Cogn Affect Behav Neurosci
DOI 10.3758/s13415-015-0393-0

Trying to trust: Brain activity during interpersonal social attitude change

Megan M. Filkowski[1] · **Ian W. Anderson**[1] · **Brian W. Haas**[1,2]

Abstract Interpersonal trust and distrust are important components of human social interaction. Although several studies have shown that brain function is associated with either trusting or distrusting others, very little is known regarding brain function during the control of social attitudes, including trust and distrust. This study was designed to investigate the neural mechanisms involved when people attempt to control their attitudes of trust or distrust toward another person. We used a novel control-of-attitudes fMRI task, which involved explicit instructions to control attitudes of interpersonal trust and distrust. Control of trust or distrust was operationally defined as changes in trustworthiness evaluations of neutral faces before and after the control-of-attitudes fMRI task. Overall, participants (n = 60) evaluated faces paired with the distrust instruction as being less trustworthy than faces paired with the trust instruction following the control-of-distrust task. Within the brain, both the control-of-trust and control-of-distrust conditions were associated with increased temporoparietal junction, precuneus (PrC), inferior frontal gyrus (IFG), and medial prefrontal cortex activity. Individual differences in the control of trust were associated with PrC activity, and individual differences in the control of distrust were associated with IFG activity. Together, these findings identify a brain network involved in the explicit control of distrust and trust and indicate that the PrC and IFG may serve to consolidate interpersonal social attitudes.

Keywords Interpersonal trust · Cognitive control · Social attitudes · Emotion · Functional connectivity · Neural network

Interpersonal trust is an important component of human social interaction. Trusting other people is associated with the strength of social relationships and the way many interpersonal and economic decisions are made (Delgado, Frank, & Phelps, 2005; Riedl & Javor, 2012). Within the human population, considerable variability exists in the ways and the extent to which people trust each other (Borum, 2010; Riedl & Javor, 2012). Interpersonal trust has been operationalized as a relatively stable trait that varies within the human population (Fleeson & Leicht, 2006). However, attitudes of interpersonal trust can dynamically change over time and can be affected by a variety of factors, including competence or attractiveness (Chang, Doll, van't Wout, Frank, & Sanfey, 2010). Although several studies have investigated the brain basis of trust-related decision-making, very little is currently known regarding how people can exhibit control over their interpersonal attitudes of trust and distrust. For example, if one were being introduced to a person for the first time, and had been told, immediately prior, "this person is trustworthy, you can really count on them," how would this statement influence the subsequent social interaction? Do individual differences exist in terms of how willing people are to accept that new people (i.e., strangers) are trustworthy or untrustworthy, in response to being instructed to do so? This study was designed to

Electronic supplementary material The online version of this article (doi:10.3758/s13415-015-0393-0) contains supplementary material, which is available to authorized users.

✉ Brian W. Haas
bhaas@uga.edu

[1] Behavioral and Brain Sciences Program, Department of Psychology, University of Georgia, Athens, GA, USA

[2] Interdisciplinary Neuroscience Graduate Program, University of Georgia, Athens, GA, USA

Published online: 13 November 2015

investigate the way people exhibit cognitive control over social attitudes of interpersonal trust and distrust.

Interpersonal trust is operationally defined as the willingness to put oneself in a vulnerable position dependent on another person's actions (Borum, 2010; Lewicki, Tomlinson, & Gillespie, 2006). Many researchers have investigated the tendency to trust others by using tasks that involve the trustworthiness evaluation of faces or through economic games that infer trust via cooperation (Adolphs, 2002; Dimoka, 2010, 2011; Krueger et al., 2007; Riedl, Hubert, & Kenning, 2010; Riedl, Mohr, Kenning, Davis, & Heekeren, 2014; Rilling et al., 2002). These studies have highlighted the role of limbic regions, such as the amygdala, when faces are judged to be either very trustworthy or very untrustworthy (Rule, Krendl, Ivcevic, & Ambady, 2013; Winston, Strange, O'Doherty, & Dolan, 2002). When using paradigms such as the prisoner's dilemma or the trust game, studies have shown a mixed pattern of results. Economic trust games indicate that the initial decision to trust is associated with increased activation within the paracingulate cortex and septal area, whereas decisions to defect are associated with activation within the ventral tegmental area (Krueger et al., 2007). Other studies of trust have indicated that the amygdala, insula, and orbitofrontal cortex are active during the evaluation of trust and distrust (Adolphs, 2002; Dimoka, 2010, 2011). Other research has shown that trust (or reciprocity) is associated with greater activity within dorsomedial prefrontal cortex (dmPFC), precuneus (PrC), and temporoparietal junction (TPJ; Emonds, Declerck, Boone, Seurinck, & Achten, 2014; Emonds, Declerck, Boone, Vandervliet, & Parizel, 2011; Morishima, Schunk, Bruhin, Ruff, & Fehr, 2012; Rilling & Sanfey, 2011; Watanabe et al., 2014). Distrust (or unreciprocated trust) is associated with greater activity within the anterior insula (Rilling & Sanfey, 2011). Interestingly, several of these areas, such as the medial prefrontal cortex (mPFC) and insula, are also implicated in emotion regulation research (Goldin, McRae, Ramel, & Gross, 2008; Grecucci, Giorgetta, van't Wout, Bonini, & Sanfey, 2013; Ochsner & Gross, 2005; Ochsner et al., 2004), which is also relevant in the conscious control/regulation of these interpersonal attitudes.

Trust and distrust are both interpersonal social attitudes; however, trust and distrust also may represent distinct psychological constructs (Cho, 2006; Lewicki, McAllister, & Bies, 1998). In terms of emotional valence, trust is a positively valenced interpersonal attitude, whereas distrust represents a negatively valenced interpersonal attitude. There is evidence that social and emotional information is processed differently according to valence. For example, negative social information tends to be prioritized and is more salient when forming interpersonal impressions than is positive social information (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Peeters & Czapinski, 1990). There is also evidence that the brain detects and processes negative emotional stimuli (such as fearful faces or negative emotional words) differently than positive emotional stimuli (Baumeister et al., 2001; Rozin & Royzman, 2001; Smith, Cacioppo, Larsen, & Chartrand, 2003). Combined, these studies suggest that trust and distrust may rely on dissociable psychological and neural systems (Dimoka, 2010), and that distrust may be a more salient social cue than trust. Therefore, in this study, we predicted that trust and distrust would be processed differently and that distrust would be a more salient cue than trust.

In this study, we used a novel experiment to measure the effortful control of interpersonal trust and distrust, behaviorally and within the brain. The design is based on the extant empirical neuroimaging and behavioral research on automatic trustworthiness evaluations of faces (Bzdok et al., 2011; Bzdok et al., 2012; Todorov, 2008; Todorov, Baron, & Oosterhof, 2008; Todorov, Mende-Siedlecki, & Dotsch, 2013; Winston et al., 2002), interpersonal attitude formation (Cunningham, Raye, & Johnson, 2004, Kuzmanovic et al., 2012), and emotion regulation (Goldin et al., 2008; Ochsner & Gross, 2005; Ochsner et al., 2004; Ochsner, Silvers, & Buhle, 2012). Participants evaluated the trustworthiness of faces before and after a control-of-trust task. Each trustworthiness evaluation task was completed outside an fMRI scanner, and the control-of-trust task was completed while fMRI data were being collected. This approach allows "control of trust and distrust" to be operationalized as the change in trustworthiness evaluations (from pre to post scanning). Additionally, the collection of fMRI data provides insight as to what brain regions are involved when people attempt to control interpersonal attitudes of trust and distrust. Finally, combined trustworthiness evaluation and fMRI data provide the opportunity to investigate how individual differences in the control of trust or distrust may be associated with individual differences in brain activity.

On the basis of existing neuroimaging studies on trust, interpersonal attitude formation, and emotion regulation (Bzdok et al., 2011; Bzdok et al., 2012; Cunningham et al., 2004; Kuzmanovic et al., 2012; Goldin et al., 2008; Ochsner & Gross, 2005; Ochsner et al., 2004; Ochsner et al., 2012; Todorov et al., 2008; Winston et al., 2002), we predicted that the control of trust and distrust would be subserved by activity within the mPFC, TPJ, PrC, insula, and dorsolateral and inferior frontal cortices. Furthermore, on the basis of prior evidence that negative social information is more salient than positive social information (Baumeister et al., 2001; Rozin & Royzman, 2001; Smith et al., 2003), we predicted that the control-of-distrust instruction would have a greater impact on changes in trustworthiness evaluations than would the control-of-trust instruction.

Method

Participants

Sixty healthy, right-handed, English-speaking participants (37 female, 23 male; mean age 20.25 ± 2.46 years) were recruited

from the University of Georgia and the surrounding community. All participants were screened for neurological and psychiatric illnesses as well as for MRI contraindications. Participants provided written informed consent prior to participation, and the University of Georgia Institutional Review Board approved all study procedures. One participant was dropped from the analyses that involved changes in trust evaluations due to a computer error during collection of the trustworthiness evaluation data (n = 59; 22 male/37 female, mean age = 20.42 ± 2.47). Therefore, all analyses that include "change-of-trust" scores as a variable of interest are reported on the basis of n = 59, whereas all other analyses are reported on the basis of n = 60. There was no significant difference between the numbers of males and females [$\chi^2(1) = 3.32$, $p > .05$]. Finally, age did not differ between the male and female participants [$t(58) = 1.879$, $p > .05$].

Procedure

The study consisted of two separate sessions. During the first session, participants completed two behavioral tasks, an initial evaluation of trustworthiness and a face name memory task. The initial evaluation of trustworthiness task provided a baseline measure of how trustworthy participants believed each face to be. The face name memory task was used to quantify each participant's ability to remember the names of faces (subsequently used as a covariate, nuisance variable in the regression analysis). During the second session, participants underwent fMRI while completing the control-of-trust task. Postscan, participants completed the trustworthiness evaluation task again. Postscan ratings of the same faces provided the data necessary to calculate change-in-trust scores. The average duration between the initial evaluation-of-trustworthiness task and the fMRI data collection was 40 ± 30.96 days (range = 8–132).

Initial evaluation of trustworthiness

The design of the task was based on prior behavioral research on the trustworthiness evaluation of faces (Rule et al., 2013; Todorov, 2008). Participants were presented with a series of 36 neutral faces and were instructed to evaluate how trustworthy they believed each face to be, on a 7-point Likert scale, with 1 being *untrustworthy* and 7 being *trustworthy*. Faces were presented until the participant gave a response. Therefore, the task duration varied among participants. The color photographs of the faces were selected from a standardized database (Minear & Park, 2004) and presented on a white background via a computer monitor using E-Prime software (www.pstnet.com/eprime.cfm). The selected faces comprised 18 males and 18 females (mean age = 26.72 ± 6.9 years). Mean trustworthiness evaluations were calculated for each participant.

Face name memory task

Participants were presented with a name followed by a face and were explicitly instructed to remember the name paired with each face. Twenty neutral Ekman faces (Ekman & Friesen, 1971) were used and repeated twice in random order. The names for the faces were randomly generated using a standardized Web-based name generator (www.namegenerator.biz). Approximately 10 min following encoding, each participant was presented with 10 out of the 20 faces and three possible names and was instructed to choose the correct name that had previously been paired with the face. Of the three possible names, one was correct, one was incorrect but had previously been presented during encoding with a different face, and one name was incorrect and had not previously been presented during encoding. For each participant, accuracy was calculated by dividing the total number of correct responses by the total number of possible responses. The faces presented in this paradigm were selected from a different database from those used during the control-of-trust task in order to reduce habituation and carryover effects between the two tasks.

The face name memory task, a recognition memory paradigm, was included in order to quantify each participant's ability to remember words paired with faces. More specifically, an important aspect of this study was to characterize individual differences in the ability to control the interpersonal attitudes of others, on the basis of semantic instructions (trust or distrust) paired with the images of faces. Face name memory scores were included as a covariate in the individual differences analyses to improve the validity of the change-of-trust or -distrust scores. See the Discussion section for limitations of the task and this approach.

Control-of-trust task

During a separate session, each participant underwent fMRI while completing the control-of-trust/distrust task. Prior to entering the scanner, each participant was read the following instructions:

> In the following experiment, we are interested in your ability to control how much you either Trust or Distrust someone. When you see the instruction "Trust" (or the letter "T") your job will be to do your best to imagine trusting that this person has your best interests in mind. When you see the instruction "Distrust" (or the letter "D"), your job will be to do your best to imagine that this person is motivated to take advantage of you. This type of trust or distrust could be related to friendship, financial issues or academic or professional advice. Lastly, when you see the instruction "Age" (or the letter

"A"), your job will be to do your best to determine the age of the person depicted in the photograph.

Each participant then completed a brief practice version of the task before entering the scanner, to ensure that participants understood the directions and the procedure of the task. Once in the scanner, and immediately prior to fMRI data collection, the participants were once again presented the written instructions. During the control-of-trust task, each participant was presented with a cue indicating the condition (control of trust, control of distrust, or age evaluation), followed by a photograph of a face (Supplementary Fig. 1). Participants were instructed to imagine trusting, distrusting, or evaluate the age of each face. No behavioral responses were collected during the fMRI task.

The faces included in the control-of-trust task were identical to the faces included in the evaluation-of-trustworthiness task. However, the 36 face stimuli were divided into three categories: 12 of the faces were paired with a "control-of-trust" instruction, 12 were paired with a "control-of-distrust" instruction, and 12 were paired with an "evaluation-of-age" instruction. The faces paired with the "control-of-trust" and "control-of-distrust" conditions were counterbalanced across participants, and participants were randomly assigned to one of the two versions. Changes in the trustworthiness evaluations were not different according the version of the task [$t(58) = -0.841$, $p = .404$]. We selected age evaluation as the control condition, on the basis of prior trustworthiness face-processing research (Winston et al., 2002). The age evaluation condition involves the same perceptional characteristics and is also "evaluative," however age is not specifically an interpersonal attitude, while trust and distrust are specifically interpersonal attitudes. For each condition (trust, distrust, and age), equal proportions of male and female faces were presented, and there were no significant differences between conditions in the ages of the people presented in the photographs [$F(2, 35) = 0.05$, $p = .95$].

After the initial set of instructions, participants were presented with a fixation cross for 15 s in order to stabilize and calibrate collection of the MRI data. These scans (five TRs) were not modeled and were not included in any of the statistical analyses. This was followed by the presentation of seven blocks for each condition, resulting in a total of 21 blocks. These blocks varied in the numbers of faces presented (three or four faces per block). For each condition, there were four blocks with three faces per block and three blocks with four faces per block. The blocks varied in number of trials in order to reduce the tendency of participants to anticipate the timing parameters within the task and to increase each participant's attentiveness to each trial. Each face image was presented for a total of 5 s (with a 1-s cue prior to presentation). All stimuli were presented twice throughout the experiment. Thus, each participant was presented with a total of 72 stimuli (36 × 2). The number of stimuli was selected so as to preserve the balance between reliably measuring the psychological construct of interest (control of interpersonal social attitudes) and attention throughout the task. The duration of the task was 7 min 30 s. An additional fixation of 15 s was presented at the end of the task.

Postscan evaluation of trustworthiness

Immediately following the fMRI scanning (about 10 min after completing the control-of-trust task), each participant completed a postscan evaluation-of-trustworthiness task. The postscan evaluation task was identical to the initial evaluation-of-trustworthiness task (i.e., the same 36 faces). Participants were seated in front of a computer screen and asked to evaluate how trustworthy they believed the face presented appeared to be, using a 7-point Likert scale, with 1 being *untrustworthy* and 7 being *trustworthy*.

Control of trust: Behavioral measure

To operationalize each participant's ability to control attitudes of trust and distrust, the initial evaluation-of-trustworthiness scores were subtracted from the postscan scores, for the faces paired with the "control-of-trust" or "control-of-distrust" instructions during fMRI scanning. Thus, a positive change in the trustworthiness evaluations represent sucessfully modifying attitudes of trust toward faces following the control-of-trust condition, and negative values represent sucessfully modifying attitudes of distrust toward faces following the control-of-distrust condition.

fMRI data acquisition

Whole-brain imaging data were acquired on a GE-Signa 3-T scanner (General Electric, Milwaukee, WI) at the University of Georgia Bio-Imaging Research Center (birc.uga.edu). A total of 195 functional images were acquired using a gradient echo T2*-weighted echoplanar imaging scan and were obtained using a flip angle of 90°, repetition time = 2.0 s, echo time = 25 ms, 40 slices, and field of view = 220 × 64 mm matrix. For the structural whole-brain images, a three-dimensional high-resolution spoiled gradient scan was conducted (repetition time, 24 ms; echo time, 4.5 ms; flip angle, 20°; matrix size, 256 × 256; field of view, 25.6 cm; slice thickness, 1.0 mm; 164 contiguous slices).

fMRI data processing

The fMRI data were preprocessed and statistically analyzed using SPM8 (Wellcome Department of Imaging Neuroscience, London, UK) and implemented through MATLAB

R2012a (www.mathworks.com). The images were temporally realigned to the middle slice, spatially realigned to the first in the time series. The images were then co-registered and spatially normalized into a standard stereotactic space (MNI template) and spatially smoothed with an 8-mm full-width-at-half-maximum isotropic Gaussian filter. Three dummy scans were discarded prior to the analysis (http://fil.ion.ucl.ac.uk/spm/doc/manual.pdf).

fMRI data statistical analysis

A series of planned contrasts between conditions were performed to identify task-specific changes in brain activity. Each contrast was performed on whole-brain data, using a family-wise error (FWE) corrected threshold of $p < .05$, ten-voxel extent. In order to identify regions involved during the explicit control of social interpersonal attitudes, BOLD responses during the control of attitude conditions were compared to the age evaluation condition (trust + distrust > age, trust > age, and distrust > age). Next, BOLD responses during the control-of-trust condition were compared to BOLD responses during the control-of-distrust condition (i.e., trust > distrust and distrust > trust). For each main contrast, we also report the results of a region-of-interest (ROI) based analysis within the mPFC, TPJ, PrC, insula, and dorsolateral and inferior frontal cortices, using an FWE-corrected statistical threshold.

The mPFC, insula, and dorsolateral and inferior frontal cortical ROIs were selected from a standardized atlas used for functional neuroimaging (Maldjian, Laurienti, Kraft, & Burdette, 2003). The TPJ ROI was specified as a sphere (16-mm radius) surrounding the coordinates (MNI: 54, –55, 26 and –54, –55, 26) reported in Mars et al. (2012), and the PrC ROI was specified as a sphere (20-mm radius) surrounding the coordinates (MNI: –12, –56, 32 and 12, –56, 32) reported by Kuzmanovic et al. (2012).

Next, exploratory analyses were performed designed to investigate individual differences in the ability to control attitudes of trust and distrust (e.g., changes in the trustworthiness evaluations). We also investigated the association between individual differences in "change of trust" and "change of distrust" scores and functional connectivity within the brain. Because there was considerable variability and a large range in the numbers of days between pre and post, all analyses that involved "change scores" were also performed with "number of days between" as a covariate. Additionally, each individual-difference analysis was performed with face name memory scores entered as a covariate. Change-in-trustworthiness scores were not correlated with the number of days between sessions [$r(57) = –.089$, $p = .497$].

For the individual-differences analyses, each participant's change-in-trustworthiness scores were entered as the independent variable, and contrast images (trust > age evaluation and distrust > age evaluation) were entered as the dependent variable. For functional connectivity analyses, seed regions were specified on the basis of whole-brain regression analysis, with change-of-trust (or change-of-distrust) scores predicting BOLD signal change between the control of trust versus age evaluation condition, and the control of distrust versus age evaluation condition, respectively. On the basis of prior research on trust and interpersonal attitude formation (Bzdok et al., 2011; Bzdok et al., 2012; Cunningham et al., 2004; Kuzmanovic et al., 2012; Todorov et al., 2008; Winston et al., 2002), we performed ROI analyses within the mPFC, TPJ, PrC, insula, and dorsolateral and inferior frontal cortices. For the exploratory statistical analysis within a-priori-specified ROIs, a statistical threshold of $p < .001$, 20-voxel extent was used. This combination is sufficient to preserve the balance between sensitivity and false-positive rates (Lieberman & Cunningham, 2009; Woo et al., 2014).

Results

Behavioral results

This section is divided into two subsections. First, we report on the initial evaluations of faces and the face name task, followed by the effect of the control-of-trust task on changes in the trustworthiness evaluations.

The mean rating for all faces was 4.36 ($SD = 0.92$, range = 2.28–6.47, variance = 0.850). We observed no associations between mean trustworthiness evaluations and the sex [$t(58) = 0.250$, $p = .803$] or age [$r(58) = .059$, $p = .655$] of the participants. There were no significant differences between the mean ratings of faces that were subsequently paired with a "trust" or "distrust" instruction during the imaging task [$t(58) = 0.897$, $p = .374$]. Finally, female faces were rated as more trustworthy than male faces (male mean = 4.06 ± 0.949, female mean = 4.67 ± 0.972) [$t(58) = 8.744$, $p < .01$].

The mean face name memory score across participants was 8.31 ± 1.51. All participants scored at or above chance, with a minimum score of 5 and a maximum score of 10. We found no significant differences between the sexes [$t(58) = .250$, $p = .803$], and scores were also not associated with the ages of participants [$r(58) = –.110$, $p = .406$], trust evaluation scores [$r(58) = .062$, $p = .642$], change-in-trust scores [$r(58) = .145$, $p = .274$], or change-in-distrust scores [$r(58) = .003$, $p = .980$].

Following the control-of-trust task, the faces paired with the distrust instruction were evaluated as being less trustworthy than at the initial evaluation (pre: $M = 4.34$, $SD = 0.96$; post: $M = 3.62$, $SD = 0.71$) [paired t test: $t(1, 57) = 4.74$, $p < .001$] (Fig. 1). The change scores for "distrust faces" (from pre to post) remained statistically significant when face name memory scores and the number of days between sessions were

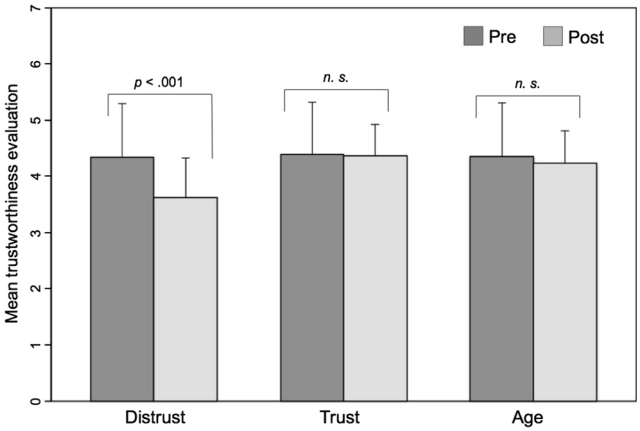

Fig. 1 Bar graph showing the mean ratings of trustworthiness for each condition (trust, distrust, age) between the pre and post scans. Error bars represent standard deviations. n.s., not significant

entered as covariates [$F(1, 56) = 13.14$, $p = .004$]. No significant change in trustworthiness was observed for faces paired with the trust instruction (pre: $M = 4.39$, $SD = 0.93$; post: $M = 4.37$, $SD = 0.56$) [paired t test: $t(1, 57) = 0.21$, $p = .83$]. Finally, for the faces paired with the "evaluation-of-age" instruction, no significant change was observed in the trustworthiness evaluations (pre: $M = 4.36$, $SD = 0.95$; post: $M = 4.24$, $SD = 0.58$) [paired t test: $t(1, 57) = 0.96$, $p = .34$]. Combined, these findings indicate that the distrust instruction served to reduce trustworthiness evaluations across participants, whereas the trust and age instructions did not consistently affect trustworthiness evaluations across participants. This finding may suggest that some participants became more trusting of the faces following the control-of-trust task, whereas other participants did not.

Neuroimaging results

Control of interpersonal social attitudes BOLD responses when participants were asked to control interpersonal social attitudes (both trust and distrust combined) were compared to results from the age evaluation condition. Whole-brain results (unmasked) are listed in Table 1 and are presented in Fig. 2A. Within ROIs, we found that the control of attitudes was associated with greater activity within bilateral TPJ [$t(57) = 10.17$, $p < .001$, 2,225 voxels, and $t(57) = 6.54$, $p < .001$, 442 voxels], left anterior cingulate cortex (ACC)/mPFC ($t = 8.30$, $p < .001$, 3,664 voxels), bilateral IFG [$t(57) = 8.08$, $p < .001$, 1,215 voxels, and $t(57) = 7.36$, $p < .001$, 78 voxels], and left PrC [$t(57) = 5.36$, $p < .001$, 43 voxels]. All results remained statistically significant when face name memory scores and the number of days between the initial evaluation and fMRI scanning were entered as covariates ($p < .001$).

Table 1 Regions of increased activation in the control of social attitudes

		MNI				
Anatomical Region	L/R	x	y	z	t	k
Attitude > Age						
ACC/mPFC	L	−10	40	48	8.30	3,664
TPJ	L	−46	−64	26	10.17	2,225
	R	56	−68	12	6.54	442
IFG	L	−48	26	2	8.08	1,215
	R	54	30	−2	7.36	78
Cerebellum–(Crus-I)	R	26	−82	−34	7.27	216
MCC	R	−2	−14	40	6.45	117
PrC	L	−6	−52	34	5.36	43
Superior temporal sulcus	R	52	−6	−16	5.52	34
Middle temporal gyrus/pole	R	42	12	−36	5.85	13
Thalamus	L	−12	−8	16	4.70	12
Trust > Age						
ACC/mPFC	L	−2	36	2	8.44	3,236
TPJ	L	−50	−72	38	8.78	1,145
	R	60	−58	24	6.01	91
IFG	L	−46	−28	−4	5.71	277
	R	54	30	−2	5.90	33
Middle temporal gyrus	L	−52	−4	−30	5.83	162
Cerebellum–(Crus-I)	R	26	−82	−34	6.57	91
PrC	L	−6	−52	32	5.16	40
MCC	L	−2	−14	40	5.58	31
Distrust > Age						
TPJ	L	−44	−62	26	10.04	2,859
	R	66	−44	26	6.54	514
ACC/mPFC	L	−8	52	42	8.60	2,297
Middle temporal gyrus	L	−48	−2	−36	8.49	1,427
Cerebellum–(Crus-I)	R	24	−76	−36	7.07	242
IFG extending to insula	R	54	30	−2	7.08	66
	L	−42	−10	54	5.31	26
PCC	L	−2	−14	40	5.75	65
Precentral gyrus	R	56	−6	52	5.86	50
Superior temporal gyrus	R	56	−4	−16	5.29	26
Temporal pole	R	52	10	−20	5.09	19

Results of the whole-brain analysis for each attitude contrast (attitude > age, trust > age, distrust > age), FWE-corrected $p = .05$, 10-voxel extent. ACC, anterior cingulate cortex; mPFC, medial prefrontal cortex; TPJ, temporoparietal junction; IFG, inferior frontal gyrus; MCC, medial cingulate cortex; PrC, precuneus; PCC, posterior cingulate cortex; L, left hemisphere; R, right hemisphere

Control of trust BOLD signals were compared during the control-of-trust condition and the age evaluation condition. Whole-brain (unmasked) results are listed in Table 1 and shown in Fig. 2B. Within ROIs, the control-of-trust condition was associated with greater activity within bilateral TPJ [$t(59) = 8.78$, $p < .001$, 1,145 voxels, and $t = 6.01$, $p < .001$, 74

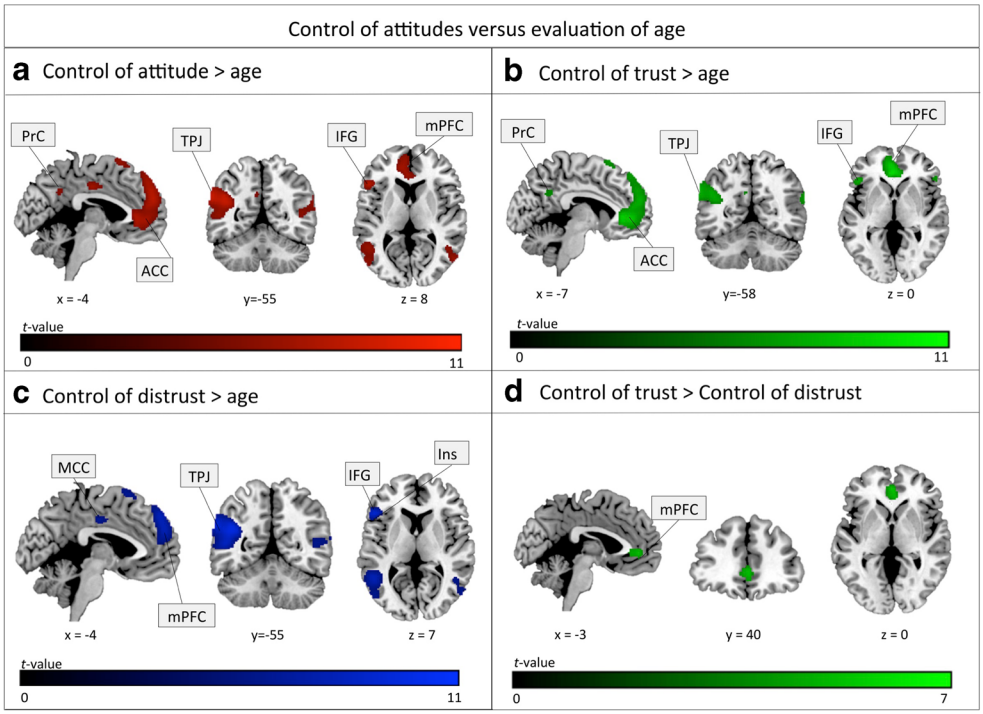

Fig. 2 (A) Whole-brain analysis showing changes in brain activity during the control of social attitudes (control of trust and distrust combined) versus the evaluation-of-age condition. (B) Whole-brain analysis showing changes in brain activity during the control of trust versus the evaluation-of-age condition. (C) Whole-brain analysis showing changes in brain activity during the control of distrust versus the evaluation-of-age condition. (D) Direct comparison between the control-of-trust and control-of-distrust conditions. Areas of significant changes of BOLD signal are overlaid on a standardized template of the brain. PrC, precuneus; ACC, anterior cingulate cortex; TPJ, temporoparietal junction; IFG, inferior frontal gyrus; mPFC, medial prefrontal cortex; MCC, medial cingulate cortex; Ins, insula

voxels], bilateral IFG [$t(59) = 5.71$, $p < .001$, 277 voxels, and $t(59) = 5.90$, $p < .001$, 33 voxels], left ACC/ventromedial PFC (vmPFC) [$t(59) = 8.44$, $p < .001$, 3,236 voxels], and left PrC [$t(59) = 5.16$, $p < .001$, 40 voxels]. All results remained statistically significant when face name memory scores and the number of days between sessions were entered as covariates ($p < .001$).

Control of distrust Whole-brain (unmasked) results for the control-of-distrust condition versus the age evaluation condition are listed in Table 1 and shown in Fig. 2C. Within ROIs, the control-of-distrust condition was associated with greater activity within bilateral TPJ [$t(59) = 10.04$, $p < .001$, 2,859 voxels, and $t(59) = 6.54$, $p < .001$, 514 voxels], left ACC/mPFC [$t(59) = 8.60$, $p < .001$, 2,297 voxels], and bilateral IFG extending to insula [$t(59) = 7.08$, $p < .001$, 66 voxels, and $t(59) = 5.31$, $p < .001$, 26 voxels]. All results remained statistically significant when face name memory scores and the number of days between sessions were entered as covariates ($p < .001$).

Control of trust versus control of distrust Whole-brain (unmasked) results are listed in Table 2 and shown in Fig. 2D. When the conditions were directly compared, the control-of-trust condition was associated with greater vmPFC/ACC [$t(59) = 6.51$, $p < .001$, 269 voxels] activity than was the control-of-distrust condition. The control-of-distrust condition was not associated with increased activation. This result remained statistically significant when face name memory scores and the number of days between sessions were entered as covariates ($p < .001$).

Individual differences in control of trust Whole-brain (unmasked) analysis revealed that greater change in trustworthiness evaluations was associated with greater PrC activity [$t(57) = 3.67$, $p < .001$, 41 voxels] (Table 3 and Fig. 3A).

Table 2 Differences in the control of trust versus distrust

		MNI				
Anatomical Region	L/R	*x*	*y*	*z*	*t*	*k*
Trust > Distrust						
ACC/mPFC/vmPFC	R	0	44	0	6.51	269
Distrust > Trust						
No significant clusters						

Results of the whole-brain analysis for the control of trust versus distrust (trust > distrust, distrust > trust), FWE-corrected $p = .05$, 10-voxel extent. ACC, anterior cingulate cortex; mPFC, medial prefrontal cortex; vmPFC, ventromedial prefrontal cortex; R, right hemisphere

Individuals who tended to evaluate faces as more trustworthy following the control-of-trust task exhibited greater PrC activity during the control-of-trust task than did individuals who did not tend to evaluate faces as more trustworthy following that task [$r(57) = .430$, $p = .001$]. The association between change in trust and PrC activity remained statically significant when face name memory scores and days between sessions were entered as covariates [$t(57) = 3.61$, $p < .001$, 26 voxels].

Individual differences in control of distrust Whole-brain (unmasked) analysis revealed that greater change in trustworthiness evaluations was associated with greater activation within the left IFG [$t(57) = 4.58$, $p < .001$, 221 voxels] and left inferior temporal gyrus [$t(57) = 3.74$, $p < .001$, 21 voxels] (Table 3 and Fig. 3B). Individuals who tended to evaluate faces as less trustworthy following the control-of-distrust task exhibited greater IFG activity during that task than did individuals who did not tend to evaluate faces as less trustworthy following the task ($r = –.518$, $p < .001$). These results remained significant when we included face name memory and the number of days between sessions as covariates [$t(55) = 4.45$, $p < .001$, 221 voxels].

Exploratory functional connectivity analysis

A psychophysiological interaction (PPI) analysis (Friston et al., 1997) was performed to identify the functionally connected neural networks associated with the control of trust or distrust. This analysis focused on brain regions where BOLD signal was found to be associated with individual differences in control-of-trust and control-of-distrust scores. Therefore, the PrC was specified as the seed region for the control-of-trust condition, and the left IFG was specified as the seed region for the control-of-distrust condition. For the PrC, we extracted signals by performing a conjunction analysis (based on the group analysis) using the trust > age evaluation contrast (across all participants), and a regression analysis with change-in-trust scores predicting control-of-trust activity (MNI: –10, –44, 32) [$t(57) = 3.28$, $p < .001$, 54 voxels]. For the distrust PPI analysis, BOLD signal changes from the left IFG were extracted by performing a conjunction analysis (based on the group analysis) using the distrust > age evaluation contrast (across all participants) (MNI: –50, 24, 16), [$t(57) = 4.55$, $p < .001$, 153 voxels]. For each conjunction group analysis, the peak cluster was selected using a $p = .01$ statistical threshold for each analysis, resulting in a per-voxel statistical threshold of $p = .0001$. For each participant, the signal was extracted from the entire PrC cluster (MNI: –10, –44, 32; 54 voxels) and the entire IFG cluster (MNI: –50, 24, 16; 153 voxels). Next, we performed a regression analysis with change-of-trust or change-of-distrust scores as the independent variable, and either PrC or IFG connectivity entered as the dependent variable. This approach served to elucidate the patterns of brain connectivity associated with individual differences in change-of-trust or change-of-distrust scores, respectively.

The results of a whole-brain (unmasked) analysis are listed in Table 4 and presented in Supplementary Fig. 2. Individual differences in the change-of-trust scores were associated with increased PrC connectivity within a large distributed network, including the left mPFC, bilateral dorsolateral PFC (dlPFC), left inferior parietal lobe (IPL), left ACC, and right superior frontal gyrus. This finding indicates that the PrC is more connected with these brain regions in individuals who tended to evaluate faces as more trustworthy than in individuals who did not tend to evaluate faces as more trustworthy. No brain regions were less connected with the PrC.

For the control-of-distrust condition, individual differences in the change-of-distrust scores were associated with increased IFG negative connectivity with two brain regions: the right dmPFC and right temporopolar area. It is important to note that this pattern of "negative connectivity" refers to the IFG exhibiting a negative association with activity in other brain regions during the control-of-distrust condition, which may represent regulatory function. No brain regions were positively connected with the IFG.

Table 3 Individual differences in the control of interpersonal social attitudes

		MNI				
Anatomical Region	L/R	*x*	*y*	*z*	*t*	*k*
Trust						
PrC/PCC	L	−12	−42	32	3.67	41
Distrust						
IFG	L	−50	26	16	4.58	221
Inferior temporal gyrus	L	−38	6	−36	3.74	21

Results of the whole-brain analysis for individual differences in the control of social attitudes. Significance threshold: $p = .001$, 20-voxel extent. PrC, precuneus; PCC, posterior cingulate cortex; IFG, inferior frontal gyrus; L, left hemisphere

Discussion

This study was designed to investigate the ways that people exhibit control over the interpersonal social attitudes of trust and distrust. Both the control-of-trust and control-of-distrust conditions were associated with greater TPJ, mPFC, insula, and inferior and lateral frontal cortical activity than was the evaluation-of-age condition. The control-of-trust condition was associated with greater vmPFC activity compared to the control-of-distrust condition. Additionally, we found that individual differences in change-of-trust scores were

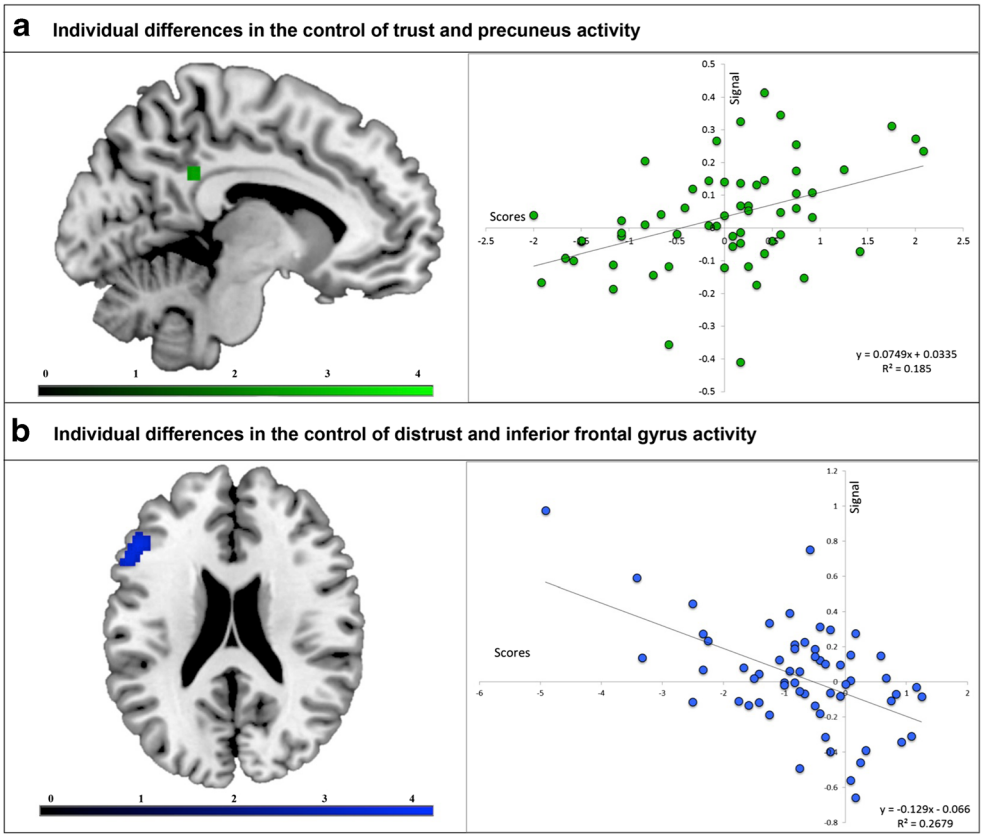

Fig. 3 (A) Individual differences in change-of-trust scores associated with precuneus (PrC) activity during the control-of-trust condition as compared to the age evaluation condition. The image depicts the cluster within the PrC. Change-in-trustworthiness values are plotted on the *x*-axis in the right panel, and PrC contrast estimates are plotted on the *y*-axis. (B) Individual differences in change-of-distrust scores associated with inferior frontal gyrus (IFG) activity during the control-of-distrust condition as compared to the age evaluation condition. The image depicts the cluster within the left IFG. Change-in-distrust scores are plotted on the *x*-axis in the right panel, and left IFG contrast estimates are plotted on the *y*-axis

associated with greater PrC activity, whereas individual differences in change-of-distrust scores were associated with greater IFG activity. Combined, these findings provide new information about the brain mechanisms engaged when people make efforts to change their social attitudes.

Table 4 Functional connectivity analysis

		MNI				
Anatomical Region	L/R	*x*	*y*	*z*	*t*	*k*
Precuneus Connectivity–Positive						
ACC/mPFC	L	−2	32	32	3.60	147
dlPFC	L	−40	28	24	3.76	143
	R	40	32	36	3.72	137
IPL	L	−56	−46	46	3.61	56
Superior frontal gyrus–medial segment	R	8	60	26	3.69	56
Superior frontal gyrus	R	8	50	52	3.97	52
	R	10	26	68	4.09	45
IFG Connectivity–Negative						
dmPFC	R	32	0	22	4.21	118
Temporopolar area	R	36	14	−26	3.58	30

Results of the whole-brain analysis for the functional connectivity of individual differences in control of trust and distrust. Significance threshold: $p = .001$, 20-voxel extent. ACC, anterior cingulate cortex; mPFC, medial prefrontal cortex; dlPFC, dorsolateral prefrontal cortex; IPL, inferior parietal lobule; dmPFC, dorsomedial prefrontal cortex; L, left hemisphere; R, right hemisphere

Behaviorally, faces paired with the distrust instruction were subsequently evaluated as less trustworthy. This finding is consistent with evidence that negative social cues are more salient than positive social cues (Baumeister et al., 2001; Peeters & Czapinski, 1990). Recent behavioral research on the processing of ambiguous facial expressions supports the initial-negativity hypothesis, which suggests that ambiguous emotional stimuli are initially categorized as negative, and that positive category decisions are made by "overriding" this default impulse (Neta, Davis, & Whalen, 2011; Neta & Whalen, 2010; Tottenham, Phuong, Flannery, Gabard-Durnam, & Goff, 2013). Thus, positive evaluations add an extra layer of regulatory influence. Overall, the initial-negativity hypothesis describes that negative evaluations of ambiguous emotional stimuli may reflect the engagement of *automatic*, *reflexive processes*, whereas positive evaluations of ambiguous emotional stimuli may reflect the engagement of *regulatory control processes*. These observations and interpretation are consistent with evolutionary models of human social cognition describing the prioritization of negative social information over positive social information (Baumeister et al., 2001). Combined, the present findings suggest that distrust may be a more salient negative social

cue. It may therefore be processed more automatically, quickly, and strongly than cues to trust.

Within the brain, increased TPJ, mPFC, IFG, PrC, and insula activity was observed during the control of trust and distrust than during the evaluation of age. These brain regions encompass the mentalizing system, which is thought to facilitate the way that the mental states of others are read and interpreted, such as during theory of mind and perspective taking (Frith & Frith, 2006; Van Overwalle & Baetens, 2009). The TPJ is involved in theory of mind (Apperly, Samson, Chiavarino, & Humphreys, 2004; Delgado et al., 2005; Fletcher et al., 1995; Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004; van Veluw & Chance, 2014)—that is, in thinking about the mental states of others. The present findings strengthen existing models associating the TPJ with the interpretation of mental states. When participants were asked to think about intentions (either positive or negative), they exhibited greater TPJ activity than when they were evaluating the age of people in photographs.

Extant data have demonstrated that the mPFC/vmPFC is involved in mentalizing, and in particular in determining boundaries between the self and others (Amodio & Frith, 2006). The vmPFC was more active during the control-of-trust condition than during the control-of-distrust condition. Recent evidence has linked the vmPFC with the way that trust-based decisions are made. For example, individuals with damage to the vmPFC show increased risk-taking during the trust task and decreased reciprocity, as measured by lower back-transfers during the trust game (Moretto, Sellitto, & di Pellegrino, 2013), and we have recently shown that vmPFC gray matter volume is associated with individual differences in the tendency to trust others (Haas, Ishak, Anderson, & Filkowski, 2015). The mPFC has also been implicated in emotion regulation (Goldin et al., 2008; Ochsner & Gross, 2005; Ochsner et al., 2004). Together, these findings support the role of the vmPFC in social evaluation, trust-based decision-making, and emotion regulation, and suggests that the vmPFC may be more involved in efforts to control trust than in efforts to control distrust.

Increased insula activity during the control-of-distrust condition was found, relative to the age evaluation condition. This finding is consistent with several studies associating insula activity with the processing of negative social and emotional information, such as pain and anxiety (Simmons, Matthews, Stein, & Paulus, 2004; Wiech et al., 2010). The insula is broadly involved in the subjective experience of emotions (Craig, 2011; Gu, Hof, Friston, & Fan, 2013) and has often been shown to be active during the processing of negative emotions (Duerden, Arsalidou, Lee, & Taylor, 2013). The insula is also involved in interoceptive processing (Craig, 2003), emotional forms of empathy (Gu et al., 2012), and emotion regulation (Goldin et al., 2008; Ochsner & Gross, 2005; Ochsner et al., 2004). In relation to trust, there is evidence that the insula is involved in modifying trust-based decisions (Adolphs, 2002; Castle et al., 2012; Dimoka, 2010), and some recent studies have shown that the insula plays roles in both trust (Killgore et al., 2013) and distrust (Winston et al., 2002). Together, these findings indicate that the insula may facilitate negative emotion attribution and heightened arousal when forming negative interpersonal social attitudes about others.

We did not observe greater amygdala activity during the control-of-distrust condition, where previous trust studies have shown increased amygdala activation in response to faces evaluated as both trustworthy and untrustworthy (Rule et al., 2013; Winston et al., 2002). It is currently unknown how task instructions to control the attitude of distrust may affect the way that the amygdala functions, and this remains an open question for future research. Previous research has shown that explicit cognitive tasks that involve the appraisal and/or reappraisal of faces may modulate amygdala activity (Chen et al., 2006; Habel et al., 2007; Hariri, Mattay, Tessitore, Fera, & Weinberger, 2003; Lange et al., 2003). Therefore, in this task, which involves the explicit evaluation and control of interpersonal social attitudes, we did not predict that the amygdala would be differentially activated between conditions.

In this study we also explored the association between individual differences in change-of-trust scores and brain activity. PrC activity was associated with individual differences in change-of-trust scores during the control-of-trust condition. That is, individuals who evaluated faces as more trustworthy following the control-of-trust task exhibited increased PrC activity relative to individuals who tended not to change their trustworthiness evaluations. The function of the PrC is currently unclear; however, several current theoretical models associate the PrC with mentalizing, social evaluation, and self-awareness (Immordino-Yang, 2011; Uddin, Iacoboni, Lange, & Keenan, 2007). Recent empirical evidence has demonstrated the PrC to be active during the subjective evaluation of social stimuli (Kuzmanovic et al., 2012). Kuzmanovic and colleagues showed that "verbal influence" was associated with greater PrC activity during impression formation. These findings indicate that the PrC may be particularly important to translate cognitive–semantic information to the formation of complex social attitudes, such as whether to trust another person or not.

Individual differences in change-of-distrust scores were associated with left IFG activity. This may indicate increased engagement of cognitive control in individuals who tend to change their distrust evaluations. A broad array of functions are associated with the IFG, including the processing of social context (Norris, Chen, Zhu, Small, & Cacioppo, 2004), judgments of trustworthiness and attractiveness (Bzdok et al., 2012), emotion regulation (Goldin et al., 2008), and trait attribution (Mitchell et al., 2005). Lateralization to the left IFG may be due to the region's involvement in inner speech, as a

part of Broca's area. Specifically, individuals may be engaging in inner speech while they actively think about distrusting the individual depicted. In terms of connectivity, there is evidence that the IFG is functionally connected with the TPJ during theory-of-mind tasks (McCleery, Surtees, Graham, Richards, & Apperly, 2011). The results from this study support the role of the IFG in the cognitive control and the valuation of attitudes of trust and distrust. However, our findings suggest that the left IFG may be particularly important when controlling negatively valenced social attributes (e.g., distrust).

Exploratory functional connectivity analyses designed to identify connected neural networks that are associated with individual differences in the tendency to change trustworthiness evaluations have demonstrated that individuals who changed trustworthiness evaluations tended to show increased PrC connectivity with the mPFC, dlPFC, IPL, ACC, and superior frontal gyrus. The engagement of this broad neural network may represent the way that the PrC functions to consolidate social attitudes. Additionally, it is likely that people varied in the types of strategies used to mentalize about trusting each person depicted in the photograph. Although this was not explicitly explored, future studies should investigate the specific strategies used in this task.

Several important limitations of this study warrant consideration. First, throughout this study, trust and distrust were characterized differently. During the pre- and postscan evaluation tasks, trust and distrust were operationalized along a single continuum, whereas during the fMRI control-of-trust task, trust and distrust were treated as dissociable constructs. It is currently unclear whether trust and distrust can be considered unidimensional or independent from one another. This study provides a basis to investigate the psychological factors that differentially affect evaluations of trust and distrust. Additionally, in using the 7-point Likert scale to obtain a behavioral metric to measure trust, there is a limit on the amount of change that can be achieved from the pre to post trust scores, depending on the initial rating. For example, if a participant initially rated an individual as highly trustworthy (e.g., a rating of 7), an effect of the instruction to trust would no longer be possible, since the participant had already reached the ceiling of trustworthiness.

The study included both men and women. Behaviorally, we found no significant differences in the pre or the post trust evaluation ratings between male and female participants. However, differences in the ratings based on the sex of the stimuli were observed, in that female faces were rated as more trustworthy than male faces. Previous studies have shown differences between the sexes during trust-related tasks (Riedl et al., 2010), and thus it is important to consider the association between the sexes of the rater and target for future studies. Differences within the brain have also been reported between the sexes (Riedl et al., 2010). Although it is not within the scope of the present study, it will be important for future studies to investigate potential neural differences associated with the sex of raters and the sex of the targets using the present paradigm. This study was limited to the evaluation of faces. Additional behavioral measures that involve social interaction may further elucidate the relationship between brain function and control of interpersonal trust. Trustworthiness evaluations were measured via self-report. Future models of trust behavior will be improved by the inclusion of trust tasks that involve conditions in which participants are asked to trust others implicitly and behaviorally (e.g., reaction time judgments).

The face name memory task was used to quantify each participant's ability to remember semantic information paired with images of faces. This task is in accordance with several other studies designed to measure how people remember the names of faces (Chua, Schacter, Rand-Giovannetti, & Sperling, 2007; Miller et al., 2008; Rentz et al., 2011). Given that there was a 1/3 chance of correctly guessing the name, it would be advantageous to quantify memory using other procedures, such as with implicit subconscious methods or with free recall in future studies. The face name memory task may not be ideal to capture all memory processes specific to this task. For example, another approach would be to explicitly ask each participant whether he or she remembered what instruction was paired with each face. Additionally, it may have been advantageous to incorporate trustworthiness judgements of the faces used within the face name memory task within the individual differences analyses in this study. Finally, the control-of-trust task did not involve any behavioral response, and participants were not asked about the strategies they had used during the task. This leaves open the possibility that the strategies that were used to control trust were highly variable. From one perspective, this provided additional support to investigate individual differences; however, it is also considerably challenging to clearly understand mental processes without behavioral metrics during the task.

It is also currently unknown how the results may have been affected if the instruction to trust or distrust came from a different individual. For example, the instruction to trust or distrust may be more salient if it is expressed by a close relative or friend rather than by a stranger or lesser-known acquaintance. Future research is needed to understand how the source of the instructions may affect the control of interpersonal social attitudes.

In conclusion, this study provides evidence that the way in which interpersonal attitudes of trust and distrust are changed relies on brain regions within the mentalizing system, including the TPJ, mPFC, insula, and inferior and lateral frontal cortices. Additionally, individual differences in changing attitudes of trust are associated with PrC activity, and individual differences in changing attitudes of distrust are associated with IFG activity. These findings strengthen existing social–

cognitive brain models and provide a basis for future research on the neuroscience of interpersonal social attitudes.

References

Adolphs, R. (2002). Trust in the brain. *Nature Neuroscience, 5,* 192–193.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7,* 268–277. doi:10.1038/nrn1884

Apperly, I. A., Samson, D., Chiavarino, C., & Humphreys, G. W. (2004). Frontal and temporo-parietal lobe contributions to theory of mind: Neuropsychological evidence from a false-belief task with reduced language and executive demands. *Journal of Cognitive Neuroscience, 16,* 1773–1784.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5,* 323–370. doi:10.1037/1089-2680.5.4.323

Borum, R. (2010). *The science of interpersonal trust*. McLean: Mitre Corp.

Bzdok, D., Langner, R., Caspers, S., Kurth, F., Habel, U., Zilles, K., … Eickhoff, S. B. (2011). ALE meta-analysis on facial judgments of trustworthiness and attractiveness. *Brain Structure and Function*, *215*, 209–223. doi:10.1007/s00429-010-0287-4

Bzdok, D., Langner, R., Hoffstaedter, F., Turetsky, B. I., Zilles, K., & Eickhoff, S. B. (2012). The modular neuroarchitecture of social judgments on faces. *Cerebral Cortex, 22,* 951–961.

Castle, E., Eisenberger, N. I., Seeman, T. E., Moons, W. G., Boggero, I. A., Grinblatt, M. S., … Taylor, S. E. (2012). Neural and behavioral bases of age differences in perceptions of trust. *Proceedings of the National Academy of Sciences*, *109*, 20848–20852. doi:10.1073/pnas.1218518109

Chang, L. J., Doll, B. B., van't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology, 61,* 87–105.

Chen, C. H., Lennox, B., Jacob, R., Calder, A., Lupson, V., Bisbrown-Chippendale, R., … Bullmore, E. (2006). Explicit and implicit facial affect recognition in manic and depressed states of bipolar disorder: A functional magnetic resonance imaging study. *Biological Psychiatry*, *59*, 31–39. doi:10.1016/j.biopsych.2005.06.008

Cho, J. (2006). The mechanism of trust and distrust formation and their relational outcomes. *Journal of Retailing, 82,* 25–35.

Chua, E. F., Schacter, D. L., Rand-Giovannetti, E., & Sperling, R. A. (2007). Evidence for a specific role of the anterior hippocampal region in successful associative encoding. *Hippocampus, 17,* 1071–1080.

Craig, A. D. (2003). Interoception: The sense of the physiological condition of the body. *Current Opinion in Neurobiology, 13,* 500–505.

Craig, A. D. (2011). Significance of the insula for the evolution of human awareness of feelings from the body. *Annals of the New York Academy of Sciences, 1225,* 72–82.

Cunningham, W. A., Raye, C. L., & Johnson, M. K. (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.

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience, 8,* 1611–1618.

Dimoka, A. (2010). What does the brain tell us about trust and distrust? Evidence from a functional neuroimaging study. *MIS Quarterly, 34,* 373–396.

Dimoka, A. (2011). Brain mapping of psychological processes with psychometric scales: An fMRI method for social neuroscience. *NeuroImage, 54,* S263–S271.

Duerden, E. G., Arsalidou, M., Lee, M., & Taylor, M. J. (2013). Lateralization of affective processing in the insula. *NeuroImage, 78,* 159–175.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17,* 124–129. doi:10.1037/h0030377

Emonds, G., Declerck, C. H., Boone, C., Seurinck, R., & Achten, R. (2014). Establishing cooperation in a mixed-motive social dilemma: An fMRI study investigating the role of social value orientation and dispositional trust. *Social Neuroscience, 9,* 10–22.

Emonds, G., Declerck, C. H., Boone, C., Vandervliet, E. J. M., & Parizel, P. M. (2011). Comparing the neural basis of decision making in social dilemmas of people with different social value orientations, a fMRI study. *Journal of Neuroscience, Psychology, and Economics, 4,* 11–24. doi:10.1037/a0020151

Fleeson, W., & Leicht, C. (2006). On delineating and integrating the study of variability and stability in personality psychology: Interpersonal trust as illustration. *Journal of Research in Personality, 40,* 5–20.

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. J., & Frith, C. D. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition, 57,* 109–128. doi:10.1016/0010-0277(95)00692-R

Friston, K., Buechel, C., Fink, G., Morris, J., Rolls, E., & Dolan, R. (1997). Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage, 6,* 218–229.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50,* 531–534. doi:10.1016/j.neuron.2006.05.001

Goldin, P. R., McRae, K., Ramel, W., & Gross, J. J. (2008). The neural bases of emotion regulation: Reappraisal and suppression of negative emotion. *Biological Psychiatry, 63,* 577–586.

Grecucci, A., Giorgetta, C., van't Wout, M., Bonini, N., & Sanfey, A. G. (2013). Reappraising the ultimatum: An fMRI study of emotion regulation and decision making. *Cerebral Cortex, 23,* 399–410.

Gu, X., Gao, Z., Wang, X., Liu, X., Knight, R. T., Hof, P. R., & Fan, J. (2012). Anterior insular cortex is necessary for empathetic pain perception. *Brain, 135,* 2726–2735. doi:10.1093/brain/aws199

Gu, X., Hof, P. R., Friston, K. J., & Fan, J. (2013). Anterior insular cortex and emotional awareness. *Journal of Computational Neurology, 521,* 3371–3388.

Haas, B. W., Ishak, A., Anderson, I. W., & Filkowski, M. M. (2015). The tendency to trust is reflected in human brain structure. *NeuroImage, 107,* 175–181.

Habel, U., Windischberger, C., Derntl, B., Robinson, S., Kryspin-Exner, I., Gur, R. C., & Moser, E. (2007). Amygdala activation and facial expressions: Explicit emotion discrimination versus implicit emotion processing. *Neuropsychologia, 45,* 2369–2377. doi:10.1016/j.neuropsychologia.2007.01.023

Hariri, A. R., Mattay, V. S., Tessitore, A., Fera, F., & Weinberger, D. R. (2003). Neocortical modulation of the amygdala response to fearful stimuli. *Biological Psychiatry, 53,* 494–501.

Immordino-Yang, M. H. (2011). Me, my "self" and you: Neuropsychological relations between social emotion, self-awareness, and morality. *Emotion Review, 3,* 313–315.

Killgore, W. D., Schwab, Z. J., Tkachenko, O., Webb, C. A., DelDonno, S. R., Kipman, M., … Weber, M. (2013). Emotional intelligence correlates with functional responses to dynamic changes in facial trustworthiness. *Social Neuroscience*, *8*, 334–346.

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., … Grafman, J. (2007). Neural correlates of trust. *Proceedings of the National Academy of Sciences*, *104*, 20084–20089.

Kuzmanovic, B., Bente, G., von Cramon, D. Y., Schilbach, L., Tittgemeyer, M., & Vogeley, K. (2012). Imaging first impressions: Distinct neural processing of verbal and nonverbal social information. *NeuroImage, 60,* 179–188. doi:10.1016/j.neuroimage.2011.12.046

Lange, K., Williams, L. M., Young, A. W., Bullmore, E. T., Brammer, M. J., Williams, S. C., … Phillips, M. L. (2003). Task instructions modulate neural responses to fearful facial expressions. *Biological Psychiatry*, *53*, 226–232.

Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. *Academy of Management Review, 23,* 438–458.

Lewicki, R. J., Tomlinson, E. C., & Gillespie, N. (2006). Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *Journal of Management, 32,* 991–1022.

Lieberman, M. D., & Cunningham, W. A. (2009). Type I and Type II error concerns in fMRI research: Re-balancing the scale. *Social Cognitive and Affective Neuroscience, 4,* 423–428. doi:10.1093/scan/nsp052

Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *NeuroImage, 19,* 1233–1239.

Mars, R. B., Sallet, J., Schüffelgen, U., Jbabdi, S., Toni, I., & Rushworth, M. F. (2012). Connectivity-based subdivisions of the human right "temporoparietal junction area": Evidence for different areas participating in different cortical networks. *Cerebral Cortex, 22,* 1894–1903.

McCleery, J. P., Surtees, A. D., Graham, K. A., Richards, J. E., & Apperly, I. A. (2011). The neural and cognitive time course of theory of mind. *Journal of Neuroscience, 31,* 12849–12854.

Miller, S. L., Celone, K., DePeau, K., Diamond, E., Dickerson, B. C., Rentz, D., … Sperling, R. A. (2008). Age-related memory impairment associated with loss of parietal deactivation but preserved hippocampal activation. *Proceedings of the National Academy of Sciences*, *105*, 2181–2186.

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers, 36,* 630–633. doi:10.3758/BF03206543

Mitchell, J. P., Neil Macrae, C., & Banaji, M. R. (2005). Forming impressions of people versus inanimate objects: social-cognitive processing in the medial prefrontal cortex. *Neuroimage, 26*, 251–257.

Moretto, G., Sellitto, M., & di Pellegrino, G. (2013). Investment and repayment in a trust game after ventromedial prefrontal damage. *Frontiers in Human Neuroscience, 7,* 593. doi:10.3389/fnhum.2013.00593

Morishima, Y., Schunk, D., Bruhin, A., Ruff, C. C., & Fehr, E. (2012). Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism. *Neuron, 75,* 73–79.

Neta, M., Davis, F. C., & Whalen, P. J. (2011). Valence resolution of ambiguous facial expressions using an emotional oddball task. *Emotion, 11,* 1425–1433. doi:10.1037/a0022993

Neta, M., & Whalen, P. J. (2010). The primacy of negative interpretations when resolving the valence of ambiguous facial expressions. *Psychological Science, 21,* 901–907.

Norris, C. J., Chen, E. E., Zhu, D. C., Small, S. L., & Cacioppo, J. T. (2004). The interaction of social and emotional processes in the brain. *Journal of Cognitive Neuroscience, 16,* 1818–1829.

Ochsner, K. N., & Gross, J. J. (2005). The cognitive control of emotion. *Trends in Cognitive Sciences, 9,* 242–249. doi:10.1016/j.tics.2005.03.010

Ochsner, K. N., Ray, R. D., Cooper, J. C., Robertson, E. R., Chopra, S., Gabrieli, J. D., & Gross, J. J. (2004). For better or for worse: Neural systems supporting the cognitive down-and up-regulation of negative emotion. *NeuroImage, 23,* 483–499. doi:10.1016/j.neuroimage.2004.06.030

Ochsner, K. N., Silvers, J. A., & Buhle, J. T. (2012). Functional imaging studies of emotion regulation: A synthetic review and evolving model of the cognitive control of emotion. *Annals of the New York Academy of Sciences, 1251,* E1–E24.

Peeters, G., & Czapinski, J. (1990). Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology, 1,* 33–60.

Rentz, D. M., Amariglio, R. E., Becker, J. A., Frey, M., Olson, L. E., Frishe, K., … Sperling, R. A. (2011). Face–name associative memory performance is related to amyloid burden in normal elderly. *Neuropsychologia*, *49*, 2776–2783. doi:10.1016/j.neuropsychologia.2011.06.006

Riedl, R., Hubert, M., & Kenning, P. (2010). Are there neural gender differences in online trust? An fMRI study on the perceived trustworthiness of eBay offers. *MIS Quarterly, 34,* 397–428.

Riedl, R., & Javor, A. (2012). The Biology of Trust. *Journal of Neuroscience, Psychology, and Economics, 5,* 63–91.

Riedl, R., Mohr, P. N., Kenning, P. H., Davis, F. D., & Heekeren, H. R. (2014). Trusting humans and avatars: A brain imaging study based on evolution theory. *Journal of Management Information Systems, 30,* 83–114.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron, 35,* 395–405.

Rilling, J. K., & Sanfey, A. G. (2011). The neuroscience of social decision-making. *Annual Review of Psychology, 62,* 23–48.

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *NeuroImage, 22,* 1694–1703.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5,* 296–320.

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of Personality and Social Psychology, 104,* 409–426. doi:10.1037/a0031050

Simmons, A., Matthews, S. C., Stein, M. B., & Paulus, M. P. (2004). Anticipation of emotionally aversive visual stimuli activates right insula. *NeuroReport, 15,* 2261–2265.

Smith, N. K., Cacioppo, J. T., Larsen, J. T., & Chartrand, T. L. (2003). May I have your attention, please: Electrocortical responses to positive and negative stimuli. *Neuropsychologia, 41,* 171–183.

Todorov, A. (2008). Evaluating faces on trustworthiness. *Annals of the New York Academy of Sciences, 1124,* 208–224.

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: A model based approach. *Social Cognitive and Affective Neuroscience, 3,* 119–127.

Todorov, A., Mende-Siedlecki, P., & Dotsch, R. (2013). Social judgments from faces. *Current Opinion in Neurobiology, 23,* 373–380.

Tottenham, N., Phuong, J., Flannery, J., Gabard-Durnam, L., & Goff, B. (2013). A negativity bias for ambiguous facial-expression valence during childhood: Converging evidence from behavior and facial corrugator muscle responses. *Emotion, 13,* 92–103. doi:10.1037/a0029431

Uddin, L. Q., Iacoboni, M., Lange, C., & Keenan, J. P. (2007). The self and social cognition: The role of cortical midline structures and mirror neurons. *Trends in Cognitive Sciences, 11,* 153–157.

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage, 48,* 564–584.

van Veluw, S. J., & Chance, S. A. (2014). Differentiating between self and others: An ALE meta-analysis of fMRI studies of self-recognition and theory of mind. *Brain Imaging and Behavior, 8,* 24–38.

Watanabe, T., Takezawa, M., Nakawake, Y., Kunimatsu, A., Yamasue, H., Nakamura, M., … Masuda, N. (2014). Two distinct neural mechanisms underlying indirect reciprocity. *Proceedings of the National Academy of Sciences*, *111*, 3990–3995.

Wiech, K., Lin, C.-S., Brodersen, K. H., Bingel, U., Ploner, M., & Tracey, I. (2010). Anterior insula integrates information about salience into perceptual decisions about pain. *Journal of Neuroscience, 30,* 16324–16331.

Winston, J. S., Strange, B. A., O'Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience, 5,* 277–283.

Woo, C.-W., Krishnan, A., & Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage, 91,* 412–419.