PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2012) **367**, 744–753
doi:10.1098/rstb.2011.0300

Research

Race and reputation: perceived racial group trustworthiness influences the neural correlates of trust decisions

Damian A. Stanley[1], **Peter Sokol-Hessner**[1], **Dominic S. Fareri**[2], **Michael T. Perino**[3], **Mauricio R. Delgado**[2], **Mahzarin R. Banaji**[4] **and Elizabeth A. Phelps**[5,6,*]

[1]*Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, USA*
[2]*Department of Psychology, Rutgers University, Newark, NJ 07102, USA*
[3]*Center for Autism Research, Children's Hospital of Philadelphia, Philadelphia, PA 19146, USA*
[4]*Department of Psychology, Harvard University, Cambridge, MA 02138, USA*
[5]*Department of Psychology, and* [6]*Center for Neural Science, New York University, New York, NY 10003, USA*

Decisions to trust people with whom we have no personal history can be based on their social reputation—a product of what we can observe about them (their appearance, social group membership, etc.)—and our own beliefs. The striatum and amygdala have been identified as regions of the brain involved in trust decisions and trustworthiness estimation, respectively. However, it is unknown whether social reputation based on group membership modulates the involvement of these regions during trust decisions. To investigate this, we examined blood-oxygenation-level-dependent (BOLD) activity while participants completed a series of single-shot trust game interactions with real partners of varying races. At the time of choice, baseline BOLD responses in the striatum correlated with individuals' trust bias—that is, the overall disparity in decisions to trust Black versus White partners. BOLD signal in the striatum was higher when deciding to trust partners from the race group that the individual participant considered less trustworthy overall. In contrast, activation of the amygdala showed greater BOLD responses to Black versus White partners that scaled with the amount invested. These results suggest that the amygdala may represent emotionally relevant social group information as a subset of the general detection function it serves, whereas the striatum is involved in representing race-based reputations that shape trust decisions.

Keywords: trust game; race bias; reputation; functional magnetic resonance imaging; decision-making

1. INTRODUCTION

Trust is ubiquitous and critical in human social interactions [1]. Few social situations are devoid of decisions about whom to trust because such decisions are inherent to safe and beneficial outcomes for the individual and their groups. In situations in which there exists a history based on previous outcomes with a potential 'trustee', we can rely upon the firm social basis of their reputation to guide our decisions. In the absence of any prior experience, how are trust decisions to be made? From the less certain evidence that is available, one must make decisions based on a subjective estimate of the trustee's reputation, given what we can observe about them in the moment or have learned about them from other sources (i.e. their appearance, group memberships or social reputation) [2–5]. In this study, we were interested in how individuals' perception of their partner's race, as an *a priori* proxy for reputation, might influence the neural systems involved in such trust decisions.

Characterizing the psychological and neural underpinnings of trust decisions has been the focus of a rapidly growing body of research over the past few decades. Many of these studies have used a behavioural economics paradigm called the 'trust game' [6,7]. In the trust game, there are two rounds of monetary exchange between two partners (an investor and a trustee). In the first round, the investor is given a monetary sum and can send all or a portion of it to the trustee. This sent amount, which is considered to be a measure of how much the investor trusts the trustee, is multiplied, typically three or four times, before it is received by the trustee. In the second and final round,

* Author for correspondence (liz.phelps@nyu.edu).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2011.0300 or via http://rstb.royalsocietypublishing.org.

One contribution of 12 to a Theme Issue 'The biology of cultural conflict'.

the trustee can reciprocate by sharing all, a portion, or none of their holdings with the investor. Studies examining the trust game have shown that even in anonymous, single-shot (i.e. one time only) interactions, the majority of people are trusting and trustworthy, both sending money and reciprocating [6,7].

Brain imaging studies of the trust game have consistently found decision-related activity in the striatum [3,8,9]. Specifically, blood-oxygenation-level-dependent (BOLD) signal in the striatum is increased for decisions when investors choose to trust their partner relative to when they choose to keep their money [3,8]. Furthermore, King-Casas *et al.* [8] found that as experience-based reputations were formed over the course of 10 repeated interactions, activity in the caudate nucleus of the striatum shifted temporally, from the moment when investors received the outcome to the moment of decision. The authors suggested that this pattern of activity is akin to that observed in reinforcement learning, in which prediction errors at outcome initially serve to update the value of a predictive cue during learning, but eventually shift to the time of the cue itself once learning has occurred [10–12]. In the context of a trust interaction, the participants' anticipatory response at the time of decision began to reflect their expectation of reciprocity, or in other words, their partner's reputation based on previous interactions. These data suggest that the striatum might be a candidate for a brain region that represents social reputation-related information in trust decisions.

Notably, the literature on trust decisions has generally focused on ensuring partners' anonymity and removing any social information about them in the trust game interactions. This has simplified the initial conditions of the trust interaction, enabling researchers to establish basic behavioural and neural models of trust decisions. However, it is exceedingly rare that we enter into a trust interaction that is completely anonymous. In the situations with the least information, we may have little knowledge about the individual actor, but we are presented with social category knowledge such as where the person may be from, their physical appearance, etc. Under such conditions of social uncertainty, we rely on a combination of observable social cues (e.g. group membership) and our own beliefs about such cues to generate an estimate of trustworthiness. A number of behavioural studies [2,4] have shown that social group cues, such as a partner's gender or ethnicity, do indeed shape trust decisions in the context of the trust game. Building on those group-level findings and taking into account individual differences, Stanley *et al.* [5] demonstrated that individuals' trust game decisions were based on a combination of their partner's race (Black or White) and their own implicit race attitudes (as measured by the implicit association test [13]).

To date, we are aware of only one study that has examined how reputation information external to the trust interaction itself (e.g. aspects of the trustee independent of their trust decisions) can modulate the brain signals associated with those interactions. Specifically, Delgado *et al.* [3] examined how a partner's perceived moral character influenced BOLD activity in the striatum during trust game decisions. Investor participants interacted with trustee partners of different moral character ('good', 'bad' or 'neutral' established with vignettes presented prior to data collection) in a repeated trust game. Despite equal reinforcement rates across all partners, participants persistently trusted the partners previously identified as 'good', showing evidence of reliance on reputation. Similar to King-Casas *et al.* [8], Delgado *et al.* [3] found that having a reputation, in this case based on moral character, diminished BOLD responses to the outcome of the trust decision. At the time of decision, BOLD activity in the striatum also varied depending on the morality of the partner, in effect reflecting their partner's reputation. Specifically, differential BOLD activity at the time of decision was greater when participants interacted with 'bad' partner relative to the 'good' partner. This finding further supports the hypothesis that the striatum is involved in the representation of reputation at the time of decision, irrespective of whether it is built on experience [8] or constructed from external information (i.e. vignettes [3]).

In contrast to neuroeconomic studies of trust decisions as assessed with the trust game, studies examining the neural systems mediating subjective judgements of trustworthiness have highlighted a role for the amygdala. The amygdala has been consistently implicated in both explicit and implicit assessments of trustworthiness [14–18]. Specifically, BOLD responses in the amygdala are greater to faces judged to be untrustworthy [15], and patients with amygdala damage generally rate faces deemed untrustworthy by neurologically intact participants as more trustworthy overall [14]. A recent study examining decisions in a trust game found that patients with amygdala damage were more likely to decide to trust partners in the face of betrayal [19]. These results are consistent with a larger literature suggesting the amygdala signals stimuli that represent potential threats [20]—in this case, the threat of untrustworthiness. Interestingly, a number of studies have also shown that BOLD activity in the amygdala reflects individuals' attitudes towards race groups, which is hypothesized to reflect the potential threat posed by race outgroup members ([21–24]; see [25] for review). In light of the amygdala's involvement in estimations of trust, these results suggest a possible neural substrate for race-based reputations related to trust. A recent finding that implicit race attitudes correlate with trust decisions provides some behavioural support for this hypothesis [5].

The findings that reputations related to trust are linked to the striatum [3,8], whereas estimations of trustworthiness appear to be dependent on the amygdala [14–18] suggest that these two regions may interact during trust decisions, supporting the final decision in unique ways. Research in non-human animals examining the interaction of the amygdala and striatum suggests that the amygdala may represent the threat value of a stimulus, but its projections to the striatum are critical when this threat stimulus leads to a decision to act [26]. This finding and others (see [27] for review) are consistent with a larger literature suggesting that the striatum is a site

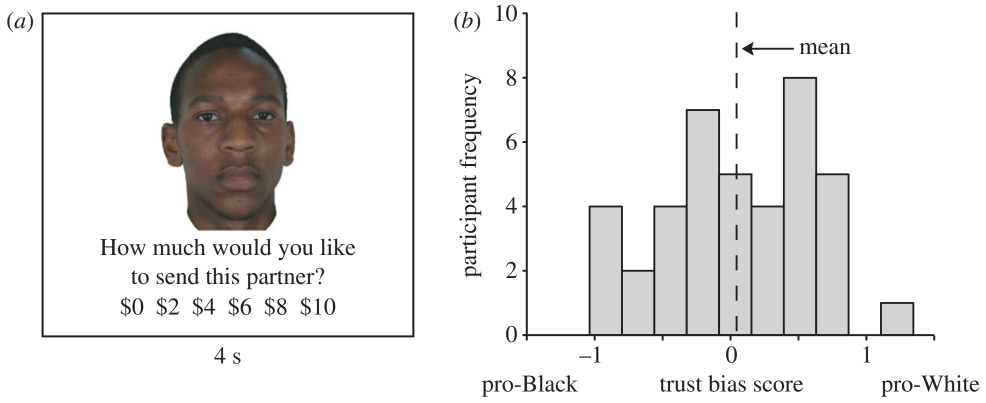

Figure 1. (*a*) Schematic of a single fMRI decision trial. Each partner's face was displayed for 4 s, during which time the participant could enter a monetary offer to send to that partner ($0 to $10, in increments of $2). Trials were separated from each other by 2–10 s of fixation. (*b*) Histogram of participants' trust bias scores (mean White offer − mean Black offer)/(standard deviation of all offers). Negative scores indicate greater mean offers to Black than White partners, and positive scores vice versa. The dotted line indicates the mean trust bias score across participants (0.04 ± 0.09).

where action and motivation are integrated during decision-making [28]. In the framework of the present study and the previous literature, we might expect the amygdala to represent race group information [21], but the striatum to integrate this information in the representation of social group reputations that underlie decisions to trust [3,8].

In order to explore how race-based reputations may alter the neural systems of trust decisions, brain-imaging participants made a series of single-shot trust game decisions with actual monetary consequences and real trustee partners whose primary distinguishing feature was their race (figure 1*a*). Given that we were specifically interested in the influence of race group on decisions to trust that were unaffected by other factors, we focused on the decision phase and did not provide outcome feedback until all decisions were complete. Because attitudes towards race groups vary widely from individual to individual, we used an individual difference approach to explore the neural systems mediating the influence of race on trust decisions. Specifically, we calculated a trust bias score for each participant, which reflected the overall monetary disparity in decisions to trust Black versus White partners. We used this trust bias score as an index of the reputation our investor participants ascribed to race groups. We were specifically interested in how the combination of partner race (Black versus White) and individual differences in trust bias, as reflected in participants' overall choices, would be reflected in the brain systems mediating trust decisions.

2. METHODS

(a) *Participants*

We collected functional magnetic resonance imaging (fMRI) data from 51 participants, 11 of whom were excluded from the final analysis (three for excessive head motion, two due to technical problems, six because they did not meet our behavioural inclusion criterion; see below and electronic supplementary material, figure S1). The remaining 40 participants (22 females and 18 males; ages 18–33, mean age = 20.2 years; 16 White, 13 Asian, 3 Black, 3 Indian, 3 Multiracial, 1 Hispanic and 1 Middle Eastern) were right-handed, had lived in the United States for more than 10 years, spoke English as their primary language and had normal or corrected-to-normal vision. All participants were paid $25 per hour in addition to their earnings from the experiment.

(b) *Procedure*

The basic procedure required participants to make a series of single-shot trust decisions with real trustee partners who were represented by facial photographs. To assure that these interactions were real, we had previously assembled a database of photos and trust game decisions obtained from participants at Rutgers University in Newark, New Jersey to serve as the 'trustee' partners in future trust game studies. For each potential trustee partner, we collected 'share' (50/50) or 'keep' decisions for all possible offers ($1–10 sent, corresponding to $4–40 received; presented in random order) from a future 'investor' in a modified trust game. After providing their decisions, they posed for a photograph (front-facing, neutral expression) and provided contact information so that they could receive possible future payments from our participants by mail. Potential trustee partners earned $5 for their participation, in addition to any future earnings from the trust game task. Data were collected using E-Prime (Psychology Software Tools, Pittsburgh, PA, USA). From this database, we selected 175 male trustee partners for the current study.

Prior to the imaging study, we prescreened potential participants. In a previous behavioural study using a similar task [5], we observed that a subset of participants based their decisions not on the individual characteristics of the partners, but instead according to a fixed rule. We distinguished those participants by investigating the number of consecutive interactions in which they sent the same amount. To reduce the exclusion rate of the fMRI study, we prescreened 176 participants in the New York University Centre for Experimental Social Science. Participants were endowed with $10, and then made single-shot trust decisions with 25 real partners. The experimental design was the same as in the fMRI experiment (described in §2*c*) with the exception of fewer trials and only one trial (instead of three) paid for each participant. Participants with 15 or more consecutive offers were not invited to participate in the fMRI study. Participants who passed this initial prescreening criterion and the medical screening were invited to participate in the fMRI experiment.

Prior to the fMRI task, participants were endowed with $30 and told that it was now 'their money to decide what to do with'. They were then instructed about the details of the trust game and completed a quiz to ensure comprehension. Following this, participants were placed in an MRI scanner. While the anatomical MRI data were collected, participants completed 100 trials of response practice (entering a specific dollar amount on each trial, no faces shown), followed by 20 practice trust game interactions with pretend partners.

During fMRI, participants made single-shot real trust game decisions to invest with 150 unique partners (50 Black, 50 White and 50 Others[1]). For each interaction, participants saw a picture of their partner's face and had 4 s to decide how much money ($0–10 in increments of $2) they wanted to send to that partner. Participants knew that the amount they sent would be quadrupled and then allocated based on their partner's decision. Stimuli were presented in a rapid event-related design. On each trust game trial, a colour photograph of the partner's face was presented (4 s) with the question 'How much would you like to share with this partner?' (figure 1*a*). Underneath the partner's face were dollar amounts ($0–10 in increments of $2) ascending from left to right. Participants entered their response using two 5-button boxes and a two-step selection process. First they indicated low ($0 and $2), middle ($4 and $6) or high ($8 and $10) with their right index, middle and ring fingers, respectively. Then they selected between the remaining two options (e.g. the lower or higher value in the selected pair) using their left middle and index fingers, respectively. A small dot above the amounts displayed on the screen indicated the current selection. Participants were asked to respond to every trial and to always select a specific amount. There was no penalty for not making a response other than the fact that the interaction would not count. Responses could be entered and changed at any time before the end of the trial. Each trial was followed by an inter-trial interval (2–10 s, in increments of 2 s, in a decreasing exponential distribution; randomly ordered). Trial order was randomized with the following constraints: each scan had 16 or 17 partners from each race group (Black, White, Others; summing to 50 interactions total) and no more than three consecutive interactions with partners of the same race. Participants were instructed to keep their eyes pointed towards a fixation cross (present throughout the scan) to control for eye movements. Visual stimuli were presented using PsychToolbox [29,30] and projected onto a rear-projection screen that participants viewed in a mirror mounted on the scanner.

Following the decision task, participants were given the outcome of each decision in a single session. The real partners drawn from our Rutgers database had previously indicated whether they wanted to 'share' (split 50/50) or 'keep' (all) of the money they received for each possible investor offer. After the experiment was over, three interactions were randomly selected and the outcomes paid to the participant (in person) and to their trustee partners (by mail).

(c) *Functional magnetic resonance imaging acquisition*

Imaging was conducted at the NYU Center for Brain Imaging, using a 3 tesla Siemens Allegra head-only scanner and a Nova Medical head coil (transmitter/receiver, model NM011). Scanning sessions began with an MPRAGE anatomical scan with 176 T1-weighted slices collected in the sagittal plane (repetition time (TR) = 2.5 s, echo time (TE) = 4.38 ms, flip angle = 8°, slice thickness = 1 mm, in-plane resolution = 1 × 1 mm, field of view (FOV) = 256 mm^2). Following this, functional T2*-weighted images (TR = 2 s, TE = 25 ms, flip angle = 80°, slice thickness = 3 mm, in-plane resolution = 3 × 3 mm, field of view (FOV) = 192 mm^2) were acquired. We collected 40 slices oriented parallel to the anterior and posterior commissures and covering the ventral temporal lobe. Slice acquisition order was interleaved, and the first two acquisitions of each functional sequence were discarded. Foam and a SecureVac Immobilization System (Bionix Radiation Therapy) were used to minimize head motion. Responses were collected using two Rowland USB 5-button boxes.

(d) *Behavioural analysis*

As a second level of behavioural screening, we again counted the number of decisions for which each fMRI participant sent the same amount on consecutive interactions [5]. Participants who made identical consecutive offers on more than 60 per cent of the interactions were excluded from further analysis (electronic supplementary material, figure S1). The remaining participants rarely missed a response (median number of missed trials = 2/150, maximum = 15/150; mean reaction time = 2.12 s ± 0.5 (s.e.)). For each participant, we calculated a behavioural trust bias score: trust bias = (mean White offer − mean Black offer)/(standard deviation of all offers).

(e) *Functional magnetic resonance imaging analysis*

fMRI data were preprocessed and analysed using SPM8 (Wellcome Trust Centre for Neuroimaging, University College London, UK). Each image in each functional run was first temporally corrected for slice acquisition time, and then all images in each run were realigned to the first image of that run using an affine transformation (three-dimensional motion correction). Following this, data were normalized to the Montreal Neurological Institute's standard EPI template (including voxel-size resampling to 2 × 2 × 2 mm) and then spatially smoothed with a three-dimensional Gaussian filter (6-mm full width at half maximum). Finally, a high-pass temporal filter (width = 128 s) was applied to the data.

Each participant's data were fit with a general linear model corrected for serial autocorrelations (AR(1) + w). The model contained three main effect regressors: All (decisions collapsed across all three racial groups), Black (decisions with Black partners only) and White (decisions with White partners only). These regressors were modelled as boxcar functions, with their duration set to participants' reaction times. Parametric modulators of the amount sent for each interaction were included for each of the main effect regressors. In addition, the model included constants for each functional run and the motion correction estimates as regressors of no interest (three translations and three rotations). First-level single-subject contrasts were calculated for the main effect of Black > White trials and the parametric modulators of amount sent to Black > amount sent to White. Second-level covariate analyses were performed on contrasts of interest using a random effects (participant) general linear model with trust bias as a participant-level covariate. Second-level group contrasts (i.e. those without a covariate) were calculated using a one-sample, two-tailed *t*-test of the first-level contrast beta weights.

To further explore the pattern of BOLD response in the striatum and amygdala, we conducted region of interest (ROI) analyses. Using the automated anatomical labelling (AAL) atlas [31], we defined the right and left caudate nucleus, putamen and amygdala. For each participant, the average beta value from the first-level whole-brain estimates in each anatomical ROI was calculated for contrasts of interest (Black > White, amount sent, amount sent to Black > amount sent to White). Statistical tests were performed on these beta values for group-level effects (one-sample, two-tailed *t*-test), and for covariance with trust bias (robust-fit regression and Pearson's correlation). Pearson's correlation values are reported only if the robust regression analysis was also significant ($p < 0.05$) or trending ($p < 0.10$).

In addition to these primary analyses, we also conducted two exploratory, supplementary analyses. The first was an ROI analysis examining BOLD responses in the ventromedial prefrontal cortex (vmPFC)—a region previously implicated in the representation of value ([32,33]; see [34] for review), and the second examining the influence of participant race (see electronic supplementary material for more details about these analyses and results).

3. RESULTS

(a) *Behaviour*

Each of the fMRI participants made trust game decisions to send \$0–10 (in increments of \$2) to 150 real partners (50 Black, 50 White and 50 Others). The partners received four times the amount sent, and had previously decided to 'share' (50/50) or 'keep' (all) of the money for each possible amount they might receive. The participants offered a mean of \$4.10 ± 0.29 (s.e.) per trial. There was no group-level significant difference between mean offers to Black (\$4.09 ± 0.30) versus White (\$4.21 ± 0.30) partners, replicating Stanley *et al.* [5].[2] To quantify Black/White trust bias for each participant, we subtracted the mean offer to Black partners from the mean offer to White partners and normalized the difference by the standard deviation of all offers (Black, White and Others) [5]. Positive scores indicate pro-White and negative scores indicate pro-black trust bias. The resulting distribution of trust bias scores (figure 1*b*) was not significantly different from zero (mean = 0.04 ± 0.09, $t_{39} = 0.51$, $p = 0.61$), indicating that there was no overall Black/White trust bias in our sample,[3] again replicating Stanley *et al.* [5].

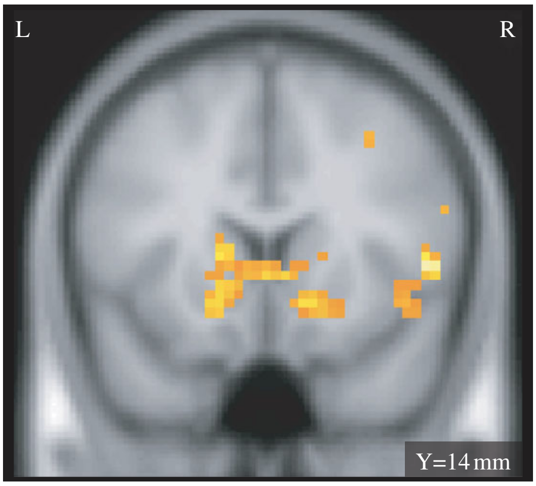

Figure 2. BOLD signal correlates of trust bias in the striatum at decision. A whole-brain voxelwise correlation of the contrast Black > White with individuals' trust bias scores identified positive correlations in the striatum (along with other regions; table 1), indicating greater responses in this area to whichever race group a participant trusted less ($n = 40$, p(cluster) < 0.05 whole brain-corrected (voxelwise $p < 0.005$)).

(b) *Functional magnetic resonance imaging*

We conducted a series of whole-brain correlation analyses to investigate the representation of individuals' trust bias in the BOLD activity linked to trust decisions. To examine the relationship between behavioural bias and BOLD signals, we constructed a general linear model with six regressors. Three main effect regressors (i.e. boxcar functions from the trial onset to response) represented the act of making a decision, one each for All (irrespective of race), Black-only and White-only trials. Correspondingly, three parametric modulators represented the dollar amount sent by the fMRI participant to their partners across All, Black-only and White-only trials. We focused our analysis on two specific contrasts: the main effect contrast of Black > White identified regions in which activity was greater when making decisions with Black versus White partners; the parametric effect contrast of amount sent to Black > amount sent to White identified regions in which the relative difference in amount sent to Black partners elicited a stronger differential BOLD response when compared with the same relative difference in amount sent to White partners. In particular, we were interested in the covariation of those two contrasts with individuals' Black/White trust bias.

The voxelwise correlation of the contrast of the main effect of Black > White trials, with individuals' behavioural Black/White trust bias identified a number of regions, most notably the striatum, as seen in figure 2 (all fMRI results were initially thresholded at $p < 0.005$ uncorrected, and were

Table 1. Cortical regions in which BOLD activity positively correlated with trust bias. Clusters were identified using a whole brain-voxelwise correlation of the two contrasts of interest with individuals' trust bias scores. Columns (left-to-right): cluster significance value (p(clust)), number of 2 × 2 × 2 mm voxels in cluster ($n$vox), MNI coordinates of peak voxel (pk vox MNI), peak voxel AAL atlas label (peak vox region; number of voxels in parentheses), other AAL regions greater than five voxels in cluster ($n = 40$, p(cluster) < 0.05 whole brain-corrected (voxelwise $p < 0.005$)). No clusters were identified that correlated negatively with trust bias at this threshold.

		pk vox MNI				
p(clust)	nVox	x	y	z	peak vox region ($n$vox)	other AAL regions >5 voxels
Black > White (ME)						
<0.001	497	−6	−22	40	cingulum mid L(69)	precuneus L(123), cingulum mid R(72), paracentral lobule R(66), precuneus R(45), cuneus L(32), paracentral lobule L(15)
<0.001	273	−3	41	7	cingulum ant L(94)[a]	cingulum ant R(104), frontal med orb L(11), frontal sup medial L(9), frontal medial orb(8), frontal sup R(7)
<0.001	179	51	−34	34	supramarginal R(87)	temporal sup R(38), angular R(24), parietal inf R(12)
<0.001	152	39	−22	−5	temporal sup R(30)[b]	temporal mid R(21), insula R(18), temporal inf R(16), Heschl R(7), hippocampus R(7)
0.01	86	0	−85	4	calcarine L(58)	calcarine R(11), occipital sup L(11)
0.01	86	42	32	1	frontal inf tri R(29)	frontal mid R(22), frontal inf orb R(12)
0.029	72	51	14	4	frontal inf oper R(33)	frontal inf tri R(16), insula R(15),
0.039	68	39	20	40	frontal mid R(60)	frontal sup R(7)
amount sent to Black > amount sent to White						
<0.001	140	−3	−34	70	paracentral lobule L(54)	paracentral lobule R(33), postcentral L(15), precuneus L(6)
0.003	106	45	−7	31	precentral R(52)	postcentral R(46)

[a] To limit the extent of this cluster to cortex, a mask of medial cortex was used.
[b] The peak voxel of this cluster was in white matter, so the largest contributing AAL region is reported.

subsequently cluster-thresholded at $p < 0.05$, whole brain-corrected). In this area, including portions of the caudate and putamen, individuals had greater activity when deciding about partners from whichever race group they personally trusted less. Additional areas identified by this whole brain correlation included a set of cortical clusters (table 1), many previously implicated in trust decisions and mentalizing about others [35,36]. Importantly, the only significant subcortical clusters of BOLD activity for this analysis were localized to the caudate and putamen. No regions exhibited a negative correlation with trust bias at this threshold.

We then conducted a second, similar whole-brain analysis to examine whether there were regions in which BOLD activity to the parametric contrast of amount sent to Black > amount sent to White correlated with trust bias. In contrast to the findings for the main effect regressors, only two clusters were found to be correlated with trust bias, one centred on the right ventral pre- and post-central gyri, and the other on the paracentral lobule (bilaterally; table 1). No regions exhibited a negative correlation with trust bias at this threshold.

Although the focus of our study was on BOLD activity reflecting individual differences in behavioural Black/White trust bias, we also examined the overall group-level contrasts (e.g. not correlated with trust bias; figure 3 and table 2). The contrast of Black > White identified bilateral regions of the occipital lobe. The contrast of amount sent to Black > amount sent to White identified three clusters. Two bilateral clusters covered the anterior temporal lobe, insula cortex and the amygdala (figure 3), and another covered the precuneus and posterior cingulate cortex. No significant clusters of activity were found for the reverse of either contrast.

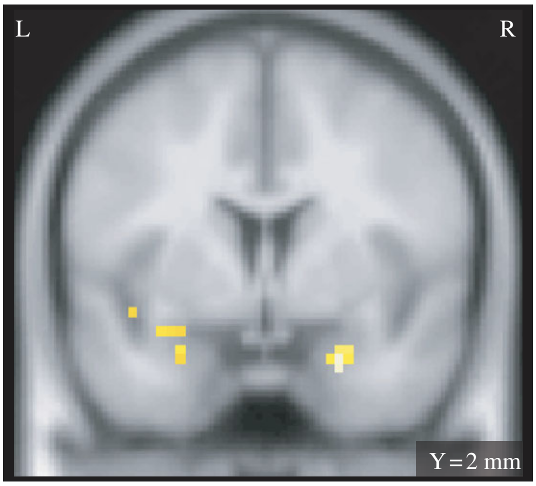

Figure 3. Partner race-related activity in the amygdala at decision. The contrast of amount sent to Black > amount sent to White identified positive clusters in the amygdala (along with other regions; table 2), indicating increased scaling in this region of the representation of the amount sent to Black versus White partners, irrespective of individuals' trust bias ($n = 40$, p(cluster) < 0.05 whole brain-corrected (voxelwise $p < 0.005$)).

To further explore the pattern of BOLD responses in the amygdala and striatum, we conducted a targeted set of analyses using anatomically defined ROIs.

Table 2. Cortical regions in which fMRI activity was influenced by partner race. Clusters were identified using simple group-level contrasts. Columns (left-to-right): cluster significance value (p(clust)), number of voxels in cluster ($n$vox), MNI coordinates of peak voxel (pk vox MNI), peak voxel AAL atlas label (peak vox region; number of voxels in parentheses), other AAL regions greater than five voxels in cluster; $n = 40$, p(cluster) < 0.05 whole brain-corrected (voxelwise $p < 0.005$). No clusters were identified in which activity was higher for White relative to Black partners at this threshold.

		pk vox MNI				
p(clust)	$n$vox	x	y	z	peak vox region (nVox)	other AAL regions >5 vox
Black > White (ME)						
<0.001	203	33	−94	1	occipital mid R(58)	occipital inf R(91), fusiform R(29), cerebellum crus1(15)
<0.001	142	−39	−82	−11	occipital inf L(37)	occipital mid L(55), fusiform L(27)
amount sent to Black > amount sent to White						
<0.001	166	27	11	−29	parahippocampal R(25)	frontal inf orb R(54), temporal pole sup R(41), insula R(23), temporal pole mid R(7)
0.005	159	−42	8	−14	insula L(25)	temporal pole sup L(39), frontal inf orb L(15)
0.023	77	0	−34	22	cingulum mid R(22)[a]	precuneus L(16), precuneus R(12), cingulum post L(7)

[a] The peak voxel of this cluster was in the interhemispheric space, so the largest contributing AAL region is reported instead.

Specifically, we identified the right and left amygdala, caudate and putamen and performed simple contrasts ($p < 0.05$) and correlations with beta values for these regions. All of the specified ROIs, with the exception of the left amygdala, showed evidence of being modulated to some extent by the amount sent ($0–10) to partners on each trial, irrespective of the partner's racial identity. This finding is consistent with previous studies implicating these ROIs in the representation of value. Consistent with our whole-brain analysis, only the striatal response to Black > White was significantly correlated with individuals' trust bias score: right caudate ($r(38) = 0.44$, $p = 0.004$); left caudate ($r(38) = 0.41$, $p = 0.008$); right putamen ($r(38) = 0.34$, $p = 0.031$); and left putamen ($r(38) = 0.29$, $p = 0.072$).

There was no evidence of correlation with individuals' trust bias in the amygdala, though both left and right amygdala showed a significant positive response to amount sent to Black > amount sent to White. In other words, the magnitude of the response to value in the amygdala bilaterally was on average increased when participants were in an interaction with a Black compared with a White partner. The findings from the ROI analysis support and extend the whole-brain contrasts and show that the relative differences in the reputations individuals hold of different race groups are represented at the time of decision in regions of the brain that process value.

4. DISCUSSION

Reputations can be acquired through a number of means, including first-hand experience with previous interactions or other knowledge of specific past actions. However, in the absence of specific relevant information about an individual, a reputation may be derived from more general knowledge, such as social group membership. In this study, we used the trust game to investigate how the influence of race group on reputation modulated BOLD activity in regions of the brain known to be involved in trust decisions and judgements, specifically the striatum and the amygdala. The trustee partners in our study were real people the investor participants had never met, whose faces were presented in photographs during a trust game interaction. As a group, the partners had little to distinguish them from each other in that they were all young adult males with neutral expressions. The primary distinguishing feature was that they varied by race, with two-thirds judged to be either Black or White based on facial features. In the absence of other information, we hypothesized that race group may play a role in the variability observed in decisions to trust [5].

To determine the influence of race on trust reputations, we assessed trust bias, reflecting the overall difference in the amount invested with Black versus White partners for each individual participant. In our sample, trust bias varied among participants with no overall tendency to trust either race group more. However, we found evidence that this race-based measure of reputation correlated with BOLD activity in the striatum during trust decisions, such that BOLD activity was elevated during interactions with members of the race group the individual participant found less trustworthy overall. In contrast, BOLD responses in the amygdala, when scaled by the amount of the investment, were greater during interactions with Black versus White partners, but did not correlate with trust bias. These results suggest that the amygdala may represent race group information, but activity in the striatum is more directly linked to trust decisions and likely reflects the integration of information from multiple sources (including the amygdala) to represent race-based reputations.

The finding that striatal activation correlated with trust bias provides further support for the hypothesis that BOLD activity in the striatum may represent partners' trust reputations at the time of decision [3,8]. The nature of our study is similar to that of Delgado *et al.* [3], in that we examined how reputation based on social information external to the trust interaction itself can influence the decision and the underlying neural circuitry. Although there were several differences in the experimental design between the studies, one similarity in the results is that both studies observed greater relative BOLD activity at the time of decision when the trustee chose to invest with the partner (or group of partners) that was trusted less overall. Delgado *et al.* [3] used a repeated trust game with three fictional partners that varied by moral character, and the investor participants simply decided on each trial if they did or did not want to invest all their money with the partner. The participants trusted the 'bad' partner less often than the 'good' partner, but showed greater differential striatal BOLD activity when they chose to invest with this partner versus keep their money. In our study, the participants made a single investment decision for each real partner and were allowed to invest a range of amounts. We found a shift in the average level of decision-related activity in the striatum based on the partner's race group, which did not scale with the value of the investment. Specifically, striatal responses were higher for all dollar amounts sent to members of the less trusted group compared with the more trusted group, but the relative BOLD difference for sending $2 compared with $10 was not influenced by the partner's race.

One possible interpretation of the pattern of results we find in the striatum within the context of the previous study by Delgado *et al.* [3] is that reputation-related shifts in overall striatal activity may reflect a weighting of the circuitry representing the potential subjective value of the trust decision. In the study by King-Casas *et al.* [8], in which trust reputations were acquired through repeated interactions, they found that once a reputation was acquired, the BOLD response at the time of decision reflected the potential expected outcome of that interaction. It may be that social reputation alters the subjective value of the potential outcome such that less predictable rewards are more valuable. In other words, even though the less trustworthy partner is, by definition, trusted less overall, when a decision is made to trust this partner, the potential profit from money sent is subjectively worth more to the investor than that sent to a more trusted partner. In this framework, the potential reward from sending $2 or $3 to a less-trusted partner is similar in subjective value of sending $3 or $4 to the more trusted partner.

In addition to the striatum, our whole-brain analysis identified a network of cortical regions in which overall BOLD activity for Black versus White partners was correlated with trust bias (table 1). In all these regions, similar to the striatum, BOLD signal was higher when making decisions about partners from the less-trusted race group relative to the more trusted group. Generally speaking, the regions identified are part of a network that has been proposed to underlie mentalizing about others and their intentions (see [35,36] for review). For the sake of brevity, we do not discuss the sizeable literature concerning the function of each of these regions. However, particularly noteworthy are clusters of activity that included the anterior cingulate cortex, found in previous trust game studies to reflect mentalizing about one's partner [8,37], and the right superior temporal sulcus, engaged both during explicit trustworthiness estimations [15] and by social trust prediction errors [38]. The increased activity in this network suggests that trust decisions with partners from a less-trusted race group may have elicited more effortful mentalizing about the partner's intentions compared to interactions with their more trusted counterparts.

Given the extensive literature linking the amygdala to judgements of trustworthiness from facial characteristics [14–18] and its modulation by race group [21–24], it is somewhat surprising that we did not find evidence that BOLD responses in the amygdala reflected trust bias in investor decisions. However, activity in the amygdala was sensitive to components of both the decision and social group membership. The ROI analysis showed that responses in the right amygdala scaled with the amount sent, such that sending larger amounts resulted in greater BOLD signal. Interestingly, when examining overall BOLD responses in the amygdala as modulated by partner race, we did not find a Black versus White difference, but rather a race group difference that was scaled by the amount sent. Specifically, in the amygdala (as well as the insula, anterior temporal lobe and orbitofrontal cortex; table 2), the relative difference in amount sent to Black partners elicited a stronger differential BOLD response compared with the same relative difference in amount sent to White partners. In other words, for each additional dollar sent to a Black partner, activity increased more than it did for each additional dollar sent to a White partner. The decision to send money in the trust game represents both trust in the partner, but also risk. The larger the amount sent, the larger the potential gain, but also the larger the potential loss. One interpretation of these results is that as the potential risk of the decision increased, BOLD signal in the amygdala increased relatively more for Black versus White partners. Our findings indicate that in the context of a trust decision, the amygdala codes a combination of decision variables and race group information.

Interpreting our findings in the broader context of the striatum and amygdala's functional roles suggests a possible conceptual framework. As mentioned earlier, the amygdala and striatum interact when the potential threat or emotional value of a stimulus results in a decision to act [39]. In our task, activity in the amygdala may have reflected an initial, automatic evaluation of a given partner based on salient physical characteristics (e.g. facial characteristics and group membership) combined with the potential value and risk of the decision. This information may have been integrated with one's own beliefs about reputation as represented in cortical regions including those implicated in mentalizing about others. Activity in the striatum, proposed to represent choice value at the time of decision, may reflect a combination of conscious beliefs about partner reputation, automatic evaluations based on race group, and actual choice value. While our data are consistent with this interpretation of the neural systems mediating the influence of race group on reputation, they are not conclusive and we propose this solely as a hypothetical framework for future investigations.

All procedures were approved by the University Committee on Activities Involving Human Subjects at New York University. The collection of photos and responses for the trustee partner database was additionally approved by the Institutional Review Board of Rutgers University.

We thank K. Sanzenbach and the Center for Brain Imaging at New York University for technical assistance. This work was supported by a grant from the Macarthur Foundation Law and Neuroscience Network.

ENDNOTES

[1] The 'Other' category consisted of people from heterogenous racial backgrounds other than Black or White. These partners were included so that participants were unaware of our primary interest in attitudes towards Black and White partners [5].

[2] There also was no significant difference in mean offers to Black versus Other ($4.00 ± 0.30) partners, but there was a significant difference between mean offers to White versus Other partners (paired $t_{39} = 2.08$, $p = 0.049$). Because of our focus on Black/White trust bias as well as the heterogeneity of the 'Other' partners, we do not report further on trials featuring 'Other' partners.

[3] Of our individual participants, 22 of 40 had trust bias scores that were significantly different from 0 (8 pro-Black, 14 pro-White; two-sample *t*-tests on amount sent to Black versus White partners), while 18 of 40 showed no significant trust bias towards either racial group.

REFERENCES

1 Coleman, J. S. 1990 *Foundations of social theory*. Cambridge, MA: Harvard University Press.

2 Fershtman, C. & Gneezy, U. 2001 Discrimination in a segmented society: an experimental approach. *Q. J. Econ.* **116**, 351–377. (doi:10.1162/003355301 556338)

3 Delgado, M. R., Frank, R. H. & Phelps, E. A. 2005 Perceptions of moral character modulate the neural systems of reward during the trust game. *Nat. Neurosci.* **8**, 1611–1618. (doi:10.1038/nn1575)

4 Slonim, R. & Guillen, P. 2010 Gender selection discrimination: evidence from a trust game. *J. Econ. Behav. Organ.* **76**, 385–405. (doi:10.1016/j.jebo.2010.06.016)

5 Stanley, D. A., Sokol-Hessner, P., Banaji, M. R. & Phelps, E. A. 2011 Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proc. Natl Acad. Sci. USA* **108**, 7710–7715. (doi:10.1073/pnas.1014345108)

6 Berg, J., Dickhaut, J. & McCabe, K. 1995 Trust, reciprocity, and social history. *Games Econ. Behav.* **10**, 122–142.

7 McCabe, K. A. & Smith, V. L. 2000 A comparison of naïve and sophisticated subject behavior with game theoretic predictions. *Proc. Natl Acad. Sci. USA* **97**, 3777–3781. (doi:10.1073/pnas.040577397)

8 King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R. & Montague, P. R. 2005 Getting to know you: reputation and trust in a two-person economic exchange. *Science* **308**, 78–83. (doi:10.1126/science.1108062)

9 Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U. & Fehr, E. 2008 Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron* **58**, 639–650. (doi:10.1016/j.neuron.2008.04.009)

10 Schultz, W., Dayan, P. & Montague, P. R. 1997 A neural substrate of prediction and reward. *Science* **275**, 1593–1599. (doi:10.1126/science.275.5306.1593)

11 McClure, S. M., Berns, G. S. & Montague, P. R. 2003 Temporal prediction errors in a passive learning task activate human striatum. *Neuron* **38**, 339–346. (doi:10.1016/S0896-6273(03)00154-5)

12 O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H. & Dolan, R. J. 2003 Temporal difference models and reward-related learning in the human brain. *Neuron* **38**, 329–337. (doi:10.1016/S0896-6273(03)00169-7)

13 Greenwald, A. G., McGhee, D. E. & Schwartz, J. L. K. 1998 Measuring individual differences in implicit cognition: the implicit association test. *J. Personality Social Psychol.* **74**, 1464–1480. (doi:10.1037/0022-3514.74.6.1464)

14 Adolphs, R., Tranel, D. & Damasio, A. R. 1998 The human amygdala in social judgment. *Nature* **393**, 470–474. (doi:10.1038/30982)

15 Winston, J. S., Strange, B. A., O'Doherty, J. & Dolan, R. J. 2002 Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nat. Neurosci.* **5**, 277–283. (doi:10.1038/nn816)

16 Engell, A. D., Haxby, J. V. & Todorov, A. 2007 Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *J. Cogn. Neurosci.* **19**, 1508–1519. (doi:10.1162/jocn.2007.19.9.1508)

17 Todorov, A. & Engell, A. D. 2008 The role of the amygdala in implicit evaluation of emotionally neutral faces. *Social Cogn. Affect. Neurosci.* **3**, 303–312. (doi:10.1093/scan/nsn033)

18 Said, C. P., Baron, S. G. & Todorov, A. 2009 Nonlinear amygdala response to face trustworthiness: contributions of high and low spatial frequency information. *J. Cogn. Neurosci.* **21**, 519–528. (doi:10.1162/jocn.2009.21041)

19 Koscik, T. R. & Tranel, D. 2011 The human amygdala is necessary for developing and expressing normal interpersonal trust. *Neuropsychologia* **49**, 602–611. (doi:10.1016/j.neuropsychologia.2010.09.023)

20 Davis, M. & Whalen, P. J. 2001 The amygdala: vigilance and emotion. *Mol. Psychiatry* **6**, 13–34.

21 Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C. & Banaji, M. R. 2000 Performance on indirect measures of race evaluation predicts amygdala activation. *J. Cogn. Neurosci.* **12**, 729–738. (doi:10.1162/089892900562552)

22 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. *Psychol. Sci.* **15**, 806–813. (doi:10.1111/j.0956-7976.2004.00760.x)

23 Hart, A. J., Whalen, P. J., Shin, L. M., McInerney, S. C., Fischer, H. & Rauch, S. L. 2000 Differential response in the human amygdala to racial outgroup vs. ingroup face stimuli. *Neuroreport* **11**, 2351–2355. (doi:10.1097/00001756-200008030-00004)

24 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. *Nat. Neurosci.* **8**, 720–722. (doi:10.1038/nn1465)

25 Stanley, D. A., Phelps, E. A. & Banaji, M. R. 2008 The neural basis of implicit attitudes. *Curr. Direct. Psychol. Sci.* **17**, 164–170. (doi:10.1111/j.1467-8721.2008.00568.x)

26 Amorapanth, P., LeDoux, J. E. & Nader, K. 2000 Different lateral amygdala outputs mediate reactions and actions elicited by a fear-arousing stimulus. *Nat. Neurosci.* **3**, 74–79. (doi:10.1038/71145)

27 Hartley, C. A. & Phelps, E. A. 2010 Changing fear: the neurocircuitry of emotion regulation. *Neuropsychopharmacology* **35**, 136–146. (doi:10.1038/npp.2009.121)

28 Delgado, M. R. 2007 Reward-related responses in the human striatum. *Ann. NY Acad. Sci.* **1104**, 70–88. (doi:10.1196/annals.1390.002)

29 Brainard, D. H. 1997 The psychophysics toolbox. *Spatial Vision* **10**, 433–436. (doi:10.1163/156856897X00357)

30 Pelli, D. G. 1997 The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision* **10**, 437–442. (doi:10.1163/156856897X00366)

31 Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B. & Joliot, M. 2002 Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* **15**, 273–289. (doi:10.1006/nimg.2001.0978)

32 Chib, V. S., Rangel, A., Shimojo, S. & O'Doherty, J. P. 2009 Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *J. Neurosci.* **29**, 12 315–12 320. (doi:10.1523/JNEUROSCI.2575-09.2009)

33 Lin, A., Adolphs, R. & Rangel, A. 2011 Social and monetary reward learning engage overlapping neural substrates. *Soc. Cogn. Affect. Neurosci.* (doi:10.1093/scan/nsr006)

34 Rangel, A. & Hare, T. A. 2010 Neural computations associated with goal-directed choice. *Curr. Opin. Neurobiol.* **20**, 262–270. (doi:10.1016/j.conb.2010.03.001)

35 Behrens, T. E. J., Hunt, L. T. & Rushworth, M. F. S. 2009 The computation of social behavior. *Science* **324**, 1160–1164. (doi:10.1126/science.1169694)

36 Rilling, J. K. & Sanfey, A. G. 2011 The neuroscience of social decision-making. *Ann. Rev. Psychol.* **62**, 23–48. (doi:10.1146/annurev.psych.121208.131647)

37 Tomlin, D., Kayali, M. A., King-Casas, B., Anen, C., Camerer, C. F., Quartz, S. R. & Montague, P. R. 2006 Agent-specific responses in the cingulate cortex during economic exchanges. *Science* **312**, 1047–1050. (doi:10.1126/science.1125596)

38 Behrens, T. E. J., Hunt, L. T., Woolrich, M. W. & Rushworth, M. F. S. 2008 Associative learning of social value. *Nature* **456**, 245–249. (doi:10.1038/nature07538)

39 LeDoux, J. E. & Gorman, J. M. 2001 A call to action: overcoming anxiety through active coping. *Am. J. Psychiatry* **158**, 1953–1955. (doi:10.1176/appi.ajp.158.12.1953)