Multivoxel Pattern Analysis Reveals a Neural Phenotype for Trust Bias in Adolescents

Sarah M. Tashjian, João F. Guassi Moreira, and Adriana Galván

Abstract

■ The extent to which individuals are inclined to judge unfamiliar others as trustworthy can have important implications for social functioning. Using multivariate pattern analysis, a neural phenotype of trust bias was identified in 48 human adolescents (ages 14–18 years, 26 female). Adolescents who exhibited more similar brain response to faces at the extremes of a trustworthy gradient were more likely to rate neutral faces as trustworthy. This relation between neural pattern representation and trust bias was evinced in the amygdala. Amygdala–insula connectivity dissimilarity to faces at the extremes of the trustworthy gradient was associated with greater trust bias to neutral faces, serving as a distinct circuit-level contributor to decision-making over and above of amygdala pattern similarity. These findings aid understanding of neural mechanisms contributing to individual differences in social evaluations of ambiguity. ■

INTRODUCTION

Humans rapidly infer complex trait characteristics from simply viewing a social counterpart's face (Todorov, Pakrashi, & Oosterhof, 2009). Neural representations of prior experience (van den Bos, van Dijk, & Crone, 2011; Fehr & Camerer, 2007; Sutter & Kocher, 2007) or overt displays of emotion (Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015; Knutson, 1996) can help guide categorization of these social signals as either positive or negative to inform subsequent social interactions. However, the brain has an interesting puzzle to solve under conditions of social ambiguity, when the facial expression is ambiguous (e.g., a neutral face) and there is no prior experience with the face. In these circumstances, there are vast individual differences in whether the face is perceived positively or negatively. Positive and negative information must be simultaneously represented and integrated to facilitate decision-making under conditions of ambiguity (Hsu, Bhatt, Adolphs, Tranel, & Camerer, 2005). Research on ambiguous surprised faces demonstrates that socially ambiguous stimuli provide a relevant model for examining trait-like individual differences in judgment bias (Neta, Kelley, & Whalen, 2013) and that the brain resolves this ambiguity by engaging the amygdala (Kim et al., 2004) and the cingulo-opercular network (Neta et al., 2013). The current study tested whether a greater tendency to distrust under conditions of ambiguity is related to neurally distinct representations of trustworthy and untrustworthy faces in the amygdala, anterior insula, fusiform face area (FFA), and broader face processing network (FaPrN). Trustworthy judgments of the most ambiguous face type (neutral, nonemotional faces) were used as a marker of individual differences in trust biases (the tendency to trust or distrust when no explicit information is available) because they comprise a frequently encountered social stimulus, moreso than surprised faces (Said, Haxby, & Todorov, 2011; Said, Sebe, & Todorov, 2009; Hassin & Trope, 2000; Fridlund, 1994). When evaluating neutral faces, scarce contextual information exists from which to determine a "correct" or "incorrect" judgment, making these stimuli well suited for investigating individual differences in neural contributions to behavior. Differences in trust bias were tested as a function of neural similarity to trustworthy and untrustworthy faces using fMRI and multivoxel pattern analysis (MVPA).

Contemporary developmental models propose that social and motivational changes occurring at the onset of puberty are the impetus for developmental changes in encoding of social information from faces (Scherf, Behrmann, & Dahl, 2012). A key feature of adolescence is social reorientation toward unknown social counterparts (van den Bos, 2013), and developmentally advantageous behaviors, like motivation to engage in social exploration, can be facilitated by biasing social judgments toward trusting unknown others. For example, displays of trusting behavior evoke reciprocal trust and cooperation from others (King-Casas et al., 2005). The general tendency to trust unknown others increases from adolescence to adulthood (Fett, Gromann, Giampietro, Shergill, & Krabbendam, 2014), and greater trust bias has been associated with greater concurrent and longitudinal well-being in adults (Poulin & Haase, 2015), perhaps due in part to greater social connectedness. Trust bias in older

University of California, Los Angeles

Journal of Cognitive Neuroscience 31:11, pp. 1726–1741
https://doi.org/10.1162/jocn_a_01448

adults has been linked to less discrimination at the neural level when rating untrustworthy and trustworthy faces (Castle et al., 2012). Despite research on detection of some complex emotions from faces, including contempt and sexual interest (Motta-Mena & Scherf, 2017), research on generalized trustworthy judgments in adolescents is comparatively sparse. Adolescents are less perceptually sensitive to subtle changes in facial expression and perceive ambiguous facial expressions as being less emotional (i.e., more neutral) than adults (Lee, Perino, McElwain, & Telzer, 2019). Thus, the results of adult studies do not automatically apply to adolescent samples. A bias toward initially inferring positive traits when evaluating others may be advantageous for adolescents by increasing approach behavior helpful for achieving the developmental task of exploring an increasingly complex social environment (Crone & Dahl, 2012). Throughout adolescence, self-reported perceptions of social trust become more stable (Flanagan & Stout, 2010), suggesting establishment of individual differences in social decision-making. The neural systems contributing to these individual differences have yet to be identified, despite the relevance for understanding how the adolescent brain contributes to inferences about ambiguity. The importance of this inquiry is bolstered by work linking decreased judgments of trustworthiness during face evaluation with social anxiety and behavioral avoidance in adults (e.g., Gutiérrez-García & Calvo, 2016).

Face evaluation relies on a distributed network of brain regions that develop during adolescence, including visuoperceptual systems (occipital cortex and FFA; Gobbini & Haxby, 2007; Kanwisher & Yoval, 2006), socioemotional systems (amygdala and insula; Gavert, Friston, Dolan, & Garrido, 2014), and cognitive systems (medial pFC, paracingulate; Fusar-Poli et al., 2009). The amygdala and insula act to imbue stimuli with motivational relevance, including whether to approach or avoid potentially appetitive or aversive stimuli. This process is important when assessing the potential trustworthiness of social counterparts. For example, patients with bilateral amygdala damage demonstrate impaired judgment of trustworthy and untrustworthy faces (Adolphs, Tranel, & Damasio, 1998). The role of the amygdala in representing the affective significance of stimuli is well established. Specifically, the amygdala responds to uncertainty or atypicality in faces (Todorov, 2012) and regulates attention by detecting salient or motivationally relevant stimuli (Adolphs, 2010). In adults, amygdala functioning is associated with social dimensions of face processing predicted to emerge with puberty (Scherf et al., 2012; Rule et al., 2011). The anterior insula has also been implicated in trustworthy judgments, with blunted activation to both trustworthy and untrustworthy faces associated with age-related increases in trust bias in adults (Castle et al., 2012). The insula provides information about aversive stimuli, which signals prefrontal systems involved in allocation of attention and execution of action (Paulus & Stein, 2006), and insular projections to the amygdala convey social information from emotional expressions (Critchley et al., 2000). Functional connectivity between the amygdala and insula tracks with trustworthy ratings such that these regions demonstrate greater functional coupling in response to untrustworthy faces (Kragel, Zucker, Covington, & LaBar, 2015). How these neural systems relate to independent behavioral judgments is an open question. Of particular interest is the relevance of amygdala and insula functioning to judgments of neutrality when no obvious answer exists and scarce information is available from which to form judgments. This study examines this question by investigating how neural pattern response to faces at the poles of a normed trustworthy gradient (Oosterhof & Todorov, 2008) relates to trust bias for neutral faces.

Previous research using traditional univariate methods has identified heightened amygdala activation to both highly trustworthy and highly untrustworthy faces (e.g., Said, Dotsch, & Todorov, 2010). Other studies have found increased amygdala and insula response as perceived trustworthiness decreases (Kragel et al., 2015; Engell, Haxby, & Todorov, 2007; Winston, Strange, O'Doherty, & Dolan, 2002). Although informative for understanding nonmonotonic neural response to faces (Said et al., 2010), this methodology is not well equipped to demonstrate whether the neuronal populations encode the same information when presented with trustworthy and untrustworthy faces. Evidence suggests there is discrete processing in the amygdala whereby distinct populations of neurons support the processing of positive and negative information specifically (for a review, see O'Neill, Gore, & Salzman, 2018). Univariate approaches mask this information. A more precise approach is MVPA, which is used to elucidate the extent to which perceptually different stimuli (e.g., trustworthy and untrustworthy faces) are represented in different ways at the neural level (Etzel, Zacks, & Braver, 2013; Xue et al., 2010). Although MVPA does not have comparable resolution to electrophysiological recordings in animals, it is a more sophisticated way of distinguishing individual differences in patterns of neural activation in humans, which may be more nuanced than group-level effects. Individual differences in the representation of social information at the neural level are an important missing piece of how individuals form judgments about ambiguity. By pairing behavioral data with representational similarity analysis (RSA; Kriegeskorte, Mur, & Bandettini, 2008), this study extends beyond traditional investigations of face evaluation to better understand how the brain undergirds social cognition and individual differences therein.

In the current study, the extent to which perceptually distinct stimuli (i.e., trustworthy and untrustworthy faces; Oosterhof & Todorov, 2008) evoked similar patterns of voxelwise neural activation was assessed. The relative similarity in psychological encoding of those facial extremes was then tested as a predictor of trustworthy inferences about neutral faces. The most contextually scarce face type (neutral) served as the dependent measure to evaluate

individual differences in trust bias. The amygdala was selected as an a priori ROI based on prior literature indicating that the amygdala tracks both valence and arousal (Wang et al., 2017; Phelps & LeDoux, 2005; Kim et al., 2004) and demonstrates heightened activation magnitude to extremes of the trustworthy gradient (Said et al., 2010). The anterior insula was also tested as another relevant socioemotional region implicated in resolving ambiguity (Singer, Critchley, & Preuschoff, 2009), as well as face processing and trustworthy judgments. To isolate contributions from neural regions implicated in socioemotional judgments, the FFA and FaPrN were tested as regions associated with processing configural information about faces (Gobbini & Haxby, 2007). Post hoc beta-series connectivity analyses assessed functional connectivity between the amygdala and insula given prior evidence that functional connectivity between the amygdala and insula modulates with trustworthy ratings (Kragel et al., 2015). To expand upon this prior work, connectivity dissimilarity between states (trustworthy and untrustworthy faces) was examined as a predictor of neutral face bias scores. Univariate replication analyses were also conducted to determine whether amygdala activation magnitude tracked the trustworthy gradient in a quadratic fashion evincing increased response to the poles of the continuum (Said et al., 2010).

We hypothesized that neural representations of poles of the trustworthy gradient would inform evaluations under conditions of ambiguity (neutrality). Based on recent work showing that similar patterns of amygdala activation to faces reflect similar psychological representations of the trustworthiness of those faces (FeldmanHall et al., 2018), greater representational similarity was interpreted as blunted sensitivity to subtle differences in nonemotional faces along the trustworthy gradient such that neuronal representations of trustworthy and untrustworthy faces were less distinct. We proffered competing hypotheses regarding the directionality of the association between neural similarity and behavior. Given adolescence is a period of social exploration, the success of which may be influenced by approach behavior, a lack of differentiation (i.e., greater similarity) may reflect a transfer of trustworthy inferences to untrustworthy faces (and subsequently neutral faces) resulting in greater positivity bias. However, because individuals are primed to appraise emotional ambiguity as negative (Neta & Whalen, 2010) and given that the amygdala consistently activates to negative stimuli (Sergerie, Chochol, & Armony, 2008), a competing hypothesis is that greater similarity to the nonemotional trustworthy and untrustworthy faces in this study may result in an overall negative appraisal of neutral faces (reduced trust bias).

METHODS

Participants

Forty-eight healthy human adolescents between the ages of 14 and 18 years (26 female; M_{age} = 15.83 years, SD = 1.15) completed the study. High school adolescents were tested because of considerable social and neurobiological development during this time (Kragel et al., 2015; Crone & Dahl, 2012), as well as increasing autonomy in development of new social relationships. Sample size was based on prior work using RSA approaches to probe social categorization (Stolier & Freeman, 2016), and the present sample was increased by approximately 50% to permit greater variability and power. Participants were recruited via flyers and prior participation in laboratory studies. After receiving approval from the university's institutional review board, participant eligibility was determined by a phone screening with a parent to ensure all enrolled participants reported no current medical, psychological, or neurological disorders and did not have any conditions contraindicated for scanning (e.g., metal braces). Adolescent participants provided informed written assent, and their parent or guardian provided informed written consent. Participants were treated in accordance with the ethical standards of American Psychological Association. Included participants were a distinct sample not measured repeatedly and were tested between November 2016 and December 2017.

Experimental Procedures

Participants performed a face judgment task (Figure 1) while undergoing MRI. Faces were selected from a published database of face stimuli computer-generated using the FaceGen Modeler program Version 3.1. Each face identity was morphed using a trust computer model (Oosterhof & Todorov, 2008) to create seven versions of each identity that ranged from untrustworthy to trustworthy (Figure 1A). Participants were randomly assigned to a subset of 10 identities. Participants were presented with 70 faces composed of seven face types at each of the points on the trustworthy gradient across 10 identities. All faces were perceptively male and white.

Each face was presented once in a single event-related run, which consisted of 70 trials. Each of the 70 trials consisted of a 1.5-sec face stimulus presentation, followed by a 2.25-sec decision screen instructing participants to classify the face as either "trustworthy" or "not trustworthy" using a button press (Figure 1B). Trials were pseudorandomly shuffled such that no face identity or gradient point was presented more than twice in a row. Trials began with an ISI presented for a jittered duration (max duration = 2.3 sec, min = 1.3, mean = 1.8, durations randomly distributed). After each decision screen, a blank screen was presented for 200 msec. Thirty baseline intertrial intervals (ITIs) were pseudorandomly interspersed between face trials for a jittered duration (max duration = 5 sec, min = 1.5, mean 2.5, durations randomly distributed). The ISIs, blank screens, and ITIs served as baseline contrast for comparison. Inclusion of baseline events optimized the design for detection of the BOLD response function (Dale, 1999). OptSeq2 (Greve, 2002) was used

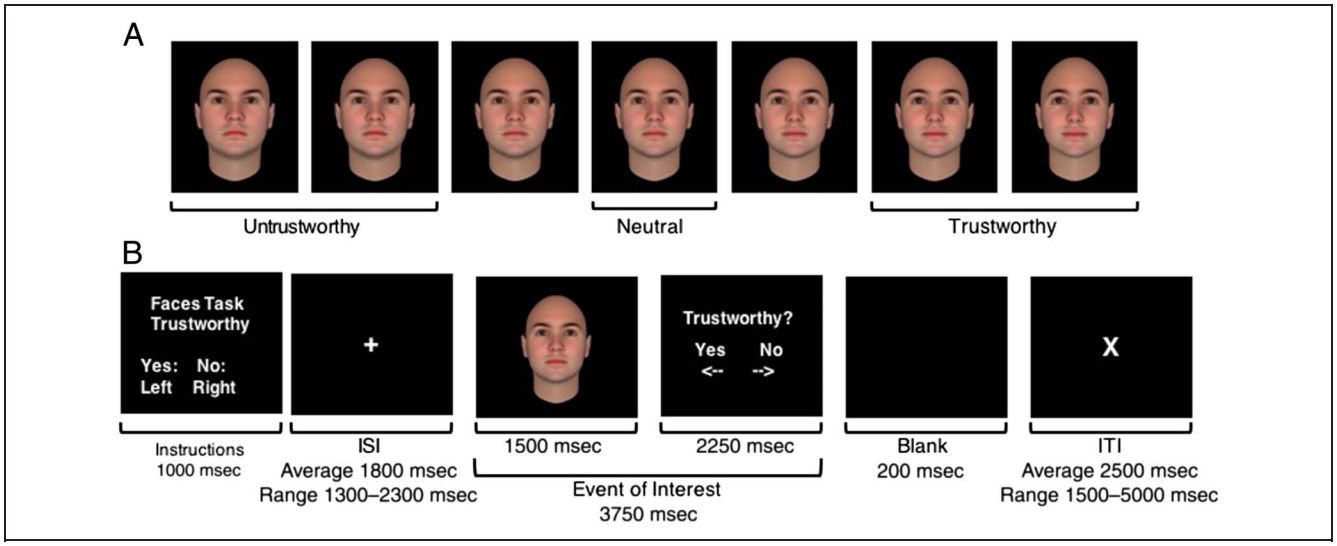

Figure 1. Face judgment task. (A) Example of face stimuli gradient ranging from untrustworthy to trustworthy in accordance with predetermined independent ratings (Oosterhof & Todorov, 2008). (B) Example trial beginning with an ISI presented for a variable jittered duration (max = 2.3 sec, min = 1.3, mean = 1.8, durations randomly distributed). After the ISI, a single face was presented for 1.5 sec, followed by a 2.25-sec response period during which participants judged the preceding face as either "trustworthy" or "not trustworthy" using a button press. Face and decision screens were the event of interest. After the decision offset, a brief blank screen (200 msec) was presented. Pseudorandomly interspersed between trials were ITIs presenting an "X" for a jittered duration (max = 5 sec, min = 1.5, mean = 2.5, durations randomly distributed). ISIs, blank screens, and ITIs served as the baseline contrast to events of interest. Instructions were presented four times throughout the task for three TRs at each presentation and were modeled as events of no interest.

to determine the optimal length and distribution of jittered fixations.

Faces were distributed evenly at seven intervals across an independently predetermined trustworthy gradient (10 faces per gradient point; gradient was morphed at ±3 *SD*s on either side of a normalized average, which was represented as neutral). For analyses, gradient points were coded as −3 to 3. Faces ranked on the lowest 2 points were classified as untrustworthy, whereas faces ranked as the highest 2 points were classified as trustworthy (20/70 faces for each of trustworthy and untrustworthy). Faces ranked as the middle point were classified as neutral (10/70 faces). The trustworthy gradient was determined based on a preestablished criterion (Oosterhof & Todorov, 2008).

Analytic Plan

Before analysis, behavioral and neural analytic models were set, and thresholds for ROIs were defined and reviewed by all authors. p Values below .050 were regarded as statistically significant, and p values between .050 and .100 (inclusive) were regarded as marginally significant.

fMRI Data Acquisition and Analysis

Whole-brain fMRI data were acquired on a 3T Siemens Magnetom Prisma scanner: voxel size = 3.0 × 3.0 × 4.0 mm, slices = 34, slice thickness = 4.0 mm, repetition time (TR) = 2000 msec, echo time = 30.0 msec, flip angle = 90°, interleaved slice geometry, oblique axial orientation, field of view = 192 mm. AutoAlign was used for automated positioning and alignment of anatomy-related slices using alignment perpendicular to the midsagittal plane and tilted along the corpus callosum contour. Images were slice aligned along the anterior/posterior commissure line to allow for interrogation of whole-brain effects (Neta et al., 2013). Structural images were acquired using a high-resolution MPRAGE sequence for registration (TR = 1900 msec, echo time = 2.26 msec, field of view = 250 mm, slice thickness = 1 mm, 176 slices).

Stimuli were presented using E-Prime Professional 2.0 and were projected onto a flat screen mounted in the scanner bore. Participants viewed the screen using a mirror mounted on a 32-channel head coil. Extensive head padding was used to minimize participant head motion and to enhance comfort. Participants made their responses with their right hand using a four-finger button response box. Finger response (index and middle) for trustworthy and not trustworthy judgments, respectively, were randomized across participants.

Regions of Interest

Thresholds were based on anatomical and functional alignment. Images were visually inspected, and Z thresholds were adjusted based on expertise of the authors to conform each ROI to appropriate anatomical locations

and maintain approximately equivalent ROI sizes under 400 voxels (with the exception of the FaPrN ROI, which was larger).

Amygdala. Based on prior work (Said et al., 2010) finding no lateralization effects in response to trustworthy judgments, bilateral amygdala representational similarity was tested. The bilateral amygdala ROI (274 voxels; Figure 4A) was defined using a meta-analysis map (association test) of voxels associated with "amygdala" from the online database NeuroSynth (neurosynth.org; Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011), which contained 54,873 activations from 1579 studies (downloaded October 30, 2018). Neurosynth images are thresholded for a false discovery rate (FDR) criterion of .01. The mask was further thresholded at $Z > 30.0$ and included only voxels also identified as the amygdala by the Harvard–Oxford 50% probability subcortical structural atlas. The amygdala ROI was created in MNI (Montreal Neurological Institute) space (peaks left $x = -22, y = -4, z = -22$, right $x = 26$, $y = -4$, $z = -22$). Standard space masks were transformed to individual functional space using FLIRT linear registration and resampled at 3.0 × 3.0 × 4.0 mm (acquisition parameters).

Anterior insula. To test the possibility that the mechanism responsible for pattern similarity relevance to behavior was salience detection (Uddin, 2015), the anterior insula was also examined. The bilateral insula ROI (398 voxels; Figure 4B) was defined using a meta-analysis map (association test) of voxels associated with "faces" from NeuroSynth, which contained 29,833 activations from 864 studies (downloaded October 30, 2018). NeuroSynth images are thresholded for an FDR criterion of .01. The mask was further thresholded at $Z > 7.0$ and included only voxels also identified as the insula by the Harvard–Oxford 50% probability cortical structural atlas. The insula ROI was created in MNI space (peaks left $x = -36$, $y = 14$, $z = -14$, right $x = 40$, $y = 14$, $z = -12$). Standard space masks were transformed to individual functional space using FLIRT linear registration and resampled at 3.0 × 3.0 × 4.0 mm (acquisition parameters).

Fusiform face area. To test whether results were driven by neural similarity during socioemotional decision-making (amygdala, insula), rather than configural face perception, representational patterns in the FFA were tested. A separate localizer task was not obtained. At the individual level, face-selective regions were defined by the contrast of faces (excluding decision screens) versus baseline fixation. Using a cluster extent threshold of $Z > 2.0$, the peak activation in the FFA was identified and 6-mm spheres (created in subject space 3.0 × 3.0 × 4.0 mm acquisition parameters) were defined around peak coordinates for each participant. The bilateral FFA was active in 41 of the 48 participants (246 voxels; Figure 4C). Five of the full 48 participants showed lateralized FFA activation (four demonstrated left lateralization and one demonstrated right lateralization; 123 voxels). Two participants did not show an FFA response even at lower thresholds and were subsequently excluded from analysis of the FFA ROI but retained for group analysis of whole-brain data and the other ROIs. FFA analyses included the 46 participants for whom the FFA could be identified.

Face processing network. Given that a network of neural regions is involved in face processing, including the amygdala, insula, and FFA, similarity in the canonical FaPrN was examined. The FaPrN ROI (4146 voxels; Figure 4D) was defined using a meta-analysis map (association test) of voxels associated with "faces" from NeuroSynth, which contained 29,833 activations from 864 studies (downloaded October 30, 2018). NeuroSynth images are thresholded for an FDR criterion of .01. The mask was further thresholded at $Z > 10.0$. The FaPrN ROI was created in MNI space. Standard space masks were transformed to individual functional space using FLIRT linear registration and resampled at 3.0 × 3.0 × 4.0 mm (acquisition parameters).

To isolate contributions from the main ROIs of interest, all amygdala and insula voxels were removed from the FaPrN mask, and results were replicated using the FaPrN revised ROI (2728 voxels; Figure 4E).

Preprocessing

For all analyses, preprocessing was conducted using FEAT (fMRI Expert Analysis Tool) Version 6.00, part of FSL (FMRIB Software Library). Preprocessing consisted of nonbrain removal using BET (Brain Extraction Tool), high-pass filtering (100-sec cutoff), and spatial smoothing using a Gaussian kernel of FWHM 5 mm for traditional univariate analyses; data were not spatially smoothed for RSAs. The first three volumes were discarded to allow for image stabilization. Motion correction was performed with MCFLIRT (intramodal motion correction tool) using 24 standard and extended regressors and additional individual spike regressors created using *fsl_motion_outliers* (frame displacement threshold = 75th percentile plus 1.5 times the interquartile range). The average number of spike regressors included was 17.35 (6.59%), ranging from 3 (1.16%) to 38 (14.73%). Average absolute displacement was 0.48 mm, average relative displacement was 0.10 mm. Analyses were duplicated excluding one participant whose average absolute displacement was over 2 mm (3.36 mm) and results remained the same; thus, the full sample of 48 is reported here.

For whole-brain and univariate analyses, functional data were registered to participants' MPRAGE using boundary-based registration (Greve & Fischl, 2009) and then to MNI 2.0 × 2.0 × 2.0 mm stereotaxic space with 12 *df* via FLIRT (FMRIB's Linear Image Registration Tool), consistent with standard univariate analysis procedures. FILM (FMRIB's Improved Linear Model) prewhitening was performed to estimate voxelwise autocorrelation

and improve estimation efficiency. Results for univariate analyses remained the same using ROIs resampled in subject space at 3.0 × 3.0 × 4.0 mm acquisition parameters, with parameter estimates extracted from first level models.

Whole-brain Analysis

Whole-brain analyses were performed to identify neural activation associated with face evaluation. Events were modeled with a canonical (double-gamma) hemodynamic response function. Temporal derivatives were included as covariates of no interest for all regressors. At the individual level, one general linear model (GLM; Friston et al., 1994) was defined for each participant. All contrasts of interest were face presentation events versus baseline (Figure 1B). Face presentation events were modeled for a duration from face stimulus onset to offset of the decision screen (3.75 sec).

One group-level whole-brain analysis was performed using FMRIB Local Analysis of Mixed Effects (FLAME-1) module in FSL (Beckmann, Jenkinson, & Smith, 2003). Outliers were de-weighted in the multisubject statistics using mixture modeling (Woolrich, 2008). A cluster-forming threshold of $Z > 3.1$ and an extent threshold of $p < .05$ familywise error corrected (Poline, Worsley, Evans, & Friston, 1997) were used.

Multivoxel Neural Pattern Analyses

Using CoSMoMVPA (Oosterhof, Connolly, & Haxby, 2016), RSA was conducted as a form of MVPA. Single-trial activation patterns were examined for untrustworthy (20 of 70 trials, 28.57% of the task) and trustworthy faces (20 of 70 trials, 28.57% of the task) using least squares-single methods (Mumford, Turner, Ashby, & Poldrack, 2012). To preserve the fine-grained spatial details required for MVPA, data were not smoothed. Each single-trial GLM included regressors for the face event of interest, all other remaining face events, and all other events of noninterest (e.g., instruction screens). The ISI, ITI, and blank screens (Figure 1B) were not explicitly modeled and therefore served as the implicit baseline. For each participant, voxelwise pattern of amygdala activation represented by z-transformed parameter estimates was extracted on a trial-by-trial basis for each face type (trustworthy, untrustworthy). Pairwise Pearson correlation coefficients were calculated for vectors for all trials, collapsed across face type. Fisher's *r*-to-*z* transformation was then applied as a variance-stabilizing processing step, producing a 40 × 40 similarity matrix for each participant with higher values representing relatively greater similarity and lower values representing relatively greater dissimilarity. The average correlation coefficients collapsed across all comparisons of trustworthy versus untrustworthy faces were used as the independent variable (Visser, Scholte, & Kindt, 2011).

Connectivity Analyses

Functional connectivity between amygdala and insula ROIs was examined using a beta-series approach to construct a time-series for each ROI (Rissman, Gazzaley, & D'Esposito, 2004). Magnitude of task-related BOLD response was estimated separately for each trial using the least squares-single method described above. This approach yields a set of parameter estimates for each trial in every voxel across the whole brain. These values can then be concatenated to form a time series, also known as a beta series. Beta series within each ROI were extracted from each trial-specific GLM resulting in an $n \times p$ matrix for each subject where n is the number of trials (20 trustworthy, 20 untrustworthy) and p is the number of ROIs (two: amygdala and insula). Correlation matrices were constructed separately for connectivity during trustworthy and untrustworthy trials using Pearson's correlation coefficient. Following standardization using the Fisher transform, the connectivity map for untrustworthy faces was subtracted from that of trustworthy faces. The absolute value of the result was taken to examine the dissimilarity of functional coupling between psychological contexts of evaluating trustworthy faces and untrustworthy faces. Higher scores indicate more connectivity dissimilarity between states whereas difference scores of zero reflect identical connectivity values between states.

Univariate Analyses

To examine average activation during untrustworthy and trustworthy trials, two second level fixed-effects voxelwise analyses were created for each participant combining the 20 first level untrustworthy trials and the 20 first level trustworthy trials. Two group-level whole-brain analyses were performed, one for each of the untrustworthy and trustworthy events. Additional group-level analyses were performed for each face type (trustworthy, untrustworthy) to test quadratic amygdala activation across the trustworthy gradient. All group analyses used the FLAME-1 module in FSL (Beckmann et al., 2003). Outliers were deweighted in the multisubject statistics using mixture modeling (Woolrich, 2008). A cluster-forming threshold of $Z > 3.1$ and an extent threshold of $p < .05$ familywise error corrected (Poline et al., 1997) were used. Average ROI activation was extracted using *fslmeants*.

Behavioral Analyses

Behavioral data analyses were conducted to assess whether (1a) the sample evinced a trust or distrust bias across the full task and to neutral faces, (1b) decisions differed as a function of age or sex, (2) decisions differed as a function of the trustworthy gradient consistent with prior work (Oosterhof & Todorov, 2008), (3) neutral face trust bias scores differed as a function of neural response

(3a) pattern similarity to trustworthy and untrustworthy faces (Representational Similarity Results), (3b) connectivity dissimilarity to trustworthy and untrustworthy faces (Connectivity Results), or (3c) activation magnitude to trustworthy or untrustworthy faces (Univariate Results), and (4) univariate activation magnitude differed as a function of the trustworthy gradient consistent with prior work (Said et al., 2010).

Behavioral data analyses were performed using R statistical software (Version 3.5.0) and the lme4 (Bates, Mächler, Bolker, & Walker, 2015; Version 1.1-17) and nlme (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2018; Version 3.1-137) packages.

Bias Score Calculation

Bias scores were calculated for each participant and each face point along the trustworthy gradient. Bias scores were calculated by subtracting the proportion of faces that participants classified as not trustworthy from the proportion of faces that participants classified as trustworthy. Possible scores range from −1 to 1. A zero value represents equal numbers of faces classified as trustworthy and not trustworthy. A negative value represents more faces being classified as not trustworthy than trustworthy, whereas a positive value represents more faces being classified as trustworthy than not trustworthy.

Bias scores were calculated for the entire task and separately for each face type (trustworthy, untrustworthy, neutral). Neutral faces served as the dependent variable in analyses (10 trials, 14.29% of the task; Figure 1A).

Decisions and Trustworthy Gradient

To determine how decisions varied as a function of the trustworthy gradient, linear mixed-effects logistic regression was used with the glmer function in lme4 (link = logit, fitted by Laplace approximation) including random intercepts. Data were analyzed using a multilevel modeling framework because the data consisted of repeated-measures (trials) nested within individuals. Random intercepts were included to account for individual differences in general propensity to bias judgments toward either trustworthy or not trustworthy. Decisions (1 = *trustworthy*, 0 = *not trustworthy*) for the jth participant (j) at the ith trial (i) were modeled as a function of the trustworthy gradient (7 points, −3 to 3) as follows: $\text{Logit}(\text{Decision}_{ij}) = \gamma_{00} + \gamma_{10}\text{Gradient}_{ij} + \gamma_{20}\text{Gradient}_{ij}^2 + u_{0j} + e_{ij}$.

Bias Scores and Neural Response

Regression models were tested for trust biases as a function of neural pattern similarity, connectivity dissimilarity, and average activation magnitude. Linear models with trust bias scores as the dependent variable were tested using the *lm* function. Orthogonal quadratic polynomial models were tested using the *poly* function. Quadratic models were tested given prior work demonstrating nonlinear associations between neural response to face valence (Said et al., 2010; Engell et al., 2007; Winston et al., 2002). Model comparisons were performed using the *anova* function.

Cross-validation of Bias Scores and Neural Response

For significant regression models, caret (Classification and Regression Training; Kuhn, 2008) was used to conduct repeated k-fold cross-validation as a test of out-of-sample error with $k = 10$, 10 repeats, seed = 48. k-Fold cross-validation (Kohavi, 1995) entails drawing multiple subsamples from within the existing data and refitting the regression models to each sample. Observations were randomly partitioned into 10 folds or subsets of roughly equal size. Each model was individually fit using 9 folds (including data for 42–44 participants) of the 10 folds with the first fold (including data for four to six participants) serving as an independent test set for estimating model performance. The first fold was then returned to the training set, and the procedure was repeated until each of the 10 folds were held out. This 10-fold cross-validation was repeated 10 times, resulting in a total of 100 random folds being used to estimate model performance. The results for all 10 folds were averaged to obtain an estimated metric of model performance.

Univariate Activation and Trustworthy Gradient

Prior work demonstrated heightened amygdala activation to faces at the poles of the trustworthy continuum compared with neutral faces at the middle of the continuum (Said et al., 2010). Replication analyses were conducted testing whether univariate amygdala activation differed across the trustworthy gradient; linear and quadratic mixed-effects regression models were tested using the lme function in lme4 with random intercepts. Data were analyzed using a multilevel modeling framework because the data consisted of repeated-measures (trials) nested within individuals. Amygdala activation values for the jth participant (j) at the ith trial (i) were modeled as a function of the trustworthy gradient using orthogonal quadratic polynomial models tested using the *poly* function as follows: $\text{Activation}_{ij} = \gamma_{00} + \gamma_{10}\text{Gradient}_{ij} + \gamma_{20}\text{Gradient}_{ij}^2 + u_{0j} + e_{ij}$.

Data Availability

Study materials, unthresholded statistical maps for Figure 3, ROI masks depicted in Figure 4, and raw data for Figures 2, 5, and 6 are available on the Open Science Framework (OSF; https://osf.io/2uf4w).

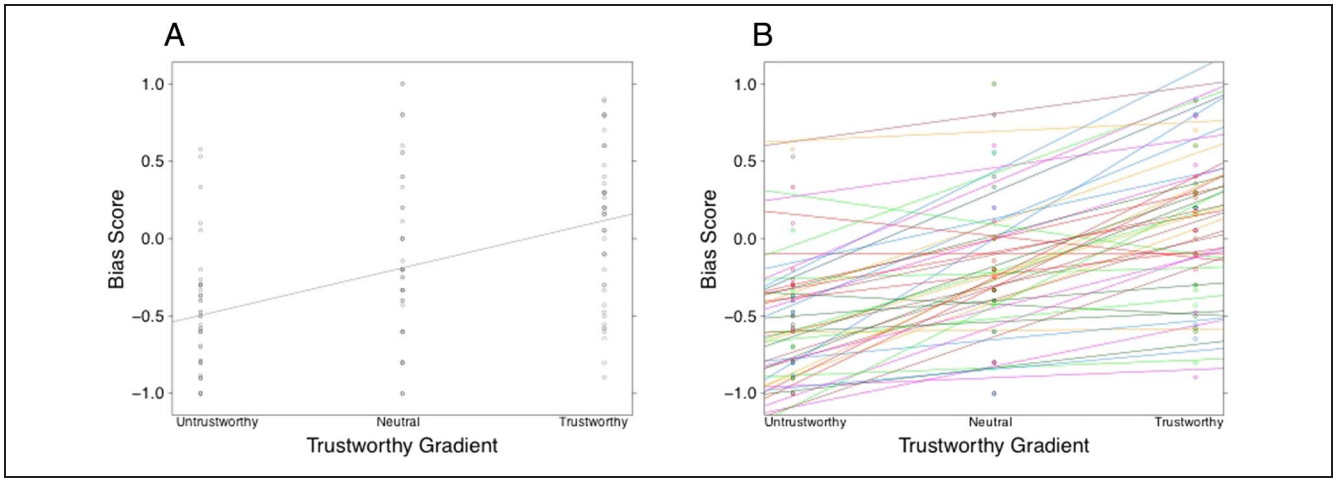

Figure 2. Trust bias scores by face. *x*-Axis represents the independently predetermined trustworthy gradient. *y*-Axis values are raw bias scores: Negative scores indicate a greater number of faces classified as not trustworthy compared with trustworthy, zero scores indicate an equal number of faces classified as trustworthy and not trustworthy, positive scores indicate a greater number of faces being classified as trustworthy compared with not trustworthy. $n = 48$. (A) Average group slope. (B) Slopes for each participant.

RESULTS

Behavioral Results

Decision Bias, Age and Sex

Across the task, participants demonstrated a negative bias indicating fewer faces rated as trustworthy compared with not trustworthy (Table 1). For neutral faces, participants also evinced a negative bias indicating fewer faces rated as trustworthy compared with untrustworthy (Table 1). Neutral face bias scores were significantly positively correlated with untrustworthy, trustworthy, and total bias scores indicating a person-specific individual difference in the tendency to classify faces as trustworthy or untrustworthy during the task (Figure 2B and Table 1).

Age was not significantly correlated with bias scores (linear and quadratic, ps $> .163$). Girls rated a greater proportion of untrustworthy faces as not trustworthy than boys, $M_{female} = -.632$, $M_{male} = -.381$, $t(46) = 2.374$, $p = .022$, $M_{difference} = .252$, $SE_{difference} = .106$, 95% CI [.038, .465], $d = .671$. There were no sex differences for neutral faces, trustworthy faces, or bias scores across the full task, ps $> .417$. As such, age and sex effects were not included in analyses using bias as the dependent variable.

Decisions and Trustworthy Gradient

Multilevel regression revealed that trustworthy judgments (1 = *trustworthy*, 0 = *not trustworthy*) were significantly linearly related to the trustworthy gradient: estimate = 0.319, $SE = 0.021$, $Z = 15.100$, $p < .001$. Testing quadratic associations revealed a significant linear and quadratic effect: linear estimate = 37.956, $SE = 2.513$, $Z = 15.105$, $p < .001$; quadratic estimate = −9.047, $SE = 2.444$, $Z = -3.702$, $p = .004$. Model comparisons revealed the quadratic model was the better fitting model, log likelihood linear model = −1863.1, log likelihood quadratic model = −1856.1, $\chi^2(4) = 14.117$,

Table 1. Bias Scores by Face Type

	Untrustworthy	*Neutral*	*Trustworthy*	*Total*
M	−.517	−.145	.095	−.181
SD	.384	.543	.479	.388
Range	−1 to .578	−1 to 1	−.895 to .900	−.940 to .744
r Untrustworthy	—	.710***	.432**	.796***
r Neutral		—	.705***	.907***
r Trustworthy			—	.848***

$n = 48$. *r* = Pearson bivariate correlation.

*$p < .05$.

**$p < .01$.

***$p < .001$.

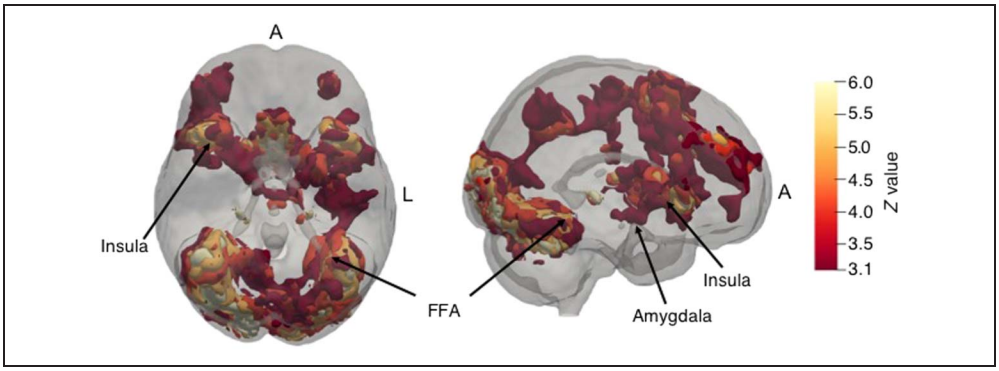

Figure 3. Whole-brain activation of faces versus baseline. Visualization of significant group activation for faces > baseline. FLAME-1, $Z > 3.1$, $p < .05$, outliers deweighted, $n = 48$. L = left hemisphere; A = anterior.

$p < .001$. Judgments of trustworthiness tracked the predetermined trustworthy gradient, although participants were more sensitive to changes at the low end of the gradient. Bias scores increased from negative (more "not trustworthy" decisions) to positive (more "trustworthy" decisions) as faces moved along the predetermined trustworthy gradient from untrustworthy to trustworthy (Figure 2A and Table 1).

Imaging Results

Whole-brain analyses replicated prior work, revealing that faces compared with baseline elicited greater activation in regions identified as important for socioemotional judgments and resolving ambiguity, including the bilateral amygdala (superficial subregion), paracingulate gyrus, inferior frontal gyrus, and bilateral insula. Regions implicated in processing visuoperceptual information from the face, including the occipital lobe and bilateral FFA, were also active, as well as numerous other cortical and subcortical structures (Figure 3 and Table 2).

Representational Similarity Results

To test the prediction that similarity in representation of untrustworthy and trustworthy faces would relate to neutral face judgments, neural profile similarity in response to faces at the extremes of the trustworthy gradient were evaluated using RSA.

Four ROIs (Figure 4A–D) were assessed: (1) amygdala, (2) anterior insula, (3) FFA, and (4) FaPrN. Tests for normality examining skewness and the Shapiro–Wilks test, indicated similarity scores for each ROI were statistically normal, $\text{skew}_{\text{amygdala}} = -0.357$, $\text{kurtosis}_{\text{amygdala}} = 0.718$, $W_{\text{amygdala}} = 0.974$, $p_{\text{amygdala}} = 0.368$, $\text{skew}_{\text{insula}} = -0.304$, $\text{kurtosis}_{\text{insula}} = 0.758$, $W_{\text{insula}} = 0.975$, $p_{\text{insula}} = 0.396$, $\text{skew}_{\text{FFA}} = 0.148$, $\text{kurtosis}_{\text{FFA}} = 0.665$, $W_{\text{FFA}} = 0.970$, $p_{\text{FFA}} = 0.283$, $\text{skew}_{\text{FaPrN}} = 0.064$, $\text{kurtosis}_{\text{FaPrN}} = 1.576$, $W_{\text{FaPrN}} = 0.965$, $p_{\text{FaPrN}} = 0.158$. Age was not associated with similarity in any of these ROIs, linear and quadratic ps > .301. There were no sex differences in similarity, ps > .506.

Amygdala. Neural similarity in the amygdala (Figure 4A) for trustworthy versus untrustworthy faces was significantly related to bias scores for neutral faces. Linear and quadratic associations were tested based on prior literature, but only linear associations were significant (Table 3 and Figure 5A). k-Fold cross-validation with $k = 10$ and 10 repeats (Table 3) was used as an additional test of model fit, which demonstrated comparable out-of-sample fit.

Table 2. Significant Clusters for the Contrast of Face Events Versus Baseline

		Peak MNI Coordinates						
Region	*R/L*	*X*	*Y*	*Z*	*Max Z Value*	*Voxels*	*Volume, mm³*	*p*
Occipital pole	R	14	−92	−4	8.52	23689	189512	<.001
Inferior frontal gyrus	R	48	20	−4	7.13	5642	45136	<.001
Paracingulate gyrus	R	6	20	44	7.75	3938	31504	<.001
Superior parietal lobule; Lateral occipital cortex	R	28	−56	42	4.40	610	4880	<.001
Frontal pole	L	−34	52	18	5.41	459	3672	<.001
Thalamus	R	22	−28	0	6.78	266	2128	<.001
Cingulate gyrus	L	−6	−24	28	4.78	180	1440	.003

$n = 48$. R = right hemisphere; L = left hemisphere.

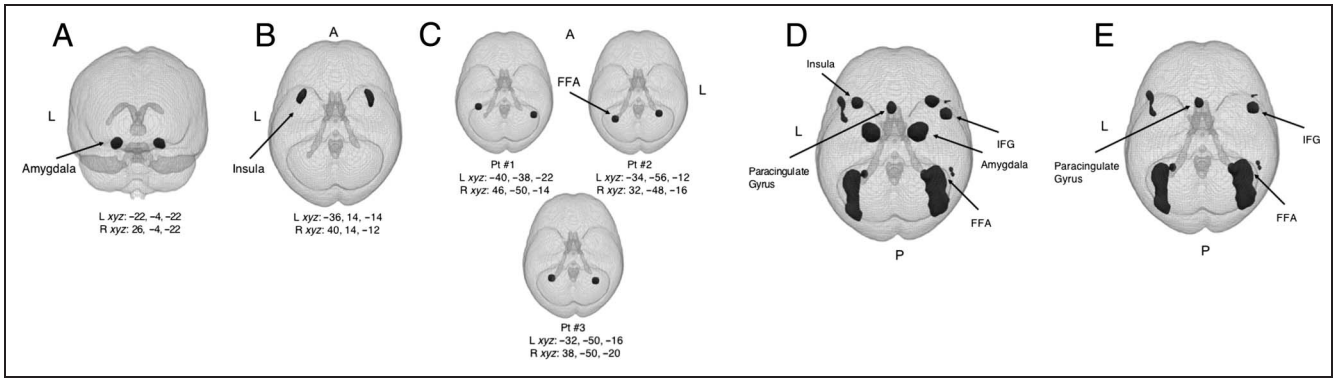

Figure 4. ROIs. (A) Bilateral amygdala ROI defined using a meta-analysis map (association test) of voxels associated with "amygdala anatomical" from the online database NeuroSynth (neurosynth.org; Yarkoni et al., 2011), which contained 54,873 activations from 1579 studies (downloaded October 30, 2018). Neurosynth images are thresholded for an FDR criterion of .01. Masks were further thresholded at $Z > 30.0$ and included only voxels also identified as the amygdala by the Harvard–Oxford 50% probability subcortical structural atlas; 274 voxels. (B) Bilateral insula ROI defined using a meta-analysis map (association test) of voxels associated with "faces" from NeuroSynth, which contained 29,833 activations from 864 studies (downloaded October 30, 2018). NeuroSynth images are thresholded for an FDR criterion of .01. Masks were further thresholded at $Z > 7.0$ and included only voxels also identified as the insula by the Harvard–Oxford 50% probability cortical structural atlas; 398 voxels. (C) Randomly selected example FFA ROI for three participants. The FFA was functionally defined for each participant for the contrast of faces > baseline (excluding decision screens). The FFA was identifiable in 46 participants. After identifying peaks in the FFA, a 6-mm sphere was defined in each hemisphere for each participant (Pt). If participants did not display bilateral FFA, a lateral ROI was defined ($n = 5$); 246 voxels (bilateral), 123 voxels (lateralized). (D) FaPrN ROI defined using a meta-analysis map (association test) of voxels associated with "faces" from NeuroSynth, which contained 29,833 activations from 864 studies (downloaded October 30, 2018). NeuroSynth images are thresholded for an FDR criterion of .01. Masks were further thresholded at $Z > 10.0$; 4146 voxels. (E) FaPrN revised ROI defined as the FaPrN ROI in Figure 4D excluding voxels in the amygdala and insula; 2728 voxels. L = left hemisphere; A = anterior; P = posterior. *xyz* peak coordinates are in MNI space.

Participants who demonstrated greater amygdala differentiation (i.e., less similarity) to faces at the extremes of the trustworthy gradient showed a greater bias toward judging neutral faces as not trustworthy whereas those with a higher similarity score (i.e., more amygdala pattern overlap to extremes) showed a greater bias toward judging neutral faces as trustworthy (Figure 5A).

Amygdala similarity for trustworthy versus untrustworthy faces did not relate to bias scores for either untrustworthy or trustworthy faces, $ps > .233$. Thus, similarity scores were specific to judgments about neutral faces.

Anterior insula. Neural similarity in the insula ROI (Figure 4B) for trustworthy versus untrustworthy faces was marginally significantly ($p = .056$) related to bias scores for neutral faces (Table 3) such that greater differentiation (i.e., less similarity) related to reduced trust bias.

Table 3. Linear Model of Amygdala Pattern Similarity to Trustworthy and Untrustworthy Faces Predicting Neutral Face Bias Scores

	Amygdala ROI				*Insula ROI*				*FFA ROI*				*FaPrN ROI*				*FaPrN Revised ROI*[b]			
	β	*SE*	*t*	*p*	*β*	*SE*	*t*	*p*	*β*	*SE*	*t*	*p*	*β*	*SE*	*t*	*p*	*β*	*SE*	*t*	*p*
Intercept	−.137	.076	−1.817	.076	−.134	.076	−1.758	.085	−.120	.078	−1.549	.129	−.139	.077	−1.816	.076	−.140	.077	−1.807	.077
Similarity[a]	.293	.135	2.170	.035	.280	.143	1.962	.056	.227	.153	1.743	.088	.291	.165	1.776	.084	.267	.171	1.559	.126
R^2	.093				.077				.065				.063				.050			
F	4.711				3.851				3.037				3.119				2.432			
AIC	77.877				78.698				75.441				79.408				80.084			
BIC	83.491				84.312				80.926				85.022				85.698			
RMSE	.512				.516				.515				.520				.523			
k-Fold R^2	.337				.348				.312				.337				.331			
k-Fold RMSE	.509				.523				.521				.522				.529			

$n = 48$, except FFA ROI $n = 46$. AIC = Akaike information criterion, BIC = Bayesian information criterion, RMSE = Root mean square error.

[a] Higher values indicate greater pattern similarity between trustworthy and untrustworthy faces.

[b] Excludes amygdala and insula voxels from FaPrN ROI.

Table 4. Linear Model of Amygdala–Insula Connectivity Dissimilarity between Trustworthy and Untrustworthy Faces Predicting Neutral Face Bias Scores

	Amygdala–Insula Connectivity (No Controls)				*Amygdala–Insula Connectivity (Controlling for Pattern Similarity)*			
	β	*SE*	*t*	*p*	*β*	*SE*	*t*	*p*
Intercept	−.433	.124	−3.498	.001	−.383	.127	−3.028	.004
Amygdala similarity[a]	—				.725	.762	.952	.346
Insula similarity[a]	—				−.548	.793	−.691	.493
Connectivity dissimilarity[b]	.665	.231	2.874	.006	.546	.240	2.280	.027
R^2	.152				.204			
F	8.262				3.754			
AIC	74.629				75.619			
BIC	80.242				84.975			
RMSE	.495				.479			
k-fold R^2	.411				.403			
k-fold RMSE	.489				.490			

n = 48. AIC = Akaike information criterion, BIC = Bayesian information criterion, RMSE = Root mean square error.

[a] Higher values indicate greater pattern similarity between trustworthy and untrustworthy faces.

[b] Absolute values, higher values indicate greater connectivity dissimilarity between trustworthy and untrustworthy faces.

Fusiform face area. To test whether results were based generally on similarity patterns of other brain regions implicated in face processing rather than specific to regions implicated in socioemotional information processing, the association between neutral face bias scores and similarity patterns in the bilateral FFA was examined (n = 46; Figure 4C). Neutral face bias scores were not significantly associated with neural pattern similarity in the FFA to trustworthy versus untrustworthy faces (Table 3).

Face processing network. Making judgments from faces reliably activates a distributed neural processing network consisting of the amygdala, FFA, insula, paracingulate gyrus, inferior frontal gyrus, and occipital cortex

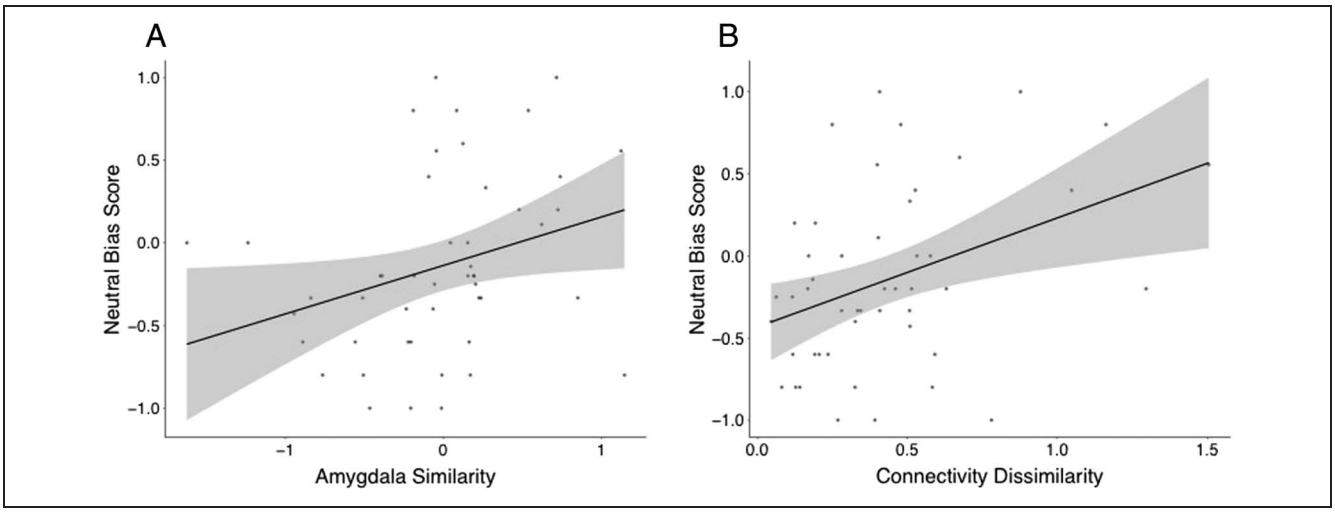

Figure 5. Best-fitting regression models predicting neutral face bias scores. (A) Greater amygdala pattern similarity (*x*-axis, higher values indicate more similarity) related to lower bias scores for neutral faces (*y*-axis, negative values indicate a greater proportion of faces rated as not trustworthy, positive values indicate a greater proportion of faces rated by participants as trustworthy, and zero values indicate an equal proportion of faces rated trustworthy and not trustworthy). (B) Greater connectivity dissimilarity between the amygdala and anterior insula (*x*-axis, higher values indicate more dissimilarity) related to higher bias scores for neutral faces (*y*-axis). *x*-Axis values are Fisher's *Z* values. n = 48. Error represents standard error.

(Figure 4D). Similarity in the broader FaPrN was tested to determine whether contribution of the core system of face perception was associated with trustworthy bias. Similarity in this network to trustworthy and untrustworthy faces did not significantly relate to neutral face bias scores (Table 3). Results remained the same using the FaPrN revised ROI excluding amygdala and insula voxels (Table 3), confirming the importance of neural regions in the extended face processing system implicated in socioemotional decision-making.

Connectivity Results

As a post hoc analysis based on significant amygdala and insula similarity results, functional connectivity between the amygdala and insula was examined as a predictor of neutral face bias scores. Connectivity dissimilarity between states was investigated using a beta-series correlation approach (Rissman et al., 2004). Average activation across voxels in amygdala and insula ROIs was extracted for each trustworthy and untrustworthy trial. Greater connectivity dissimilarity between trustworthy and untrustworthy faces was associated with a greater trust bias (i.e., more neutral faces being categorized as trustworthy; Table 4 and Figure 5B). Connectivity was marginally significantly correlated with amygdala similarity, $r(46) = .270$, $p = .063$, and insula similarity, $r(46) = .247$, $p = .091$. Connectivity differences remained significant controlling for amygdala and insula region-level similarity, representing a distinct metric of neural response (Table 4). Age was not associated with amygdala–insula connectivity dissimilarity, linear and quadratic ps $> .785$. There were no sex differences in connectivity dissimilarity, $p = .804$.

Univariate Results

Univariate analyses replicated prior results (Said et al., 2010) indicating quadratic trends in amygdala activation with heightened activation to more trustworthy faces and more untrustworthy faces compared with faces at the middle of the gradient (Figure 6; using faces at −3, −1, 1, and 3 *SD*s away from the neutral face): linear estimate = 8.322, $SE = 11.524$, $t(142) = 0.722$, $p = .471$; quadratic estimate = 26.705, $SE = 11.524$, $t(142) = 2.317$, $p = .022$. Linear models excluding quadratic terms were not significant. Insula activation did not change as a function of the trustworthy gradient, linear and quadratic $ps > .464$. Age was not associated with univariate activation in the amygdala or insula, linear and quadratic ps $> .216$. There were no sex differences in univariate activation for either ROI, $ps > .183$. Univariate associations in the FFA or FaPrN ROIs were not tested given nonsignificant RSA findings.

Additional analyses tested whether average univariate activation during judgments of trustworthy and untrustworthy faces related to neutral face bias scores.

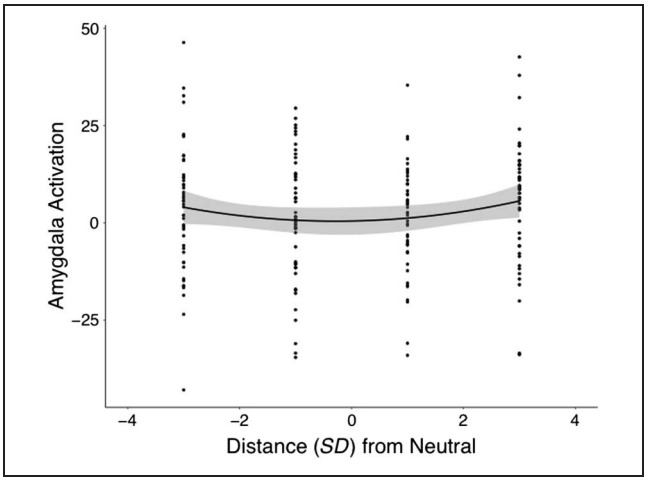

Figure 6. Average univariate amygdala activation by face. Faces points are −3, −1, 1, and 3 standard deviations from the neutral face (0 value) on the independently determined trustworthy gradient. Amygdala activation values are parameter estimates extracted from group-level univariate analyses for each face point for the contrast of faces > baseline. $n = 48$.

Amygdala. Average amygdala activation to trustworthy and untrustworthy faces did not relate to neutral face judgments, $p = .165$, nor did univariate activation to either face type relate to neutral faces judgments, $ps > .175$. Average amygdala activation was marginally significantly correlated with amygdala similarity scores, $r(48) = .243$, $p = .096$. In other words, similarity scores were distinct from average activation, such that participants with greater amygdala similarity in response to trustworthy and untrustworthy faces were not the same participants with greater average activation magnitude to those faces.

Anterior insula. Average insula activation to trustworthy and untrustworthy faces was significantly associated with neutral judgments such that individuals with higher insula activation rated a greater proportion of neutral faces as trustworthy: estimate = .007, $SE = .003$, $t(46) = 2.425$, $p = .019$. This result was driven by activation to untrustworthy faces, estimate = .008, $SE = .003$, $t(46) = 2.950$, $p = .005$, trustworthy faces, $p = .132$. Average insula activation was not significantly correlated with insula similarity scores, $r(48) = .178$, $p = .226$.

DISCUSSION

This study utilized MVPA to identify neural mechanisms linked to trust bias in adolescents; adolescents who demonstrated greater pattern similarity in amygdala responses to trustworthy and untrustworthy faces rated a greater proportion of neutral faces as trustworthy than those with lower pattern similarity. Adolescents who demonstrated greater connectivity dissimilarity between the amygdala and insula to trustworthy and untrustworthy faces rated a greater proportion of neutral faces as

trustworthy than those with less connectivity dissimilarity. Amygdala region pattern similarity and amygdala–insula connectivity dissimilarity were not significantly correlated, indicating representation at the region level and connections at the circuit level make differential contributions to decision-making under conditions of ambiguity. These findings could not be accounted for by gross estimates of amygdala activation magnitude, as pattern similarity and connectivity dissimilarity were not significantly correlated with univariate amygdala activation to untrustworthy and trustworthy faces, representing separate metrics of social judgment tendency. These results suggest the amygdala exhibits a representational structure of trustworthiness that is important for distinguishing individual differences in evaluations of neutrality in adolescents.

This study elucidates otherwise opaque behavioral tendencies: It was previously unclear whether the way individuals represent trustworthiness and untrustworthiness informs their assessments of ambiguity (i.e., neutral faces). Our findings replicated prior univariate findings revealing quadratic associations between amygdala activation magnitude and face trustworthiness (Said et al., 2010), but amygdala activation magnitude to trustworthy and untrustworthy faces was not able to identify how these faces are represented in neuronal population codes, which was relevant for judgments about neutrality. Judgments of trustworthiness were associated with a neural topography composed of two distinct patterns, one for canonical trustworthiness and another for canonical untrustworthiness. The manner in which the trustworthy and untrustworthy representations were calibrated at the individual level related to social judgments of personality traits under conditions of ambiguity (i.e., neutral faces; Cunningham, Van Bavel, & Johnsen, 2008; Somerville, Kim, Johnstone, Alexander, & Whalen, 2004). That the amygdala modulates in response to motivational goals (Canli, Silvers, Whitfield, Gotlib, & Gabrieli, 2002) may be one reason this region emerged as behaviorally relevant, particularly given the importance of exploratory social behavior during adolescence. Interestingly, although humans tend to initially appraise neutrality as negative, the findings suggest that if the two systems are undifferentiated, individuals show a higher trust bias, whereas if the systems are highly differentiated, individuals demonstrate a greater propensity to infer negative intent from neutral emotional expressions. This study did not identify age-related differences in neural response or trust bias during adolescence but rather suggests individual differences in judgments about ambiguous social information are linked to the amygdalar representation of trustworthiness. It is possible that amygdala differentiation of face trustworthiness is experience-dependent, such that trustworthy representations become more differentiated in response to individual differences in one's environment (see Green et al., 2016). This possibility is worthy of continued consideration in developmental studies assessing social appraisal.

The finding that both increased pattern similarity at the regional level and decreased connectivity similarity at the circuit level related to similar behavioral phenotypes may reflect differential roles of neuronal populations within the amygdala and insula projections to the amygdala. Although significant animal and human research points to valence encoding as a mechanism for the observed amygdala pattern similarity finding, less is known about the role of amygdala–insula connectivity in social evaluation. The insula is implicated in a wide variety of emotional processes with a common suggested function being encoding of autonomic changes necessary for conscious emotional awareness (Gu, Hof, Friston, & Fan, 2013). It has been proposed that simultaneous amygdala and insula activation to untrustworthy faces reflects amygdala generation of autonomic changes in bodily states that are mapped in the insula (Winston et al., 2002). The anterior insula also regulates physiological states in addition to perceiving internal changes in the body (Gu et al., 2013). This regulation is one possible explanation for the observed findings, given prior work implicating amygdala–insula connectivity in habituation (Denny et al., 2014). Perhaps individuals with greater neural connectivity differentiation regulated responses to untrustworthy faces to a greater extent such that encoding of trustworthy and untrustworthy faces was more approximate. Our findings support the theory that trait judgments reflect the integration of multiple bottom–up and top–down psychological processes (Stolier, Hehman, & Freeman, 2018), but future work is needed to disentangle directional influences in amygdala–insula circuitry during social evaluation. Additionally, this interpretation should be considered in light of methodological approaches: Pattern similarity was assessed at the voxel level, whereas connectivity dissimilarity was assessed using average activation across all voxels in each ROI.

Trust bias to neutral faces may be socially beneficial, particularly for adolescents as they take on the developmental task of exploring their social environment and forging new relationships. Prior work demonstrating increased trust decision accuracy throughout adolescence has conjectured that a possible mechanism for this developmental change is an improvement in facial processing expertise (De Neys, Hopfensitz, & Bonnefon, 2015). However, pattern similarities in the FFA and FaPrN were not significantly associated with trust bias to neutral faces in this study. Our data suggest functioning of socioemotional systems are better candidate mechanisms for this developmental shift. This proposal is supported by developmental models proposing that social and motivational changes occurring with the onset of puberty are related to developmental changes in encoding of social information from faces (Scherf et al., 2012). Changes in the way neural regions implicated in different aspects of face processing interact may underlie age-related differences in socioemotional processing more generally and individual differences in trust bias, as observed in this

study. Interpretation is bolstered by recent work on fear conditioning, suggesting a lack of BOLD differentiation in the amygdala to conditioned and unconditioned stimuli reflects adaptation mediated by the salience network (including the insula; Yin, Liu, Petro, Keil, & Ding, 2018). Future work should consider how studying neural systems involved in face processing may elucidate phenotypes of other individual differences in social information processing more broadly.

Neural similarity scores did not represent an inability to distinguish among faces at the poles of the trustworthy gradient, meaning that participants were able to reliably rate trustworthy faces as more trustworthy than untrustworthy faces, regardless of similarity score. These findings are relevant for the notion that the amygdala tracks the perceptual similarity and salience of social stimuli, a hypothesis supported by recent work tracking associative learning through patterns of neural representational similarity (FeldmanHall et al., 2018) and first established by foundational social neuroscience studies (e.g., Van Bavel, Packer, & Cunningham, 2008). Despite general proficiency in distinguishing stimuli along the trustworthy gradient, person-specific individual differences were evident in the tendency to classify faces as trustworthy or untrustworthy during the task (Figure 2B). Additionally, pattern similarity was not a measure of trust bias to all stimuli but was specific to neutral faces. Thus, neural similarity was not indicative of a generalized propensity toward negative inference (as might be the case for stereotype encoding; Stolier & Freeman, 2016) but was relevant for individual differences under conditions of highest ambiguity.

Some limitations should be noted. This study was not optimized to investigate amygdala subregions, despite evidence that the amygdala is not a homologous structure (Bzdok, Laird, Zilles, Fox, & Eickhoff, 2013). Additional work is needed to disentangle psychological substrates involved in trait inferences from neutral faces and the role amygdala subnuclei contribute to those complex judgments. Additionally, the task was presented in a single run preventing cross-validation of neuroimaging results. This study was not sufficiently powered to investigate effects of culture or stereotypes known to be relevant for face evaluation, and stimuli in this study were computer-generated perceptually male, white faces. Future work may seek to explore gender, race, and ethnicity interactions potentially pertinent to social inferences. Because of randomization of face presentation during the task, the present methodology does not allow for interpretations about the direction of effects. Although prior work has demonstrated continuing development of trustworthy perceptions through older adulthood (Poulin & Haase, 2015), this study focused on individual differences in trust bias, which are likely to persist despite group-level changes with age. As such, future work may seek to translate this work in a broader age range and range in pubertal status. The current sample was mostly populated by participants in the late stage of puberty so the study was not powered to examine pubertal effects. Although the sample size was modest, *k*-fold cross-validation results provide support for future predictive validity of the findings.

The current findings indicate that greater similarity in amygdala representation of trustworthy and untrustworthy faces is associated with more positively biased judgments in response to social ambiguity during adolescence. Greater dissimilarity in amygdala–insula connectivity to trustworthy and untrustworthy faces was also associated with increased trust bias to neutral faces representing a separate metric of neural functioning relevant for socio-emotional decision-making. The results underscore the complexity of the amygdala: It is not merely absolute levels of activation that are relevant for making social judgments, but the pattern of activation and the circuit-level interaction with other neural regions that might determine whether individuals approach or avoid social counterparts when no context beyond configural perceptual features are available. This study identifies different neural representational schemes for trustworthy and untrustworthy faces as a potential phenotype of trust bias and elucidates the role of the amygdala and insula in contributing to individual differences in response to ambiguity.

Reprint requests should be sent to Sarah M. Tashjian or Adriana Galván, Department of Psychology, University of California, Los Angeles, 1285 Franz Hall, Los Angeles, CA 90095-1563, or via e-mail: smtashjian@ucla.edu, agalvan@ucla.edu.

REFERENCES

Adolphs, R. (2010). What does the amygdala contribute to social cognition? *Annals of the New York Academy of Sciences*, *1191*, 42–61.

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2003). General multilevel linear modeling for group analysis in fMRI. *Neuroimage*, *20*, 1052–1063.

Bzdok, D., Laird, A. R., Zilles, K., Fox, P. T., & Eickhoff, S. B. (2013). An investigation of the structural, connectional, and functional subspecialization in the human amygdala. *Human Brain Mapping*, *34*, 3247–3266.

Canli, T., Silvers, H., Whitfield, S. L., Gotlib, I. H., & Gabrieli, J. D. (2002). Amygdala response to happy faces as a function of extraversion. *Science*, *296*, 2191.

Castle, E., Eisenberger, N. I., Seeman, T. E., Moons, W. G., Boggero, I. A., Grinblatt, M. S., et al. (2012). Neural and behavioral bases of age differences in perceptions of trust. *Proceedings of the National Academy of Sciences, U.S.A.*, *109*, 20848–20852.

Critchley, H., Daly, E., Phillips, M., Brammer, M., Bullmore, E., Williams, S., et al. (2000). Explicit and implicit neural mechanisms for processing of social information from facial expressions: A functional magnetic resonance imaging study. *Human Brain Mapping*, *9*, 93–105.

Crone, E. A., & Dahl, R. E. (2012). Understanding adolescence as a period of social-affective engagement and goal flexibility. *Nature Reviews Neuroscience*, *13*, 636–650.
Cunningham, W. A., Van Bavel, J. J., & Johnsen, I. R. (2008). Affective flexibility evaluative processing goals shape amygdala activity. *Psychological Science*, *19*, 152–160.
Dale, A. M. (1999). Optimal experimental design for event-related fMRI. *Human Brain Mapping*, *8*, 109–114.
De Neys, W., Hopfensitz, A., & Bonnefon, J.-F. (2015). Adolescents gradually improve at detecting trustworthiness from the facial features of unknown adults. *Journal of Economic Psychology*, *47*, 17–22.
Denny, B. T., Fan, J., Liu, X., Guerreri, S., Mayson, S. J., Rimsky, L., et al. (2014). Insula-amygdala functional connectivity is correlated with habituation to repeated negative images. *Social Cognitive and Affective Neuroscience*, *9*, 1660–1667.
Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience*, *19*, 1508–1519.
Etzel, J. A., Zacks, J. M., & Braver, T. S. (2013). Searchlight analysis: Promise, pitfalls, and potential. *Neuroimage*, *78*, 261–269.
Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: The neural circuitry of social preferences. *Trends in Cognitive Sciences*, *11*, 419–427.
FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences, U.S.A.*, *115*, E1690–E1697.
Fett, A. K., Gromann, P. M., Giampietro, V., Shergill, S. S., & Krabbendam, L. (2014). Default distrust? An fMRI investigation of the neural development of trust and cooperation. *Social Cognitive and Affective Neuroscience*, *9*, 395–402.
Flanagan, C. A., & Stout, M. (2010). Developmental patterns of social trust between early and late adolescence: Age and school climate effects. *Journal of Research on Adolescence*, *20*, 748–773.
Fridlund, A. J. (1994). *Human facial expression: An evolutionary view*. San Diego, CA: Academic Press.
Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*, 189–210.
Fusar-Poli, P., Placentino, A., Carletti, F., Landi, P., Allen, P., Surguladze, S., et al. (2009). Functional atlas of emotional face processing: A voxel-based meta-analysis of 105 functional magnetic resonance imaging studies. *Journal of Psychiatry & Neuroscience*, *34*, 418–432.
Gavert, M. M., Friston, K. J., Dolan, R. J., & Garrido, M. I. (2014). Subcortical amygdala pathways enable rapid face processing. *Neuroimage*, *102*, 309–316.
Gobbini, M. I., & Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia*, *45*, 32–41.
Green, S. A., Goff, B., Gee, D. G., Gabard-Durnam, L., Flannery, J., Telzer, E. H., et al. (2016). Discrimination of amygdala response predicts future separation anxiety in youth with early deprivation. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *57*, 1135–1144.
Greve, D. N. (2002). Optseq [Computer program]. Retrieved from surfer.nmr.mgh.harvard.edu/optseq.
Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary based registration. *Neuroimage*, *48*, 63–72.
Gu, X., Hof, P. R., Friston, K. J., & Fan, J. (2013). Anterior insular cortex and emotional awareness. *Journal of Comparative Neurology*, *521*, 3371–3388.
Gutiérrez-García, A., & Calvo, M. G. (2016). Social anxiety and trustworthiness judgments of dynamic facial expressions of emotions. *Journal of Behavior Therapy & Experimental Psychiatry*, *52*, 119–127.
Hassin, R., & Trope, Y. (2000). Facing faces: Studies on the cognitive aspects of physiognomy. *Journal of Personality and Social Psychology*, *78*, 837–852.
Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., & Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision making. *Science*, *310*, 1680–1683.
Kanwisher, N., & Yoval, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *361*, 2109–2128.
Kim, H., Somerville, L. H., Johnstone, T., Polis, S., Alexander, A. L., Shin, L. M., et al. (2004). Contextual modulation of amygdala responsivity to surprised faces. *Journal of Cognitive Neuroscience*, *16*, 1730–1745.
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.
Knutson, B. (1996). Facial expressions of emotion influence interpersonal trait inferences. *Journal of Nonverbal Behavior*, *20*, 165–182.
Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In C. S. Mellish (Ed.), *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 1137–1143) Montreal: Morgan Kaufmann.
Kragel, P. A., Zucker, N. L., Covington, V. E., & LaBar, K. S. (2015). Developmental trajectories of cortical-subcortical interactions underlying the evaluation of trust in adolescence. *Social Cognitive and Affective Neuroscience*, *10*, 240–247.
Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.
Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*, 1–26.
Lee, T.-H., Perino, M. T., McElwain, N. L., & Telzer, E. H. (2019) Perceiving facial affective ambiguity: A behavioral and neural comparison of adolescents and adults. *Emotion*. doi:10.1037/emo0000558
Motta-Mena, N. V., & Scherf, K. S. (2017). Pubertal development shapes perception of complex facial expressions. *Developmental Science*, *20*, e12451.
Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage*, *59*, 2636–2643.
Neta, M., Kelley, W. M., & Whalen, P. J. (2013). Neural responses to ambiguity involve domain-general and domain-specific emotion processing systems. *Journal of Cognitive Neuroscience*, *25*, 547–557.
Neta, M., & Whalen, P. J. (2010). The primacy of negative interpretations when resolving the valence of ambiguous facial expressions. *Psychological Science*, *21*, 901–907.
O'Neill, P. K., Gore, F., & Salzman, C. D. (2018). Basolateral amygdala circuitry in positive and negative valence. *Current Opinion in Neurobiology*, *49*, 175–183.
Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, U.S.A.*, *105*, 11087–11092.
Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU octave. *Frontiers in Neuroinformatics*, *10*, 27.

Paulus, M. P., & Stein, M. B. (2006). An insular view of anxiety. *Biological Psychiatry*, *60*, 383–387.
Phelps, E. A., & LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron*, *48*, 175–187.
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2018). Nlme: Linear and nonlinear mixed effects models (R package version 3.1–137). Retrieved from https://CRAN.R-project.org/package=nlme.
Poline, J. B., Worsley, K. J., Evans, A. C., & Friston, K. J. (1997). Combining spatial extent and peak intensity to test for activations in functional imaging. *Neuroimage*, *5*, 83–96.
Poulin, M. J., & Haase, C. M. (2015). Growing to trust: Evidence that trust increases and sustains well-being across the life span. *Social Psychological and Personality Science*, *6*, 614–621.
Rissman, J., Gazzaley, A., & D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage*, *23*, 752–763.
Rule, N. O., Moran, J. M., Freeman, J. B., Whitfield-Gabrieli, S., Gabrieli, J. D., & Ambady, N. (2011). Face value: Amygdala response reflects the validity of first impressions. *Neuroimage*, *54*, 734–741.
Said, C. P., Dotsch, R., & Todorov, A. (2010). The amygdala and FFA track both social and non-social face dimensions. *Neuropsychologia*, *48*, 3596–3605.
Said, C. P., Haxby, J. V., & Todorov, A. (2011). Brain systems for assessing the affective value of faces. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *366*, 1660–1670.
Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion*, *9*, 260–264.
Scherf, K. S., Behrmann, M., & Dahl, R. E. (2012). Facing changes and changing faces in adolescence: A new model for investigating adolescent-specific interactions between pubertal, brain and behavioral development. *Developmental Cognitive Neuroscience*, *2*, 199–219.
Sergerie, K., Chochol, C., & Armony, J. L. (2008). The role of the amygdala in emotional processing: A quantitative meta-analysis of functional neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, *32*, 811–830.
Singer, T., Critchley, H. D., & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Sciences*, *13*, 334–340.
Somerville, L. H., Kim, H., Johnstone, T., Alexander, A. L., & Whalen, P. J. (2004). Human amygdala responses during presentation of happy and neutral faces: Correlations with state anxiety. *Biological Psychiatry*, *55*, 897–903.
Stolier, R. M., & Freeman, J. B. (2016). Neural pattern similarity reveals the inherent intersection of social categories. *Nature Neuroscience*, *19*, 795–797.
Stolier, R. M., Hehman, E., & Freeman, J. B. (2018). A dynamic structure of social trait space. *Trends in Cognitive Science*, *22*, 197–200.
Sutter, M., & Kocher, M. G. (2007). Trust and trustworthiness across different age groups. *Games and Economic Behavior*, *59*, 364–382.
Todorov, A. (2012). The role of the amygdala in face perception and evaluation. *Motivation and Emotion*, *36*, 16–26.
Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*, 519–545.
Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, *27*, 813–833.
Uddin, L. Q. (2015). Salience processing and insular cortical function and dysfunction. *Nature Reviews Neuroscience*, *16*, 55–61.
Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2008). The neural substrates of in-group bias. *Psychological Science*, *19*, 1131–1139.
van den Bos, W. (2013). Neural mechanisms of social reorientation across adolescence. *Journal of Neuroscience*, *33*, 13581–13582.
van den Bos, W., van Dijk, E., & Crone, E. A. (2011). Learning whom to trust in repeated social interactions: A developmental perspective. *Group Processes & Intergroup Relations*, *15*, 243–256.
Visser, R. M., Scholte, H. S., & Kindt, M. (2011). Associative learning increases trial-by-trial similarity of BOLD-MRI patterns. *Journal of Neuroscience*, *31*, 12021–12028.
Wang, S., Yu, R., Tyszka, J. M., Zhen, S., Kovach, C., Sun, S., et al. (2017). The human amygdala parametrically encodes the intensity of specific facial emotions and their categorical ambiguity. *Nature Communications*, *8*, 14821.
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.
Woolrich, M. (2008). Robust group analysis using outlier inference. *Neuroimage*, *41*, 286–301.
Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J. A., & Poldrack, R. A. (2010). Greater neural pattern similarity across repetitions is associated with better memory. *Science*, *330*, 97–101.
Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*, 665–670.
Yin, S., Liu, Y., Petro, N. M., Keil, A., & Ding, M. (2018). Amygdala adaptation and temporal dynamics of the salience network in conditions fear: A single-trial fMRI study. *eNeuro*, *5*, ENEURO.0445-17.2018.