doi:10.1093/scan/nsn009 SCAN (2008) 3, 119–127

Evaluating face trustworthiness: a model based approach

Alexander Todorov, Sean G. Baron, and Nikolaas N. Oosterhof
Department of Psychology and Center for the Study of Brain, Mind and Behavior, Princeton University, Princeton, NJ 08540, USA

Judgments of trustworthiness from faces determine basic approach/avoidance responses and approximate the valence evaluation of faces that runs across multiple person judgments. Here, based on trustworthiness judgments and using a computer model for face representation, we built a model for representing face trustworthiness (study 1). Using this model, we generated novel faces with an increased range of trustworthiness and used these faces as stimuli in a functional Magnetic Resonance Imaging study (study 2). Although participants did not engage in explicit evaluation of the faces, the amygdala response changed as a function of face trustworthiness. An area in the right amygdala showed a negative linear response—as the untrustworthiness of faces increased so did the amygdala response. Areas in the left and right putamen, the latter area extended into the anterior insula, showed a similar negative linear response. The response in the left amygdala was quadratic—strongest for faces on both extremes of the trustworthiness dimension. The medial prefrontal cortex and precuneus also showed a quadratic response, but their response was strongest to faces in the middle range of the trustworthiness dimension.

Keywords: faces; person perception; trustworthiness; amygdala

People evaluate faces on multiple trait dimensions (Uleman *et al.*, 2005) and these evaluations predict important social outcomes ranging from electoral success (Todorov *et al.*, 2005; Ballew and Todorov, 2007; Little *et al.*, 2007) to sentencing decisions (Blair *et al.*, 2004; Eberhardt *et al.*, 2006). As little as 100 ms exposure to a face is sufficient for people to make a variety of person judgments such as trustworthiness, competence and aggressiveness (Willis and Todorov, 2006). In fact, the minimal time exposure after which people start discriminating between different categories of faces may be as little as 33–38 ms (Bar *et al.*, 2006; Todorov *et al.*, under review).

Although people make multiple person judgments from faces, these judgments are highly correlated with each other, reflecting the valence evaluation that underlies person judgments (Rosenberg *et al.*, 1968; Kim and Rosenberg, 1980). Oosterhof and Todorov (under review) showed that judgments of trustworthiness approximate this valence evaluation. In a series of studies, they elicited spontaneous person descriptions of faces and then identified the most frequent trait dimensions used to describe faces. Judgments on these traits were submitted to a principal components analysis. All positive trait judgments had positive loadings and all negative trait judgments had negative loading on the first principal component, which accounted for more than 60% of the variance of these judgments. Out of 13 different trait judgments, judgments of trustworthiness showed the highest correlation with this component. This correlation was practically unchanged when the principal component was obtained from all other trait judgments except trustworthiness. The correlation between trustworthiness judgments and this component—a linear combination of 12 other trait judgments—was 0.94, indicating that these judgments approximate the valence evaluation underlying multiple social judgments from faces.

This finding is consistent with prior findings about the involvement of the amygdala in the evaluation of faces on trustworthiness. Adolphs *et al.* (1998) showed that patients with bilateral amygdala damage were impaired in discriminating untrustworthy- from trustworthy-looking faces. Two subsequent functional neuroimaging studies confirmed the involvement of the amygdala in face evaluation on trustworthiness. Winston *et al.* (2002) showed that the amygdala response to faces increased as their perceived *untrustworthiness* increased. This was the case for both explicit and implicit (age judgments) evaluation of trustworthiness. Engell *et al.* (2007) used an implicit task and replicated the Winston *et al.* findings.

Whereas the perceived trustworthiness of faces in Winston *et al.* (2002) was assessed by subjective judgments of trustworthiness collected after the fMRI study, Engell *et al.* (2007) used consensus judgments of trustworthiness (averaged across raters) obtained by an independent sample of participants. The amygdala response was better predicted by the consensus judgments of trustworthiness than by the participants' own judgments of trustworthiness (collected after the imaging experiment as in Winston *et al.*' s study). Because consensus judgments reflect properties of the

Received 10 October 2007; Accepted 28 February 2008
Advance Access publication 26 March 2008

We thank Valerie Loehr for her help with the data collection and Andy Engell and Chris Said for their comments on previous versions of this paper. This research was supported by National Science Foundation Grant BCS-0446846.

Correspondence should be addressed to Alexander Todorov, Department of Psychology, Princeton University, Princeton, NJ 08540, USA. E-mail: atodorov@princeton.edu.

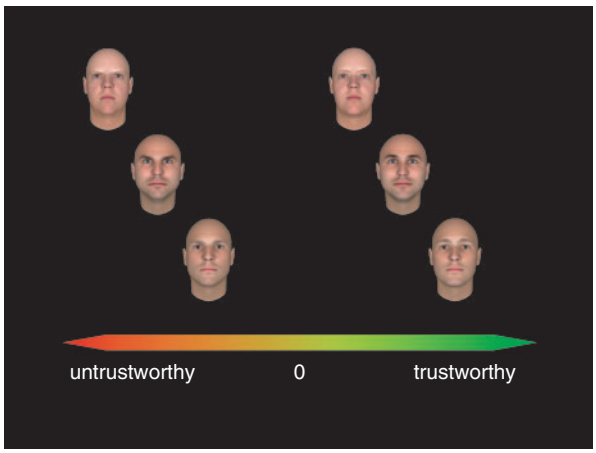

Fig. 1 Examples of faces used in the fMRI experiment. Each of the three rows shows the untrustworthy (on the left) and trustworthy (on the right) versions of a face. Their position on the trustworthiness axis indicates the trustworthiness predicted by the regression model (see text for details).

face rather than idiosyncratic perceptions of the judge (Hönekopp, 2006), Engell *et al.* argued that the amygdala response is driven by structural properties of the face that convey cues for untrustworthiness.

Engell *et al.* used statistical procedures to disentangle the contributions of idiosyncratic perceptions and consensus judgments to the amygdala's response to face trustworthiness. Following on this exploratory analysis, the first objective of this article was to develop a model-based validation approach for testing the role of the amygdala in the evaluation of face trustworthiness. First, we determined what facial features are important for judgments of trustworthiness across participants. Second, we built a 3-dimensional (3D) computer model for representing face trustworthiness based on these features. Third, using this model, we generated trustworthy- and untrustworthy-looking faces (Figure 1). Finally, using functional Magnetic Resonance Imaging (fMRI), we measured how neural activation changes as a function of the trustworthiness of these model-generated faces.

The second objective of the article was to test not only for linear but also for non-linear effects of face trustworthiness on the amygdala response. Specifically, following the computer modeling work of Oosterhof and Todorov (under review) and the findings of Said *et al.* (in press), both described subsequently, we expected that the amygdala might show increased response to faces on both extremes of the trustworthiness dimension.

In research conducted subsequently to study 1, using a data-driven statistical model for face representation, Oosterhof and Todorov built a model for representing face trustworthiness. They argued that face evaluation of emotionally neutral faces is an overgeneralization of functionally adaptive systems for detection of the emotional states of others (Knutson, 1996; Montepare and Dobish, 2003). Specifically, judgments of trustworthiness reflect detection of subtle facial features that resemble emotional expressions signaling approach/avoidance behavior (Todorov, in press). Consistent with this argument, exaggerating the facial features in the negative direction of the trustworthiness dimension produced faces expressing anger, whereas exaggerating the facial features in the positive direction of the dimension produced faces expressing happiness. These expressions signal to the perceiver whether they should avoid or approach the person displaying the emotion (cf., Fridlund, 1994).

Given that several functional neuroimaging studies have found increased amygdala response to happy than to neutral faces (Breiter *et al.*, 1996; Yang *et al.*, 2002; Winston *et al.*, 2003; Pessoa *et al.*, 2006), Oosterhof and Todorov's findings suggest that trustworthy faces can evoke a stronger amygdala response than faces in the middle of the trustworthiness dimension. Said *et al.* (in press) provided a confirmation of this prediction. They modeled both linear and quadratic components of the amygdala response to face trustworthiness and found that the quadratic components provided a better fit of the amygdala response than the linear components. The amygdala response was stronger to both trustworthy and untrustworthy faces than to faces in the middle of the trustworthiness dimension. However, consistent with the previous findings of linear amygdala response to trustworthiness (Winston *et al.*, 2002; Engell *et al.*, 2007), the amygdala response was more sensitive to differences at the negative than at the positive end of the trustworthiness dimension. We sought to replicate this finding with the model-generated faces.

STUDY 1: CREATING A MODEL OF FACE TRUSTWORTHINESS

The objectives of the first study were to: (i) empirically determine the facial features important for judgments of trustworthiness and (ii) build a parsimonious model for manipulating face trustworthiness based on these features. Using a data-driven statistical model of face representation (Blanz and Vetter, 1999; Singular Inversions, 2006), we generated faces with neutral expressions and asked participants to rate these faces on trustworthiness. Then, we regressed the mean trustworthiness judgments on the model values of the four facial features showing the highest correlation with these judgments. Consequently, the regression coefficients estimated from this analysis were used to build a regression model for predicting the trustworthiness of novel faces. We used this model to manipulate the trustworthiness of the faces used in the fMRI study.

Methods

Participants. Twenty-one undergraduate students participated in the behavioral study for partial course credit.

Statistical model of face representation. We used the Facegen Modeller program (http://facegen.com) version 3.1 (Singular Inversions, 2006). Facegen creates 3D faces whose shape and texture can be adjusted on multiple dimensions. The face model of Facegen is based on a database of male

and female human faces that were laser-scanned in 3D. Using Principal Component Analysis (PCA), a model was constructed so that each face can be represented by a limited number of independent components. The components do not correspond to specific facial attributes or features. However, feature controls, which are a linear transformation of the components, resolve this issue. For example, different controls allow for changing the nose (e.g. flat/pointed) and the eyebrows (e.g. down/up inner brow ridge). In contrast to the principal components, which are uncorrelated, the features are correlated and, thus, changing one control value changes other values. For this study, we worked with the 61 symmetric shape (features) controls of Facegen that together give complete control over the underlying components.

Face stimuli. First, we generated 96 Caucasian faces using Facegen. The faces were generated randomly with the following constraints. Facegen's race controls were set so that all faces were European. This was done because a completely random face can be of any race (including Afro-American and Asian) and we wanted to avoid judgments affected by ethnic stereotypes. Additionally, facial attractiveness was increased to make them more similar to the photo-fitted real faces used in Engell *et al.* (2007). Also, we introduced a bias towards male faces, because male faces without hair look more natural than female faces without hair. This bias resulted in mostly typical male faces, with some feminine and some extremely masculine faces. By default, the randomly generated faces are emotionally neutral. Facegen has separate controls for adding the basic emotional expressions: anger, disgust, fear, sadness, happiness and surprise. For all of the randomly generated faces, these expressions were set to neutral. Nevertheless, to further ensure that the expressions were neutral, we also set the mouth shape control, which moves the corners of the mouth up and down, to neutral.

Second, we generated another set of 96 faces derived from the first set of 96 faces by manipulating the eyebrows (lowering or raising the inner brow ridge) and the mouth (the distance between the mouth and the nose) features of each face. This was done because a pilot study determined that these two features are important for judgments of trustworthiness. Thus, for each of the first set of 96 randomly created faces, another face was created by adjusting either the brow ridge inner up/down control (± 2 s.d.), or the mouth up/down control (± 2 s.d.), or both.

Procedures. Participants were told that we were interested in first impressions and that there is no right or wrong answer. Each of the 192 faces was presented once and the order of faces was randomized for each participant. Each face was presented at the center of the screen for 500 ms and was preceded by a 1000 ms fixation cross. The inter-stimulus interval (ISI) was 1000 ms. The response scale ranged from 1 (Very untrustworthy) to 8 (Very trustworthy). The mean judgments averaged across participants were used to find the facial features most predictive for trustworthiness judgments.

Table 1 Zero-order correlations between changes in facial features and judgments of face trustworthiness, and regressions coefficients of changes in facial features as predictors of face trustworthiness

Facial feature	Correlation	Regression coefficient
Brow ridge (down/up)	0.30*	0.13*
Cheekbones (shallow/pronounced)	0.24*	0.13*
Chin (wide/thin)	−0.26*	−0.21*
Nose sellion (shallow/deep)	−0.38*	−0.09

*$P < 0.05$

Results

The trustworthiness judgments were sufficiently reliable, Cronbach's $\alpha = 0.80$. At the first stage of the analysis, we computed the correlations between the mean trustworthiness judgments and each of the 61 feature shape controls. We selected the four facial features that showed the highest correlation with trustworthiness judgments in different face regions (Table 1). Faces with high inner eyebrows, pronounced cheekbones, wide chins and shallow nose sellion looked more trustworthy than faces with low inner eyebrows, shallow cheekbones, thin chins and deep nose sellion. We also selected these features because they showed relatively weak correlations with each other, max $(|r|) = 0.24$.

At the second stage of the analysis, we regressed the mean trustworthiness judgments on the four facial features. This regression analysis was based on the mean judgments of the unambiguously male faces, as judged by three independent raters, because we used only male faces in the fMRI study. The four facial features accounted for 29.4% of the variance of trustworthiness judgments. The coefficients of the regression model (Table 1) were used to predict the trustworthiness of a new set of faces used in the fMRI study.

It should be noted that these predicted trustworthiness values were robust with respect to which faces were used in the regression analysis. A *post hoc* correlation analysis showed that the predicted trustworthiness values were very similar if either all faces (male and female) were used to estimate the regression coefficients ($r = 0.99$), or if only male faces from the original and unmanipulated face set were used ($r = 0.99$).

STUDY 2: NEURAL RESPONSES TO FACE TRUSTWORTHINESS

In this experiment, we used the same implicit task as in Engell *et al.* (2007). Participants ostensibly participated in a face memory task. They were presented with blocks of faces and asked to indicate whether a test face was presented in the block. Thus, the task did not demand explicit person evaluation. We tested for both linear and quadratic effects as a function of face trustworthiness.

Methods

Subjects. Fourteen (seven female) subjects different from the subjects in the behavioral study volunteered for the fMRI study and were paid $30 for their participation. They were between the ages of 18 and 27 (mean = 22.6). All subjects were right-handed, had normal or corrected-to-normal vision and reported no history of neurological illnesses or abnormalities. We acquired informed consent for participation approved by the Institutional Review Board for Human Subjects at Princeton University. All subjects were fully debriefed at the completion of the experiment.

Face stimuli. Ninety new faces were created randomly using the same procedure as the one described in the 'Method' section of study 1. Thirty-three non-ambiguous male faces were selected based on the sex judgments of three independent raters. To increase the variance of face trustworthiness, we used these 33 faces as a basis of 66 new faces: 33 trustworthy and 33 untrustworthy faces (see Figure 1 for examples). For the trustworthy faces, the shape controls with positive coefficient weights (brow ridge and cheekbones) were increased with about 2 s.d. and those with negative coefficient weights (chin and nose sellion) were decreased with about 2 s.d. (because the shape controls are correlated, we were unable to manipulate them with exactly 2 s.d.). For the untrustworthy faces, the shape controls for each feature were changed the same distance but in the opposite direction. To obtain a continuous measure of trustworthiness, we computed the predicted trustworthiness value for each of the 66 faces using the regression model obtained in study 1 (Table 1). These values were centered around zero and used to create the regressors for the fMRI analysis as explained subsequently.

As noted in the introduction, in research conducted after study 1, we have formally modeled a trustworthiness dimension in the 50-dimensional space defined by the 50 symmetric shape components in Facegen (Oosterhof and Todorov, under review). The predicted values from the regression model for the faces used in the fMRI study and the predicted values from the comprehensive trustworthiness model were practically indistinguishable. The correlation was 0.99. Thus, the parsimonious regression model provided a robust representation of the trustworthiness of faces. This high correlation is due to the fact that we expanded the range of trustworthiness by manipulating the facial features important for judgments of trustworthiness and that the feature controls are correlated. That is, changes in the four features manipulated in the current study are linked to changes on a number of other features. In other words, changes in the four features are linked to all 50 underlying principle (shape) components.

Procedures. Subjects were informed that they were participating in a study examining face memory. They were told that they would see six blocks of face images. A block consisted of 11 face images presented in random order. The acquisition run began with a 12 s presentation of a fixation cross. Subsequently, each face stimulus was presented for 1 s in a jittered event-related design. The ISI was chosen randomly from an exponential distribution with a target mean ISI of 3.5 s. The minimum ISI was 1.5 s. Subjects were told to 'do their best' to remember the first 11 face images and that the 12th image would be a 'test' image. They were instructed to indicate whether they remembered the 'test' image from the preceding 11 face images by pressing either a 'yes' or a 'no' button. Each block was separated by a 12-s rest period to allow hemodynamic activation to return to baseline. The order of the face images was randomized for each subject. Stimuli were projected onto a screen located at the rear of the bore of the magnet. Subjects were able to view these stimuli via an angled mirror attached to the RF coil placed above their eyes.

After the scanning session, subjects were led to a computer and asked to judge the 66 faces used in the fMRI session on trustworthiness. The order of the faces was randomized for each subject. Each face was presented at the center of the screen until the subject responded. The response scale ranged from 1 (Very untrustworthy) to 9 (Very trustworthy). We were unable to obtain judgments for one subject because he needed to leave immediately after the scanning session.

Image acquisition. Blood oxygenation level-dependent (BOLD) signal was used as a measure of neural activation. Echo planar images (EPI) were acquired using a Siemens 3.0 Tesla Allegra head-dedicated scanner (Siemens, Erlangen, Germany) with a standard 'bird-cage' head coil (TR = 2000 ms, TE = 30 ms, flip angle = 80°, matrix size = 64 × 64). By using 33 interleaved 3-mm axial slices we were able to achieve near whole brain coverage. Prior to the primary data acquisition scan, a high resolution anatomical image (T1-MPRAGE, TR = 2500 ms, TE = 4.3 ms, flip angle = 8°, matrix size = 256 × 256) was acquired for use in registering functional activity to the subject's anatomy and for spatially normalizing data across subjects.

Image analysis. All fMRI data were analyzed with Analysis of Functional Neuro-images software (AFNI; Cox, 1996). Subjects' motion was corrected using a six-parameter 3D motion-correction algorithm following slice scan-time correction. Data were then low-passed filtered with a frequency cut-off of 0.1 Hz following spatial smoothing with a 6 mm full width at half minimum (FWHM) Gaussian kernel. The signal was then normalized to percent signal change from the mean.

To test for linear and quadratic effects of face trustworthiness on neural responses, we used a polynomial regression (Buchel *et al.*, 1998). We created three time series of interest: a zero-order time series indicating the presence of a face, a first-order time series testing for linear effects of trustworthiness and a second-order time series testing for quadratic effects of trustworthiness. Both the first-order and second-order time series were centered around zero and orthogonalized to each other. The three time series were then convolved with an ideal hemodynamic response function

and entered into the General Linear Model (GLM). The model also included regressors of non-interest: time series representing subject head movement, time-dependent linear and quadratic trends caused by scanner drift and the presentation of the 'test' images.

A *t*-test was performed on the parameter estimates supplied by the GLM for each subject to test for the significance of linear and quadratic estimates across all subjects. We generated group level statistical parametric maps showing voxels that varied linearly with face trustworthiness and voxels that varied quadratically with face trustworthiness. The maps were then thresholded at an uncorrected voxelwise α-level of 0.001. To find out the minimum cluster size for corrected significance of $P < 0.05$, we conducted a whole brain Monte Carlo simulation of null-hypothesis data. These simulations determined that the minimum cluster size was 378 mm^3.

Because we made a priori predictions about the amygdala, we thresholded the statistical maps in each amygdala at an uncorrected voxelwise α-level of 0.05 and then conducted a Monte Carlo simulation in each amygdala. This simulation indicated that a minimum cluster size of 135 mm^3 was required to achieve corrected significance of $P < 0.05$. Because this experiment used the same design as the fMRI experiment conducted by Engell *et al.* (2007), we also conducted a conjunction analysis of the statistical maps of the two experiments after we submitted the Engell *et al.* data to the same GLM analysis. For the conjunction analysis, the maps were thresholded at an α-level of 0.05 and, thus, the resulting conjoint probability was 0.0025. A Monte Carlo simulation in each amygdala determined that the minimum cluster size was 54 mm^3 for corrected significance of $P < 0.05$.

To independently validate the shape of the neural response as a function of face trustworthiness, we defined regions of interests (ROI) and then conducted an additional GLM analysis to extract the signal change for each face. For the amygdala, the ROI were defined by the intersection of the statistical parametric maps with an anatomical mask of the amygdala. We also created functional masks for those regions outside the amygdala that met the criterion for corrected statistical significance.

In the additional analysis, for every subject, regressors for each face image were convolved with an ideal hemodynamic response function and entered into the GLM. The model also included regressors of non-interest: time series representing subject head movement, time-dependent linear and quadratic trends caused by scanner drift and the presentation of the 'test' images. These GLMs provided the parameter estimates for each face presented to every subject. It should be noted that this analysis does not make any assumptions about the shape of the response as a function of face trustworthiness.

Within each ROI, we pulled the mean percent signal change for each face. Because each face was presented only once, we binned the faces into six categories of 11 faces each, ranging from the 11 least trustworthy faces to the 11 most trustworthy faces. The mean signal change across subjects was plotted as a function of these categories in the ROI (Figures 2D, E, 3B and C).

Results

Behavioral judgments. The trustworthiness judgments of the faces collected after the imaging session agreed with the trustworthiness predicted by the model. The correlation between the latter and the mean behavioral judgments was 0.65, $P < 0.001$. Correlation analysis at the level of individual subjects showed that for all subjects but one, the correlation between their judgments and the model trustworthiness was positive. The average correlation 0.35 (s.e. = 0.06) was significantly higher than zero, $t\,(12) = 5.09$, $P < 0.001$.

FMRI results. Replicating Engell *et al.* (2007) findings, a cluster of voxels in the right amygdala showed a significant negative linear trend as a function of face trustworthiness (Figure 2B). The amygdala response to faces increased as the untrustworthiness of faces increased (Figure 2D). A conjunction analysis with the statistical map for the linear trend in Engell *et al.* showed that this cluster was largely overlapping with the cluster showing a negative linear response to the trustworthiness of the real faces used by Engell *et al.* (103 out of 110 mm^3). There was a small cluster of voxels in the left amygdala (31 mm^3) showing the same negative linear trend, but this cluster did not pass the significance criterion adjusted for multiple comparisons.

The only other region besides the amygdala that showed a significant linear response to face trustworthiness and passed the statistical threshold corrected for multiple comparisons was the left putamen (Table 2). Similarly, to the response of the right amygdala, the putamen's response increased as the face untrustworthiness increased. At a reduced threshold of 0.01, a large cluster (1506 mm^3) in the right putamen that extended into the right anterior insula also showed a negative linear response ($P < 0.05$ corrected for multiple comparisons).

The analysis of the quadratic trend showed a significant positive response in a cluster of voxels in the left amygdala (Table 2 and Figure 2C). As shown in Figure 2E, the amygdala response was strongest to both untrustworthy and trustworthy faces, although the response was more elevated for untrustworthy faces. However, the coefficient for the negative linear trend was not significant. This finding replicates the findings of Said *et al.* (in press) who found a similar quadratic response function to face trustworthiness in the amygdala. In the present study, the quadratic response was detectable only in the left amygdala. The cluster in the right amygdala showing a quadratic response was very small (24 mm^3) and did not pass the statistical threshold corrected for multiple comparisons.

The regions other than the left amygdala that showed a quadratic response and survived the correction for multiple comparisons were the medial prefrontal cortex (MPFC) and

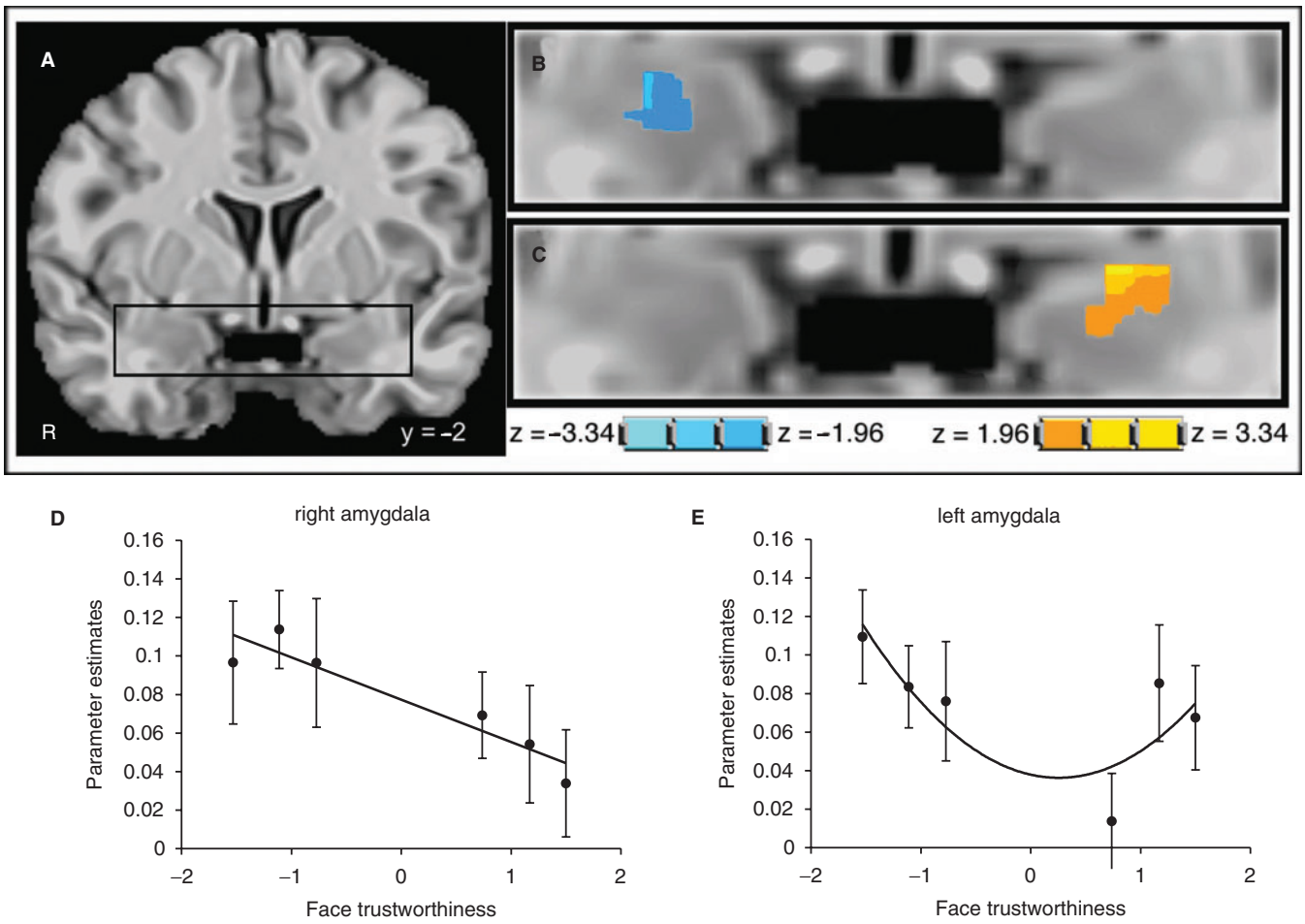

Fig. 2 Amygdala response as a function of face trustworthiness. (A) Amygdala region of a standardized brain. (B) Area in the right amygdala showing a significant negative linear change, this area showed the same linear response in Engell *et al*. (2007). (C) Area in the left amygdala showing a significant quadratic change. The statistical maps show the results of a *t*-test performed on the coefficients of the linear and quadratic trend regressors on the individual data. (D) Parameter estimates (percent signal change) in the functionally defined right amygdala as a function of face trustworthiness. (E) Parameter estimates (percent signal change) in the functionally defined left amygdala as a function of face trustworthiness. For the plots in panels D and E, the faces were binned into six categories according to their trustworthiness. The lines represent the best fitting curves.

the precuneus (Table 2 and Figure 3A). Both of these regions showed a negative quadratic response. As shown in Figure 3B and C, the response was stronger to faces in the middle range of the trustworthiness dimension than to faces at the extremes of the dimension.

DISCUSSION

Evaluating faces on trustworthiness approximates the valence evaluation of faces that underlies multiple trait judgments (Oosterhof and Todorov, under review). In this article, we used a model-based approach to test for the involvement of the amygdala in the implicit evaluation of face trustworthiness. First, based on behavioral data, we built a parsimonious model for representing face trustworthiness. Second, based on this model, we generated novel faces. Third, we used these novel faces in an fMRI study and confirmed the activation of the amygdala as a function of the trustworthiness of faces. Specifically, replicating previous studies (Winston *et al*., 2002; Engell *et al*., 2007), as the untrustworthiness of faces increased so did the response in an area in the right amygdala. Given that participants were never engaged in explicit person evaluation, this finding provides further support for the notion that faces are spontaneously evaluated on trustworthiness (Engell *et al*., 2007).

In addition to right amygdala, we also observed a linear response for bilateral putamen and right anterior insula as a function of face trustworthiness. As the untrustworthiness of faces increased, so did the response in these regions. Winston *et al*. (2002) observed a similar response in the right anterior insula. The amygdala, putamen and anterior insula are often activated in the processing of faces expressing negative emotions (Phillips *et al*., 1997; Sambataro *et al*., 2006; Dannlowski *et al*., 2007). These findings provide additional support for the hypothesis that processing of face trustworthiness is subserved by the mechanisms underlying processing of emotional expressions (Oosterhof and Todorov, under review; Todorov, in press).

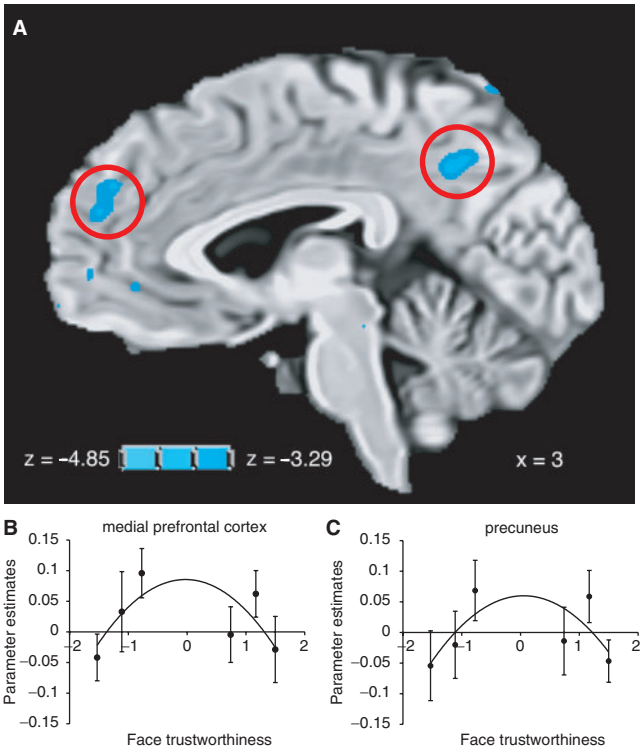

Fig. 3 (A) Regions in the medial prefrontal cortex and precuneus showing significant quadratic effects as a function of face trustworthiness. The results of a *t*-test performed on the coefficients of the quadratic trend regressors on the individual data. (B) Parameter estimates (percent signal change) in the functionally defined MPFC as a function of face trustworthiness. (C) Parameter estimates (percent signal change) in the functionally defined precuneus as a function of face trustworthiness. For the plots in panels B and C, the faces were binned into six categories according to their trustworthiness. The lines represent the best fitting curves.

Table 2 Brain regions responding significantly to face trustworthiness

Regions responding linearly	Volume (mm^3)	*x*	*y*	*z*	*t*-value
Left putamen	402	−16	12	−4	4.48
Right amygdala	110	26	1	−14	2.56
Regions responding quadratically					
Precuneus	478	−1	−57	39	4.34
Medial prefrontal cortex	458	1	58	19	4.30
Left amygdala	271	−21	−2	−10	3.81

The *t*-value for the voxel with maximum activation in the cluster is reported. Coordinates of this voxel are reported in Talairach space.

In contrast to the observed linear response in the right amygdala, the response in an area in the left amygdala changed as a quadratic function of face trustworthiness. That is, the amygdala response was strongest to faces on both extremes of the trustworthiness dimension. This replicates the findings of Said *et al.* (in press). To the extent that judgments of face trustworthiness reflect similarity of facial features to happy and angry expressions, the left amygdala's sensitivity to the extremes of the dimension is consistent with other studies finding a stronger amygdala's response to emotionally expressive faces, independent of the valence of the emotion, than to emotionally neutral faces (Breiter *et al.*, 1996; Yang *et al.*, 2002; Winston *et al.*, 2003; Pessoa *et al.*, 2006). It should also be noted that the pattern of response in the left amygdala suggests that the amygdala's response was more sensitive to differences at the negative than at the positive end of the trustworthiness dimension, although the negative linear trend did not reach significance.

It is interesting to note in this context that the relatively poor discrimination between trustworthy- and untrustworthy-looking faces of bilateral amygdala damage patients is due to a bias to perceive untrustworthy faces as trustworthy (Adolphs *et al.*, 1998). That is, although these patients show an overall positivity bias in judging faces, this bias is especially pronounced for faces at the negative end of the trustworthiness dimension (see also Todorov and Duchaine, in press). This also seems to be the case for people with Asperger syndrome (Adolphs *et al.*, 2001; White *et al.*, 2006). The findings from patient and functional neuroimaging studies suggest that the amygdala is more tuned to detecting differences in the negative than in the positive valence of faces.

In addition to left amygdala, we observed a quadratic response in the MPFC and precuneus. However, these regions showed a stronger response to faces in the middle of the trustworthiness dimension than to untrustworthy and trustworthy faces. These regions are part of the network supporting social cognition processes (Gallagher and Frith, 2003; Amodio and Frith, 2006; Mitchell *et al.*, 2005; Mitchell *et al.*, 2006) and are activated by the presence of familiar faces (Gobbini *et al.*, 2004; Gobbini and Haxby, 2007). One interpretation of their pattern of response is that the faces in the middle range of trustworthiness are relatively more familiar than the faces on the extremes of the dimension. Another interpretation is that it is more difficult to infer the intentions of these faces than faces at the extremes of the dimension and, as a result, these faces engage regions supporting theory of mind inferences. These interpretations remain to be tested.

CONCLUSIONS

We showed that it is possible to construct a model for representing faces on a specific trait dimension and to use computer model-generated 3D faces to search for the neural substrate of face evaluation. This approach has two distinct advantages. First, in contrast to correlation-based exploratory approaches (Engell *et al.*, 2007) in which faces are rated on a trait dimension and then the neural responses are regressed on these ratings, it is a theory validation approach. Second, it allows the investigator to have precise control over the facial stimuli and to generate an unlimited number of faces that vary on a particular dimension of interest. As noted in the introduction, trait judgments from faces are highly correlated with each other. For example, for the set of standardized faces used by Engell *et al.* (2007),

judgments of trustworthiness correlated 0.75 with judgments of attractiveness, −0.76 with judgments of aggressiveness and 0.63 with judgments of intelligence. These high correlations make it difficult to disentangle the contributions of face evaluation on specific dimensions to neural responses. For example, Winston *et al.* (2007) recently found a non-linear amygdala response to facial attractiveness. However, given the high correlation between face trustworthiness and attractiveness, it is possible that this response was driven by the shared variance of attractiveness with trustworthiness. The standard approach is to statistically control for the shared variance among various judgments, but this approach can reduce the statistical power of experiments and, in many cases, it would be difficult to decide on an a priori basis what judgments should be controlled. The alternative to this approach is to experimentally, rather than statistically, unconfound contributions of different dimensions of face evaluation to neural responses. This alternative approach is feasible if the variation of faces on the dimensions of interest can be modeled, as we showed here. Such models can produce an unlimited number of faces varying on specific dimensions and the faces can be orthogonalized on the dimensions of interest.

REFERENCES

Adolphs, R., Sears, L., Piven, J. (2001). Abnormal processing of social information from faces in autism. *Journal of Cognitive Neuroscience*, *13*, 232–40.

Adolphs, R., Tranel, D., Damasio, A.R. (1998). The human amygdala in social judgment. *Nature*, *393*, 470–4.

Amodio, D.M., Frith, C.D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Review Neuroscience*, *7*(4), 268–77.

Ballew, C.C., Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences of the USA*, *104*(46), 17948–53.

Bar, M., Neta, M., Linz, H. (2006). Very first impressions. *Emotion*, *6*(2), 269–78.

Blair, I.V., Judd, C.M., Chapleau, K.M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science*, *15*, 674–9.

Blanz, V., Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. Los Angeles: Addison Wesley Longman, pp. 187–94.

Breiter, H.C., Etcoff, N.L., Whalen, P.J., et al. (1996). Response and habituation of the human amygdala during visual processing of facial expression. *Neuron*, *17*(5), 875–87.

Buchel, C., Holmes, A.P., Rees, G., Friston, K.J. (1998). Characterizing stimulus-response functions using nonlinear regressors in parametric fMRI experiments. *Neuroimage*, *8*(2), 140–8.

Cox, R. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*, 162–73.

Dannlowski, U., Ohrmann, P., Bauer, J., et al. (2007). Amygdala reactivity predicts automatic negative evaluations for facial emotions. *Psychiatry Research: Neuroimaging*, *154*, 13–20.

Eberhardt, J.L., Davies, P.G., Purdie-Vaughns, V.J., Johnson, S.L. (2006). Looking deathworthy: perceived stereotypicality of Black defendants predicts capital-sentencing outcomes. *Psychological Science*, *17*, 383–6.

Engell, A.D., Haxby, J.V., Todorov, A. (2007). Implicit trustworthiness decisions: automatic coding of face properties in human amygdala. *Journal of Cognitive Neuroscience*, *19*, 1508–19.

Fridlund, A.J. (1994). *Human Facial Expression: An Evolutionary View*. San Diego, CA: Academic Press.

Gallagher, H.L., Frith, C.D. (2003). Functional imaging of 'theory of mind.' *Trends in Cognitive Sciences*, *7*, 77–83.

Gobbini, M.I., Haxby, J.V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia*, *45*, 32–41.

Gobbini, M.I., Leibenluft, E., Santiago, N., Haxby, J.V. (2004). Social and emotional attachment in the neural representation of faces. *NeuroImage*, *22*, 1628–35.

Hönekopp, J. (2006). Once more: is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 199–209.

Kim, M.P., Rosenberg, S. (1980). Comparison of two structural models of implicit personality theory. *Journal of Personality and Social Psychology*, *38*, 375–89.

Knutson, B. (1996). Facial expressions of emotion influence interpersonal trait inferences. *Journal of Nonverbal Behavior*, *20*, 165–81.

Little, A.C., Burriss, R.P., Jones, B.C., Roberts, S.C. (2007). Facial appearance affects voting decisions. *Evolution and Human Behavior*, *28*, 18–27.

Mitchell, J.P., Cloutier, J., Banaji, M.R., Macrae, C.N. (2006). Medial prefrontal dissociations during processing of trait diagnostic and nondiagnostic person information. *Social Cognitive and Affective Neuroscience*, *1*, 49–55.

Mitchell, J.P., Macrae, C.N., Banaji, M.R. (2005). Forming impressions of people versus inanimate objects: social-cognitive processing in the medial prefrontal cortex. *NeuroImage*, *26*, 251–7.

Montepare, J.M., Dobish, H. (2003). The contribution of emotion perceptions and their overgeneralizations to trait impressions. *Journal of Nonverbal Behavior*, *27*, 237–54.

Oosterhof, N.N., Todorov, A. (under review). The functional basis of face evaluation.

Pessoa, L., Japee, S., Sturman, D., Underleider, L.G. (2006). Target visibility and visual awareness modulate amygdala responses to fearful faces. *Cerebral Cortex*, *16*, 366–75.

Phillips, M.L., Young, A.W., Senior, C., et al. (1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature*, *389*, 495–8.

Rosenberg, S., Nelson, C., Vivekananthan, P.S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology*, *9*, 283–94.

Said, C.P., Baron, S., Todorov, A. (In press). Nonlinear amygdala response to face trustworthiness: contributions of high and low spatial frequency information. *Journal of Cognitive Neuroscience*.

Sambataro, F., Dimalta, S., Di Giorgio, A.D., et al. (2006). Preferential responses in amygdala and insula during presentation of facial contempt and disgust. *European Journal of Neuroscience*, *24*, 2355–62.

Singular Inversions (2006). FaceGen 3.1 Ful l SDK Documentation. http://facegen.com (last accessed 5 June 2007).

Talairach, J., Tournoux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain*. New York: Thieme.

Todorov, A. (In press). Evaluating faces on trustworthiness: an extension of systems for recognition of emotions signaling approach/avoidance behaviors. In: Miller, M., Kingstone, M.A., editors. *The Year in Cognitive Neuroscience*, *2008*, Vol. 1.

Todorov, A., Duchaine, B. (In press). Reading trustworthiness in faces without recognizing faces. *Cognitive Neuropsychology*.

Todorov, A., Mandisodza, A.N., Goren, A., Hall, C.C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*, 1623–6.

Todorov, A., Pakrashi, M., Loehr, V.R., Oosterhof, N. (Under review). Evaluating faces on trustworthiness: automatic assessment of face valence.

Uleman, J.S., Blader, S., Todorov, A. (2005). Implicit impressions. In: Hassin, R., Uleman, J.S., Bargh, J.A., editors. *The New Unconscious*. New York: Oxford University Press, pp. 362–92.

White, S., Hill, E., Winston, J., Frith, U. (2006). An islet of social ability in Asperger Syndrome: judging social attributes from faces. *Brain and Cognition*, *61*(1), 69–77.

Willis, J., Todorov, A. (2006). First impressions: making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*, 592–8.

Winston, J., O'Doherty, J., Dolan, R.J. (2003). Common and distinct neural responses during direct and incidental processing of multiple facial emotions. *NeuroImage*, *20*, 84–97.

Winston, J.S., O'Doherty, J., Kilner, J.M., Perrett, D.I., Dolan, R.J. (2007). Brain systems for assessing facial attractiveness. *Neuropsychologia*, *45*(1), 195–206.

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*(3), 277–83.

Yang, T.T., Menon, V., Eliez, S., et al. (2002). Amygdalar activation associated with positive and negative facial expressions. *NeuroReport*, *13*, 1737–41.