the first level, linear weighted contrasts were used to identify effects of interest, providing contrast images for group effects analyzed at the second (random effects) level. Given our a priori anatomical hypotheses, we predicted significant activations in the following brain regions: anterior temporal lobe, lateral (inferior frontal gyrus) and medial PFC, amygdala, striatum, and insula (see Amodio, 2014, for a review). All ROIs were anatomically predefined in the Wake Forest University PickAtlas software (www.fmri.wfubmc.edu), except the insula, which was defined in the SPM anatomy toolbox (Eickhoff et. al., 2005; www.fil.ion.ucl.ac.uk/spm/ext/#Anatomy). A threshold of $p < .05$ corrected for family-wise error and a minimum of five voxels were used for the whole-brain analysis and within ROIs. No contrasts using a threshold of $p < .05$ corrected for the whole-brain volume with a minimum of five voxels revealed any significant effects in our analysis. For completeness, we report all regions significant at a threshold of $p < .001$ (uncorrected for multiple comparisons) and a minimum of five voxels in Table 1.

RESULTS

Behavioral Results

Analysis of the mean judgments of the group labels across the two scanning sessions revealed that participants learned about the positive and negative groups. Figure 2 presents the mean valence for the majority and minority groups averaged across the two sessions of training and averaged across the two valences (positive and absolute value of the negative judgments). The pattern of data shows a differential learning effect for the majority and minority groups. A mixed ANOVA with factors of Valence (negative vs. positive) × Group size (minority vs. majority) × Blocks (5) × Scanning session (2) supported these observations. The negative groups and the majority groups received more extreme ratings overall as illustrated by the reliable main effects of Valence ($F(1, 17) = 17.276$, $p < .001$, $\eta_p^2 = 0.504$) and Group size ($F(1, 17) = 4.744$, $p < .04$, $\eta_p^2 = 0.218$). In addition, the two-way interactions between Valence and Group size as well as between Valence and Blocks were significant ($F(1, 17) = 4.934$, $p < .04$, $\eta_p^2 = 0.225$ and $F(4, 68) = 2.734$, $p < .036$, $\eta_p^2 = 0.139$, respectively). There

Table 1. Coordinates and *Z* Scores for Brain Regions Identified in the SPM Analysis

Brain Region	*x, y, z*	*Cluster Size*	*Z Score*	*Significance Threshold*
Learned Affect (O_j) (All Groups)				
R anterior temporal cortex	**27, 8, −35**	**14**	**4.01**	**FWER < .05**
L anterior temporal cortex	−33, 11, 35	15	3.10	<.005 u.c.[a]
R cerebellum	15, −55, −32	11	4.02	<.001 u.c.
L intraparietal sulcus	−36, −61, 34	12	3.56	<.001 u.c.
R middle occipital gyrus	24, −88, 4	15	3.45	<.001 u.c.
Prediction Error ($T_j - O_j$) (All Groups)				
No regions				
Learned Affect (O_j) (Negative Groups)				
L anterior temporal cortex	**−51, 14, −23**	**7**	**4.03**	**FWER < .05**
R superior parietal gyrus	21, −73, 46	172	4.05	<.001 u.c.
R superior frontal gyrus	3, 17, 55	88	3.77	<.001 u.c.
R parahippocampal gyrus	24, −31, −17	9	3.76	<.001 u.c.
R angular gyrus	39, −76, 22	31	3.75	<.001 u.c.
R inferior occipital gyrus	−27, −85, −14	6	3.70	<.001 u.c.
R middle occipital gyrus	30, −88, 4	18	3.63	<.001 u.c.
R middle frontal gyrus	42, 23, 34	23	3.61	<.001 u.c.
L cuneus	−3, −94, −5	10	3.40	<.001 u.c.
Prediction Error ($T_j - O_j$) (Negative Groups)				
L inferior frontal gyrus (lPFC)	**−48, 32, 4**	**328**	**5.28**	**FWER < .05**
L superior frontal gyrus (dmPFC)	**−6, 50, 37**	**117**	**4.14**	**FWER < .05**
R anterior temporal gyrus	**54, 14, −20**	**23**	**4.29**	**FWER < .05**
L calcarine sulcus	−15, −88, −2	44	4.56	<.001 u.c.
R calcarine sulcus	15, −79, 10	29	3.72	<.001 u.c.
R thalamus	9, −13, 7	24	4.44	<.001 u.c.
L thalamus	−12, −4, 13	28	3.90	<.001 u.c.
L middle frontal gyrus	−39, −1, −39	63	4.08	<.001 u.c.
L superior temporal gyrus	−57, −55, 16	107	3.86	<.001 u.c.
L fourth occipital gyrus	−39, −58, −11	16	3.70	<.001 u.c.
Categorical: Negative Groups > Positive Groups[b]				
R inferior frontal gyrus	**51, 23, 16**	**59**	**4.34**	**FWER < .05**
L inferior frontal gyrus	**−51, 20, 7**	**5**	**3.57**	**FWER < .05**
R ACC	**6, 35, 28**	**56**	**3.37**	**FWER < .05**

Table 1. (*continued*)

Brain Region	*x, y, z*	*Cluster Size*	*Z Score*	*Significance Threshold*
R superior frontal gyrus	12, 41, 46	31	3.81	<.001 u.c.
R middle frontal gyrus	39, 8, 46	6	3.55	<.001 u.c.
R angular gyrus	63, −43, 28	14	3.59	<.001 u.c.
Categorical: Positive Groups > Negative Groups[b]				
R cingulate sulcus	15, −37, 52	44	4.34	<.001 u.c.
R putamen	33, −16, 7	24	4.17	<.001 u.c.
L uncus	−30, 2, −23	7	4.00	<.001 u.c.
Categorical: Minority Groups > Majority Groups				
dmPFC/ACC	**0, 23, 40**	**145**	**4.82**	**FWER < .05**
R ventral striatum	**21, 17, −2**	**93**	**4.80**	**FWER < .05**
L ventral striatum	−18, 14, −5	46	3.88	<.001 u.c.
L superior parietal gyrus	−12, −67, 40	741	4.79	<.001 u.c.
R superior parietal gyrus	12, −67, 37	741	4.75	<.001 u.c.
R posterior cingulate	6, −43, 19	224	4.53	<.001 u.c.
R inferior occipital gyrus	51, −61, −14	119	4.50	<.001 u.c.
L fusiform gyrus	−45, −52, −20	165	4.13	<.001 u.c.
R pulvinar	15, −34, 1	19	4.14	<.001 u.c.
L pulvinar	−21, −25, −5	36	3.91	<.001 u.c.
L middle occipital gyrus	−30, −91, −11	19	3.98	<.001 u.c.
R middle occipital gyrus	33, −91, −2	150	3.88	<.001 u.c.
R middle frontal gyrus	36, 5, 64	16	3.92	<.001 u.c.
L middle frontal gyrus	−42, 35, 34	18	3.81	<.001 u.c.
R cerebellum	33, −46, −26	32	3.75	<.001 u.c.
L circular insular sulcus[c]	−39, 14, 4	12	3.72[c]	<.001 u.c.
R angular gyrus	39, −67, 37	20	3.52	<.001 u.c.
Categorical: Majority Groups > Minority Groups				
L dmPFC	−9, 20, 43	10	3.86	<.001 u.c.
R cingulate sulcus	12, −16, 46	25	3.70	<.001 u.c.
L superior frontal gyrus	−9, 50, 37	6	3.50	<.001 u.c.
Categorical: Minority Control Group > Majority Control Group				
L superior frontal gyrus	−30, 2, 67	36	4.99	<.001 u.c.
R superior frontal gyrus	15, 11, 67	5	3.56	<.001 u.c.
L inferior frontal gyrus	−36, 29, 28	33	4.34	<.001 u.c.
L postcentral gyrus	−48, −22, 40	15	3.99	<.001 u.c.

Table 1. (*continued*)

Brain Region	*x, y, z*	*Cluster Size*	*Z Score*	*Significance Threshold*
R middle frontal gyrus	54, −4, 43	10	3.35	<.001 u.c.
R intraparietal sulcus	33, −64, 34	6	3.27	<.001 u.c.
Categorical: Majority Control Group > Minority Control Group				
L anterior hippocampus	−30, −16, −17	6	3.45	<.001 u.c.

No brain region survived FWER correction for whole-brain volume. FWER = family-wise error corrected for anatomically defined ROI; L = left; R = right; u.c. = uncorrected threshold.

[a]Region does not survive correction for small-volume ROI but is listed here for completeness (see Figure 3). The ventral striatum refers to activity in the region of the ventral putamen/head of the caudate/nucleus accumbens.

[b]These analyses are included to help characterize the data, rather than test specific predictions.

[c]Activity in the insula did not reach significance for family-wise error correction for the insula ROI.

was also a three-way interaction between Blocks, Valence, and Group size, indicating that the learning effect across the blocks was different depending on valence and group size ($F(4, 68) = 2.738, p < .036, \eta_p^2 = 0.139$). To investigate this three-way interaction effect, we analyzed the individual blocks and found that, during Blocks 1 and 2, only the main effect of Valence was significant ($F(1, 17) = 10.66, p < .005, \eta_p^2 = 0.385$). That is, on these two blocks, the only difference was the stronger ratings for the negative groups. However, for Block 3, both the effect of

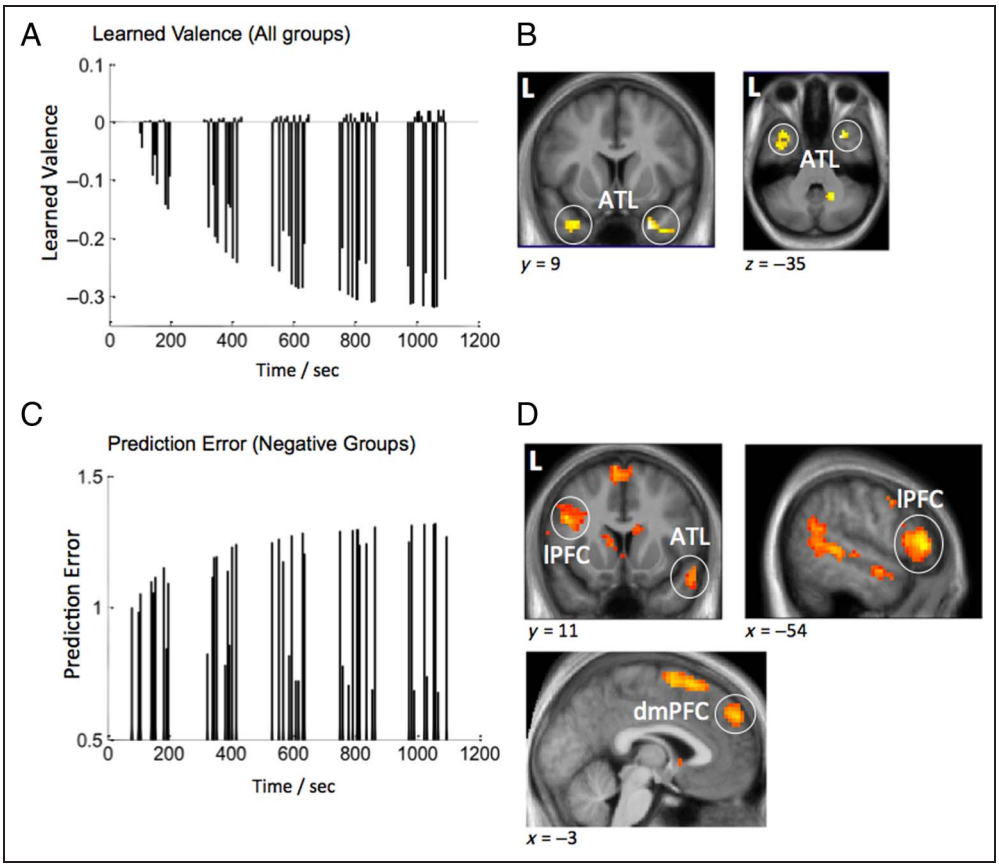

Figure 3. Brain activity that tracks learned affect and prediction error. (A) The time course of learned affect for all four social groups, plotted from the data of a single representative participant from one of the two learning sessions. The learned valence parameter was calculated as quantity O_t in Equation 1. (B) Activity in the anterior temporal lobes (white circles) significantly correlated with the learned valence parameter for all groups (see Table 1). Display threshold $p < .005$ uncorrected, shown on the mean structural image. (C) The time course for prediction error for the negative groups for the same session and participant as shown in A. The prediction error was calculated as $T_j - O_j$ (see Equation 2). (D) Activity significantly correlated with the negative group prediction error: lateral PFC (lPFC), dorsomedial PFC (dmPFC), and anterior temporal lobe (ATL; see Table 1). SPMs are displayed on the mean structural image at a threshold of $p < .005$, uncorrected. Note that the activity shown close to the caudate does not survive at $p < .001$ uncorrected or at corrected thresholds.

Valence ($F(1, 17) = 4.906, p < .041, \eta_p^2 = 0.224$) and the main effect of Group size were significant ($F(1, 17) = 6.363, p < .022, \eta_p^2 = 0.272$); no other effects were significant. This latter result suggests that, although the negative groups remained more strongly negative than the positive groups were positive, by Block 3, there was also an effect of Group size whereby the majority groups received more extreme ratings than the minority groups, replicating previous demonstrations of the illusory correlation effect (Murphy et al., 2011; Hamilton & Gifford, 1976). Figure 2 shows that, during the last two blocks (Blocks 4 and 5), the effect of Group size disappeared for the positive groups but not for the negative groups, which continued to show stronger judgments of the majority group. Specifically, during Blocks 4 and 5, the main effects of Valence and Group size interacted (minimum: $F(1, 17) = 6.986, p < .017, \eta_p^2 = 0.291$ on Block 5) such that only the negative groups were demonstrating a significant effect of Group size (minimum: $F(1, 17) = 5.591, p < .030, \eta_p^2 = 0.247$).

Modeling Results

The model was fit to the behavioral data derived from the ratings after each block and the postscan sentence ratings given to the stimuli. The fitting procedure attempted to maximize the correlation between the model's predicted valence judgment for each group (calculated as the difference between the activity of positive and negative valence output units) and the ratings provided by the participants. These model predictions were linearly transformed to best match ($r^2 = .92$) the participants' ratings (see Figure 2). In addition to the linear transformation, model parameters included a learning rate of 0.109 for the groups predominately associated with negative behaviors and a second learning rate of 0.002 for the positive groups. That is, the best-fitting learning rate for the negative groups was almost 50 times higher than that for positive groups. These fitted learning rates indicate that there was little change in valence judgments over the course of training for positive groups, whereas significant learning for the negative groups continued across training.

fMRI Results

Anterior Temporal Lobe Activity Tracks the Acquisition of the Learned Valence of the Social Groups

Within our ROIs, we found that activity in the right temporal pole was significantly positively correlated with our parametric measure of learned valence across all groups (see Figure 3A, Table 1). At a more liberal threshold ($p < .005$ uncorrected), activity was also significantly correlated with learned valence in the left temporal pole (Table 1). Because the positive groups showed little evidence of learning after the first block (Figure 2), we focused our second analysis on the negative groups, which had a pattern of more gradual learning over the five blocks in each session (Figure 2). We found a significant positive correlation between learned valence of the negative groups and activity in the left anterior temporal cortex (Table 1).

Prediction Errors for the Negative Groups Were Correlated with Prefrontal and Anterior Temporal Lobe Activity

An initial analysis involving all groups found no evidence of brain activity significantly correlated with the prediction errors, even at liberal uncorrected thresholds ($p < .001$ uncorrected). By contrast, when we examined the prediction error for the negative groups (which was more variable over the course of the two sessions; see Figure 3B), we found a network of brain regions, which included the left lateral PFC, dorsomedial PFC, and right lateral anterior temporal lobe (see Figure 3 and Table 1).

Categorical Effects of Group Status (Minority/Majority) and Valence

In addition to examining the data with parametric variables derived from our model, we also examined the

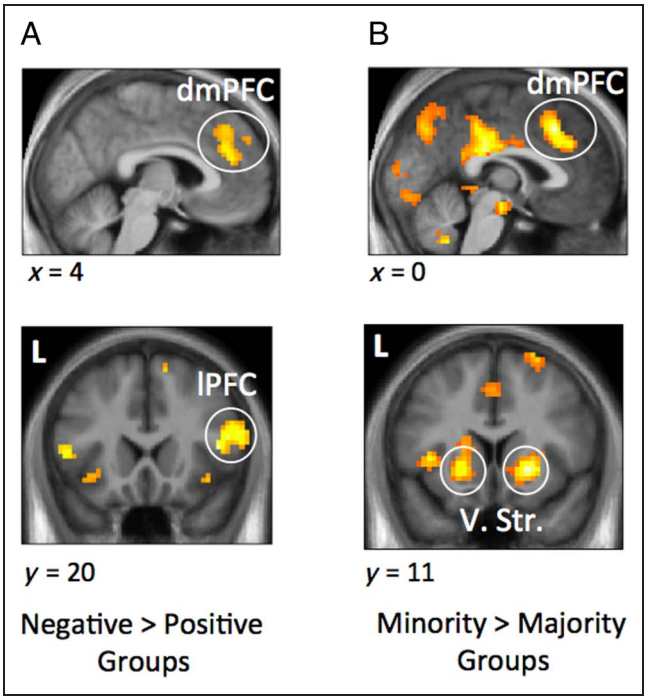

Figure 4. Categorical responses to valence and group status. All SPMs are shown displayed on the mean structural image at a threshold of $p < .005$, uncorrected. See Table 1 for Montreal Neurological Institute coordinates and Z scores. (A) Increased activity in the lateral PFC (lPFC) and dorsomedial PFC (dmPFC) in response to the negative groups compared with the positive groups. (B) Increased activity in the ventral striatum (V. Str.) and dmPFC/cingulate regions in response to the minority groups compared with the majority groups.

categorical effect of group size (minority/majority). Activity in the ventral striatum and ACC/dorsomedial PFC was significantly greater for the minority groups compared with the majority groups (Figure 4B, Table 1). None of our ROIs showed significantly greater activity for majority groups than minority groups. We next examined whether similar patterns of activity would be elicited in our control task, in which purely group names were presented alone without any accompanying descriptions of behaviors. No evidence of increased activity in any of our ROIs was detected when minority and majority control group trials were compared (Table 1), suggesting that the patterns obtained in the earlier analyses were relatively specific to the case in which participants were asked to form impressions of the social groups.

Finally, we compare responses elicited by the positive and negative groups. Because negative behaviors were associated with being less imageable and more uncommon (see Stimuli), the comparison of negative and positive groups is difficult to draw clear inferences from. Thus, we report this contrast to support future replication of this experiment. Significantly more activity was observed in the dorsomedial PFC (extending to the ACC) and bilateral lateral PFC by the negative groups compared with the positive groups (Figure 4A). None of our ROIs revealed significantly more activity for the positive groups than the negative groups.

DISCUSSION

We combined fMRI, a novel learning task, and a model of trial-by-trial behavioral data to characterize the brain regions involved in the acquisition of attitudes regarding the valence of social groups. Our results reveal that activity in the anterior temporal lobe is correlated with the emergence of prejudicial beliefs regarding the different social groups and that activity in structures including the anterior temporal lobe, lateral PFC, and dorsomedial PFC tracks the prediction error when gradual learning occurred. These findings provide new insights into the brain regions involved in the formation of prejudice and support the view that the anterior temporal lobe plays a prominent role in learning and representing social–emotional conceptual knowledge (Amodio, 2014; Olson, McCoy, Klobusicky, & Ross, 2013; Wong & Gallate, 2012; Zahn et al., 2007, 2009; Olson, Plotzker, & Ezzyat, 2007). Moreover, our data are consistent with the temporal pole learning the information via an error-correcting associative learning system.

On the basis of a range of evidence, the anterior temporal lobe has been viewed as a semantic hub where amodal information about concepts are stored (Chadwick et al., 2016; Visser, Jefferies, & Ralph, 2010; Patterson, Nestor, & Rogers, 2007). Recently, it has been argued that the anterior temporal lobe may play a privileged role in storing knowledge of a socioemotional nature (Amodio, 2014; Olson et al., 2007, 2013; Wong & Gallate, 2012). Damage to the temporal poles in humans and nonhuman primates typically results in profound deficits in social behavior and emotional regulation (Gozzi, Raymont, Solomon, Koenigs, & Grafman, 2009; Thompson, Patterson, & Hodges, 2003; Mychack, Kramer, Boone, & Miller, 2001; Miller, Darby, Benson, Cummings, & Miller, 1997; Miller, Darby, Swartz, Yener, & Mena, 1995; Kling, Tachiki, & Lloyd, 1993; Kling & Steklis, 1976; Franzen & Myers, 1973; Bucher, Myers, & Southwick, 1970; Kluver & Bucy, 1939). Previous neuroimaging research has often reported increased activity in the anterior temporal lobe during tasks that required participants to make interferences about other agents' mental states or emotions (Mitchell et al., 2006; Saxe & Powell, 2006; German, Niehaus, Roarty, Giesbrecht, & Miller, 2004; Grèzes, Frith, & Passingham, 2004; Iacoboni et al., 2004; Ohnishi et al., 2004; Calarge, Andreasen, & O'Leary, 2003; Berthoz, Armony, Blair, & Dolan, 2002; Gallagher, Jack, Roepstorff, & Frith, 2002; Moll et al., 2002; Vogeley et al., 2001; Brunet, Sarfati, Hardy-Baylé, & Decety, 2000; Castelli, Happe, Frith, & Frith, 2000; Gallagher et al., 2000) or during reported moments of spontaneously thinking about other people's mental states (Spiers & Maguire, 2006). It has been argued that such responses may relate to the role of the anterior temporal cortex in the retrieval of stored knowledge about likely mental states (Amodio, 2014; Olson et al., 2007; Frith & Frith, 2006). However, to date, most of the research in this area has examined the response of anterior temporal lobe regions to discrete stimuli rather than tracking the learning process. Thus, our finding that activity in the anterior temporal lobe tracks changes in the perceived valence of the different social groups provides, to our knowledge, the first evidence that this region is involved in the acquisition of prejudicial intergroup attitudes. Although the response we observed in the anterior temporal lobe may reflect neural activity involved in the encoding of the new information into the network, it may also represent activity associated with the activation of knowledge stores during the learning process that are required for the integration of new information. More research will be required to explore these possibilities.

Although our data suggest that the anterior temporal lobes are relatively selectively involved in learning the likely valence of the behaviors of the social groups, we found a separate network of brain regions involved in detecting behavior that deviated from what was expected for the group (prediction errors). Activity in this brain network was specifically correlated with the prediction error term in our model for the negative groups. A likely reason for this is that the prediction errors for the positive groups varied little over the course of the scanning sessions and would be predicted to relate to small fluctuations in the fMRI signal, thus making it difficult to detect them with our analysis. By contrast, for the negative groups, we observed gradual learning across the five blocks in the sessions, as evidenced by the much larger learning rate parameter in our model for the negative

groups. The network of brain regions that tracked the prediction error for the negative groups included the dorsomedial PFC, lateral PFC, and anterior temporal lobe. These areas have been reported to be responsive in situations involving making learned predictions about another person's behavior (Suzuki et al., 2012; Behrens et al., 2008) or updating impressions of other people's behavior (Mende-Siedlecki et al., 2012; Cloutier et al., 2011). Thus, our data are consistent with frontotemporal regions acting to compare incoming information with stored knowledge (prejudicial attitudes) about groups or individuals and updating these representations. Although we cannot separate responses involved in detecting the deviation from the expected valence and updating the representations of the group-related attitudes, it seems plausible that the temporal cortex is involved in updating the stored representations and PFC regions in detecting the deviation from that expectation. This is based on evidence that damage to the temporal lobes disrupts retrieval of long-term knowledge (see, e.g., Patterson et al., 2007) and the frontal lobes with novelty detection (see, e.g., Løvstad et al., 2012). Future research separating attitude updating from the surprise of encountering the behavior would be useful to determine if this perspective is valid (see, e.g., O'Reilly et al., 2013), as would experimental paradigms that can explore whether the pathway between the anterior temporal pole and medial PFC mediates the updating (Amodio, 2014; Olson et al., 2013).

Previous studies using associative learning tasks with nonsocial stimuli have often reported a correlation of activity in the ventral striatum with prediction error (see, e.g., O'Doherty, 2004). However, no such correlation was observed in the current study using social stimuli. Whereas increased striatal responses have been reported for social stimuli that were a violation from predicted norms (Harris & Fiske, 2010), other studies—like ours—have not found striatal areas to follow prediction errors for social stimuli (e.g., Cloutier et al., 2011; Behrens et al., 2008). It is possible that striatal responses emerge when the learning task involves direct corrective feedback, and this may be why we did not observe a significant striatal response to the prediction error. Several recent fMRI studies lend credence to this view by reporting striatal responses during learning tasks involving social stimuli with feedback (Powers, Somerville, Kelley, & Heatherton, 2016; Zaki, Kallman, Wimmer, Ochsner, & Shohamy, 2016; Hackel, Doll, & Amodio, 2015). Thus, it would be useful in future research to experimentally manipulate the presence or absence of feedback in social stimuli-based learning tasks to test this hypothesis. Although we found no ventral striatal response to the prediction error, for either all groups or only the negative groups, we did find increased activity in the ventral striatum (and ACC/medial PFC) for the minority groups compared with the majority groups. We found, similar to prior research (Murphy et al., 2011), that minority groups were generally rated as being less strong in valence compared with the majority groups, despite being a priori matched for valence to majority groups. This less extreme rating for minority groups in the context of changing ratings is consistent with a slower learning because of the less frequent provision of information (Murphy et al., 2011).

Three factors may explain the ventral striatal and ACC/PFC responses to minority groups. The first factor is the potential slowed learning, which may have meant that, over the course of the scanning session, the striatum was more consistently engaged by the minority groups because of its putative role in providing an associative learning signal (Schultz, Dayan, & Montague, 1997). However, it is notable that the prediction error values in our model did not significantly correlate with the ventral striatum response profile. The second factor is the pure stimulus novelty, with the minority group names appearing less frequently on the screen than the majority names, independent of any information provided by the descriptions of behavior. However, in our control task, we found no evidence that a simple difference in the frequency of encountering members of the minority and majority control groups (in the absence of descriptions of their behaviors) led to an increased activity in these brain regions, even at liberal uncorrected thresholds. Although pure stimulus novelty for group names seems unlikely to have accounted for the neural responses to the minority groups, we cannot rule out the possibility of an interaction between stimulus novelty and learning. Finally, it has been argued that attention shifts from the majority to minority groups as learning proceeds (Sherman et al., 2009). Thus, it is possible that the shift in attention to the minority groups also interacts with the other factors to result in the greater evoked ventral striatal responses to the minority groups.

Although we had predicted that our participants would show slower learning of the valence of the minority groups, mirroring past work on the "illusory correlation" effect (Murphy et al., 2011; Hamilton & Gifford, 1976), we did not predict that learning would be more extensive over the learning blocks for the negative groups than the positive groups. Indeed, previous studies have found pronounced differences between the minority and majority groups using positively valenced stimuli (Murphy et al., 2011; Hamilton & Gifford, 1976). The more extensive learning we observed for the negative groups may have occurred because negative behaviors were, in general, more unusual (counter normative) than positive behaviors, making them more salient and memorable. Furthermore, evidence suggests that negative behaviors are judged as more diagnostic of a person's true character than positive behaviors (Cone & Ferguson, 2015), which would privilege negative behaviors for learning. It will be useful in the future research to determine if, when absolute valence and familiarity of the behaviors are matched, there persists a difference in the learning

about the valence of negative and positive groups. Similarly, such research would be useful to clarify if the saliency underlies the anterior cingulate response observed in the comparison of reading negative behaviors compared with reading positive behaviors.

Because participants reported changes in their liking of the different groups, these findings relate most closely to the development of prejudice (Fiske, 1998). That said, the use of an explicit report measure means that we cannot be sure if these attitudes involved an affective commitment ("I do not like Group A") or were purely cognitive/conceptual ("I know that members of Group A perform negative behaviors, so I will give Group A a low rating of likeability"). However, previous work has used an implicit measure of evaluation to demonstrate that the kind of learning procedure used in the current experiment does produce changes in affect regarding the different groups (Le Pelley, Calvini, & Spears, 2013). The implication is that the attitudes formed in the current study may well have had an affective component, although it remains for future research to confirm this.

Taken together, our data provide support for the role of the anterior temporal lobe in the formation of prejudicial intergroup attitudes and the PFC and temporal cortex in detecting violations of these attitudes. In future work, it will be important to separate familiarity and imageability from valence in the stimuli used for learning. This will help determine whether the results reported here are specific to learning about the valence of social groups or whether familiarity and imageability also contribute to the neural responses. It will also be useful to explore whether it is possible to dissociate the response of brain regions involved in updating stored representations, from those detecting the attitude violation.

Acknowledgments

We thank Marty Sereno for help in optimizing the scanning sequences. This research was supported by a Wellcome Trust Advanced Training Fellowship and a James S. McDonnell Foundation Scholar Award to H. J. S., a Wellcome Trust Senior Investigator Award (WT106931MA) to B. C. L., and start-up costs provided by the Birkbeck-UCL Centre for Neuroimaging to H. J. S. and R. A. M. This work was partially supported by the Leverhulme Trust (Grant RPG-2014-075), the NIH (Grant 1P01HD080679), and a Wellcome Trust Investigator Award (Grant WT106931MA) to BCL.

Reprint requests should be sent to Hugo J. Spiers, Division of Psychology and Language Sciences, Department of Experimental Psychology, UCL Institute of Behavioural Neuroscience, University College London, London WC1H 0AP, United Kingdom, or via e-mail: h.spiers@ucl.ac.uk.

REFERENCES

Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews Neuroscience, 15,* 670–682.

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

Amodio, D. M., & Lieberman, M. D. (2009). Pictures in our heads: Contributions of fMRI to the study of prejudice and stereotyping. In T. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 347–366). New York: Erlbaum.

Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature, 456,* 245–249.

Berndsen, M., Spears, R., van der Pligt, J., & McGarty, C. (2002). Illusory correlation and stereotype formation: Making sense of group differences and cognitive biases. In C. McGarty, V. Y. Yzerbyt, & R. Spears (Eds.), *Stereotypes as explanations: The formation of meaningful beliefs about social groups* (pp. 90–110). Cambridge: Cambridge University Press.

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

Brunet, E., Sarfati, Y., Hardy-Baylé, M. C., & Decety, J. (2000). A PET investigation of the attribution of intentions with a nonverbal task. *Neuroimage, 11,* 157–166.

Bucher, K., Myers, R. E., & Southwick, C. (1970). Anterior temporal cortex and maternal behavior in monkey. *Neurology, 20,* 415.

Calarge, C., Andreasen, N. C., & O'Leary, D. S. (2003). Visualizing how one brain understands another: A PET study of theory of mind. *American Journal of Psychiatry, 160,* 1954–1964.

Castelli, F., Happe, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage, 12,* 314–325.

Chadwick, M. J., Anjum, R. S., Kumaran, D., Schacter, D. L., Spiers, H. J., & Hassabis, D. (2016). Semantic representations in the temporal pole predict false memories. *Proceedings of the National Academy of Sciences, U.S.A., 113,* 10180–10185.

Cloutier, J., Gabrieli, J. D., O'Young, D., & Ambady, N. (2011). An fMRI study of violations of social expectations: When people are not who we expect them to be. *Neuroimage, 57,* 583–588.

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology, 108,* 37–57.

Contreras, J. M., Banaji, M. R., & Mitchell, J. P. (2011). Dissociable neural correlates of stereotypes and other forms of semantic knowledge. *Social Cognitive and Affective Neuroscience, 7,* 764–770.

Cunningham, W., Raye, C., & Johnson, M. (2004). Implicit and explicit evaluation: fMRI correlates of valence, emotional intensity, and control in the processing of attitudes. *Journal of Cognitive Neuroscience, 16,* 1717–1729.

Eickhoff, S., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., et al. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage, 25,* 1325–1335.

Fazio, R. H., & Olson, M. A. (2003). Attitudes: Foundations, functions, and consequences. In M. Hogg & J. Cooper (Eds.), *The Sage handbook of social psychology* (pp. 139–160). London: Sage.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., Vol. 2, pp. 357–411). New York: McGraw-Hill.

Franzen, E. A., & Myers, R. E. (1973). Neural control of social behavior: Prefrontal and anterior temporal cortex. *Neuropsychologia, 11,* 141–157.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50,* 531–534.

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

Gallagher, H. L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *Neuroimage, 16,* 814–821.

German, T. P., Niehaus, J. L., Roarty, M. P., Giesbrecht, B., & Miller, M. B. (2004). Neural correlates of detecting pretense: Automatic engagement of the intentional stance under covert conditions. *Journal of Cognitive Neuroscience, 16,* 1805–1817.

Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology, 60,* 509.

Gilbert, S. J., Swencionis, J. K., & Amodio, D. M. (2012). Evaluative vs. trait representation in intergroup social judgments: Distinct roles of anterior temporal lobe and prefrontal cortex. *Neuropsychologia, 50,* 3600–3611.

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.

Grèzes, J., Frith, C. D., & Passingham, R. E. (2004). Inferring false beliefs from the actions of oneself and others: An fMRI study. *Neuroimage, 21,* 744–750.

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nature Neuroscience, 18,* 1233–1235.

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology, 12,* 392–407.

Harris, L. T., & Fiske, S. T. (2010). Neural regions that underlie reinforcement learning are also active for social expectancy violations. *Social Neuroscience, 5,* 76–91.

Hilton, J. L., & Von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology, 47,* 237–271.

Iacoboni, M., Lieberman, M. D., Knowlton, B. J., Molnar-Szakacs, I., Moritz, M., Throop, C. J., et al. (2004). Watching social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases compared to a resting baseline. *Neuroimage, 21,* 1167–1173.

Katz, D., & Braly, K. (1933). Racial stereotypes of one hundred college students. *Journal of Abnormal and Social Psychology, 28,* 280–290.

Kling, A., & Steklis, H. D. (1976). A neural substrate for affiliative behavior in nonhuman primates. *Brain, Behavior and Evolution, 13,* 216–238.

Kling, A. S., Tachiki, K., & Lloyd, R. (1993). Neurochemical correlates of the Klüver–Bucy syndrome by in vivo microdialysis in monkey. *Behavioural Brain Research, 56,* 161–170.

Kluver, H., & Bucy, P. (1939). Preliminary analysis of functions of the temporal lobes in monkeys. *Archives of Neurology and Psychiatry, 41,* 979–1000.

Kunda, Z. (1999). *Social cognition: Making sense of people*. Cambridge, MA: MIT Press.

Kutzner, F. L., & Fiedler, K. (2015). No correlation, no evidence for attention shift in category learning: Different mechanisms behind illusory correlations and the inverse base-rate effect. *Journal of Experimental Psychology: General, 144,* 58–75.

Le Pelley, M. E., Calvini, G., & Spears, R. (2013). Learned predictiveness influences automatic evaluations in human contingency learning. *Quarterly Journal of Experimental Psychology, 66,* 217–228.

Le Pelley, M. E., Reimers, S. J., Calvini, G., Spears, R., Beesley, T., & Murphy, R. A. (2010). Stereotype formation: Biased by association. *Journal of Experimental Psychology: General, 139,* 138–161.

Løvstad, M., Funderud, I., Lindgren, M., Endestad, T., Due-Tønnessen, P., Meling, T., et al. (2012). Contribution of subregions of human frontal cortex to novelty processing. *Journal of Cognitive Neuroscience, 24,* 378–395.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology, 51,* 93–120.

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2012). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience, 8,* 623–631.

Miller, B. L., Darby, A. L., Benson, D. F., Cummings, J. L., & Miller, M. H. (1997). Aggressive, socially disruptive and antisocial behavior associated with frontotemporal dementia. *British Journal of Psychiatry, 170,* 150–155.

Miller, B. L., Darby, A. L., Swartz, J. R., Yener, G. G., & Mena, I. (1995). Dietary changes, compulsions, and sexual behavior in frontotemporal degeneration. *Dementia, 6,* 195–199.

Mitchell, J. P., Ames, D. L., Jenkins, A. C., & Banaji, M. R. (2009). Neural correlates of stereotype application. *Journal of Cognitive Neuroscience, 21,* 594–604.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron, 50,* 655–663.

Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourão-Miranda, J., Andreiuolo, P. A., et al. (2002). The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. *Journal of Neuroscience, 22,* 2730–2736.

Murphy, R. A., Schmeer, S., Vallée-Tourangeau, F., Mondragon, E., & Hilton, D. (2011). Making the illusory correlation effect appear and then disappear: The effects of increased learning. *Quarterly Journal of Experimental Psychology, 64,* 24–40.

Mychack, P., Kramer, J. H., Boone, K. B., & Miller, B. L. (2001). The influence of right frontotemporal dysfunction on social behavior in frontotemporal dementia. *Neurology, 56,* 11–15.

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology, 14,* 769–776.

Ohnishi, T., Moriguchi, Y., Matsuda, H., Mori, T., Hirakata, M., Imabayashi, E., et al. (2004). The neural network for the mirror system and mentalizing in normally developed children: An fMRI study. *NeuroReport, 15,* 1483–1487.

Olson, I. R., McCoy, D., Klobusicky, E., & Ross, L. A. (2013). Social cognition and the anterior temporal lobes: A review and theoretical framework. *Social Cognitive and Affective Neuroscience, 8,* 123–133.

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.

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin, 32,* 421–433.

O'Reilly, J. X., Schüffelgen, U., Cuell, S. F., Behrens, T. E., Mars, R. B., & Rushworth, M. F. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences, U.S.A., 110,* 3660–3669.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience, 8,* 976–987.

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.

Powers, K. E., Somerville, L. H., Kelley, W. M., & Heatherton, T. F. (2016). Striatal associative learning signals are tuned to in-groups. *Journal of Cognitive Neuroscience, 28,* 1243–1254.

Quadflieg, S., Flannigan, N., Waiter, G. D., Rossion, B., Wig, G. S., Turk, D. J., et al. (2011). Stereotype-based modulation of person perception. *Neuroimage, 57,* 549–557.

Quadflieg, S., & Macrae, C. N. (2011). Stereotypes and stereotyping: What's the brain got to do with it? *European Review of Social Psychology, 22,* 215–273.

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.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II: Current Research and Theory* (Vol. 2, pp. 64–99).

Saxe, R., & Powell, L. J. (2006). It's the thought that counts specific brain regions for one component of theory of mind. *Psychological Science, 17,* 692–699.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275,* 1593–1599.

Sherman, J. W., Kruschke, J. K., Sherman, S. J., Percy, E. J., Petrocelli, J. V., & Conrey, F. R. (2009). Attentional processes in stereotype formation: A common model for category accentuation and illusory correlation. *Journal of Personality and Social Psychology, 96,* 305–323.

Spiers, H. J., & Maguire, E. A. (2006). Spontaneous mentalizing during an interactive real world task: An fMRI study. *Neuropsychologia, 44,* 1674–1682.

Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., et al. (2012). Learning to simulate others' decisions. *Neuron, 74,* 1125–1137.

Thompson, S. A., Patterson, K., & Hodges, J. R. (2003). Left/right asymmetry of atrophy in semantic dementia: Behavioral–cognitive implications. *Neurology, 61,* 1196–1203.

Van Rooy, D., Van Overwalle, F., Vanhoomissen, T., Labiouse, C., & French, R. (2003). A recurrent connectionist model of group biases. *Psychological Review, 110,* 536–563.

Visser, M., Jefferies, E., & Ralph, M. L. (2010). Semantic processing in the anterior temporal lobes: A meta-analysis of the functional neuroimaging literature. *Journal of Cognitive Neuroscience, 22,* 1083–1094.

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., et al. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *Neuroimage, 14,* 170–181.

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

Widrow, B., & Hoff, M. E. (1988). Adaptive switching circuits. In J. A. Anderson & E. Rosenfeld (Eds.), *Neurocomputing: Foundations of research* (pp. 123–134). Cambridge, MA: MIT Press.

Wong, C., & Gallate, J. (2012). The function of the anterior temporal lobe: A review of the empirical evidence. *Brain Research, 1449,* 94–116.

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, U.S.A., 104,* 6430–6435.

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

Zaki, J., Kallman, S., Wimmer, G. E., Ochsner, K., & Shohamy, D. (2016). Social cognition as reinforcement learning: Feedback modulates emotion inference. *Journal of Cognitive Neuroscience, 28,* 1270–1282.