Automatic and intentional brain responses during evaluation of trustworthiness of faces

J.S. Winston[1], B.A. Strange[2], J. O'Doherty[1] and R.J. Dolan[1,3]

[1] Wellcome Department of Imaging Neuroscience, 12 Queen Square, London WC1N 3BG, UK

[2] Institute of Cognitive Neuroscience, 17 Queen Square, London WC1N 3AR, UK

[3] Royal Free and University College Medical School, Roland Hill Street, London NW3 2PF, UK

Correspondence should be addressed to J.S.W. (j.winston@fil.ion.ucl.ac.uk)

Published online: 19 February 2002, DOI: 10.1038/nn816

Successful social interaction partly depends on appraisal of others from their facial appearance. A critical aspect of this appraisal relates to whether we consider others to be trustworthy. We determined the neural basis for such trustworthiness judgments using event-related functional magnetic resonance imaging. Subjects viewed faces and assessed either trustworthiness or age. In a parametric factorial design, trustworthiness ratings were correlated with BOLD signal change to reveal task-independent increased activity in bilateral amygdala and right insula in response to faces judged untrustworthy. Right superior temporal sulcus (STS) showed enhanced signal change during explicit trustworthiness judgments alone. The findings extend a proposed model of social cognition by highlighting a functional dissociation between automatic engagement of amygdala versus intentional engagement of STS in social judgment.

It is conjectured that human survival has depended to a large extent on accurate social judgments and that, as an evolutionary consequence, modular cognitive processes are devoted to such functions[1]. Neuropsychological studies and human functional imaging provide partial support for this idea of a dedicated 'social intelligence', particularly studies that address perception of facial expression[2–7]. However, facial emotional expression is only one aspect of social judgment made about others. In many situations, individuals must also decide whether another person is someone to approach or avoid, trust or distrust. Preliminary evidence regarding the neural underpinnings of this sort of evaluative judgment comes from studies in which patients with bilateral amygdala lesions make abnormal social judgments about others based on facial appearance[8]. These abnormalities are most pronounced in relation to faces that received the most negative ratings by control subjects. Notably, such deficits are not apparent in subjects with unilateral amygdala lesions[8]. Patients with damage to ventromedial prefrontal cortex also have difficulties with trustworthiness decisions[9,10].

The most influential neurobiological model of social cognition[11], based on inferences largely from neurophysiological recordings in non-human primates, postulates that the superior temporal sulcus acts as association cortex for processing conspecifics' behavior and that socially relevant information is subsequently labeled by the emotional systems, such as amygdala and orbitofrontal cortex. More recent models of human social cognition also include sensory regions such as the face-processing area in fusiform gyrus and somatosensory cortex (including insula, SI and SII)[12–14].

Here we used event-related functional magnetic resonance imaging (fMRI) to ascertain the neural substrates mediating evaluative social judgment. Processing of facial emotion can be implicit, occurring when subjects make judgments about facial attributes unrelated to emotion (for example, refs. 5–7, 15, 16). To establish whether trustworthiness judgments might be similarly processed, we used a task in which subjects viewed faces while making either explicit judgments whether an individual was trustworthy or an unrelated age assessment. To account for individual differences in trustworthiness judgment, we acquired ratings of trustworthiness for each stimulus from each subject after scanning and used these ratings as parametric covariates in our subsequent analysis. Based on the models of social cognition outlined above[11–14], along with the neuropsychological findings[8], we predicted that discrete brain regions, the amygdala, orbitofrontal cortex, fusiform gyrus and superior temporal sulcus, would be implicated in trustworthiness assessments. Consequently, these areas formed regions of interest in our statistical analysis. Our data indicate that social judgments about faces involve such a network and that this network is differentially modulated by implicit and explicit evaluations.

Results

Behavioral

After scanning, on average subjects labeled more than half of the 120 faces as having 'neutral' emotional expressions (mean, 65). Labeled emotional expression interacted significantly with trustworthiness score across the group of subjects (Kruskal-Wallis test, $p < 0.001$). Mann-Whitney U tests showed that the trustworthiness scores (from 1, least trustworthy, to 7) did not differ significantly between 'disgusted,' 'fearful' and 'surprised' faces and 'neutral' faces ($p > 0.05$ in all cases). 'Happy' faces (mean trust-

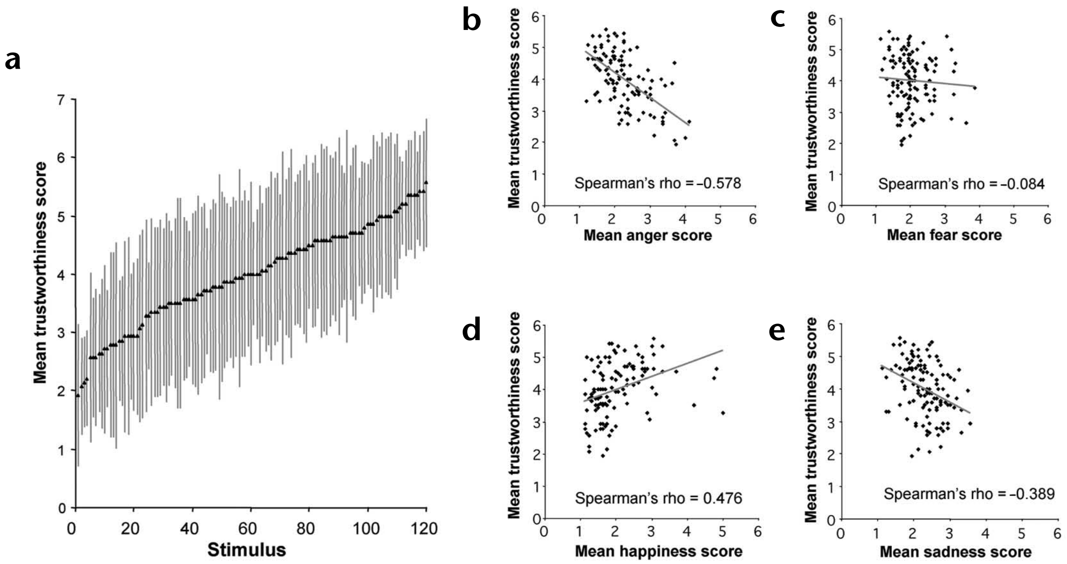


Fig. 1. Trustworthiness and emotion ratings for stimuli. (**a**) Means and standard deviations of trustworthiness scores of stimuli, rank-ordered by trustworthiness score. (**b**–**e**) Mean emotion scores (from second cohort of sixteen subjects) and mean trustworthiness scores (from cohort of subjects scanned with fMRI) for anger (**b**), fear (**c**), happiness (**d**) or sadness (**e**). Lines of best fit are derived by linear regression. Both rating scales ranged from 1 (low degree of emotion or highly untrustworthy) to 7 (highly emotional or highly trustworthy).

worthiness rating, 4.0) were rated as significantly more trustworthy than 'neutral' faces (mean rating, 3.9), and 'angry' (mean rating, 2.7) and 'sad' (mean rating, 3.6) faces as significantly less trustworthy ($p < 0.01$ in all cases). Mean trustworthiness scores (**Fig. 1a**) were significantly correlated with mean scores for anger, happiness and sadness from the second group of subjects (see Methods) ($p < 0.01$ for each, two-tailed; **Fig. 1b–e**).

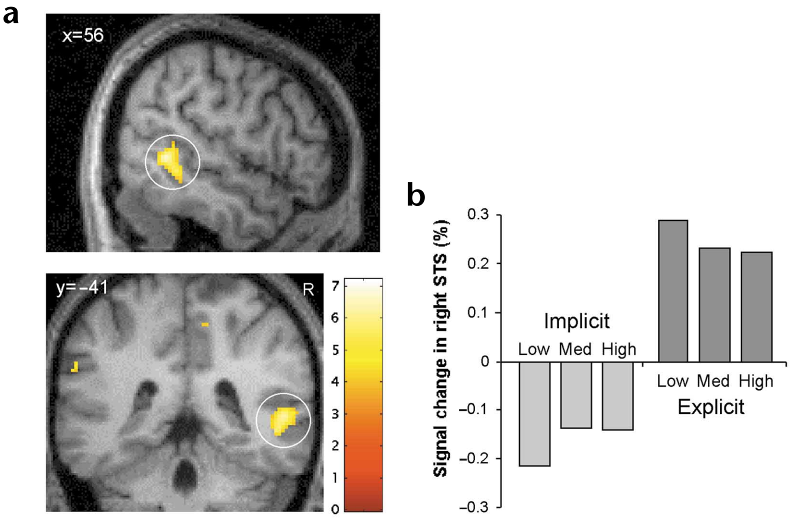
Neuroimaging

Linear contrasts were performed to produce statistical parametric maps (SPMs) of the main effect of task (explicit or implicit processing of trustworthiness), the main effect of trustworthiness and the interaction between these two factors. An additional model in which the effects of facial emotion of the stimuli were included as covariates of no interest was used to generate an SPM related to the main effect of trustworthiness independent of effects of facial emotional expression.

A significant activation in the explicit compared to implicit task, independent of trustworthiness, was found in the right posterior superior temporal sulcus (x, y, z coordinates, 56, –44, 4; $Z = 4.27$; $p < 0.05$, corrected for multiple comparisons across a small volume of interest; **Fig. 2**; **Table 1**). Additionally, primary visual cortex was significantly activated in this contrast. Attentional and emotional manipulations are known to alter neural responses in early visual cortex[17], and we propose that similar processes engendered by the explicit task account for this latter activation.

As predicted, significant bilateral amygdala activation was evident in the contrast of untrustworthy to trustworthy faces (right, –18, 0, –24; $Z = 4.29$; left, –16, –4, –20; $Z = 3.92$; both $p < 0.05$, corrected for multiple comparisons across a small volume of interest; **Fig. 3a**). This examination of parametric data based on each subject's ratings of faces indicates that more untrustworthy faces evoke greater BOLD responses in the amygdala (**Fig. 3c** and **d**).

Fig. 2. Main effect of explicit social judgments. (**a**) Random-effects SPM overlaid on a normalized structural scan from a single subject showing activation in right superior temporal sulcus region (x, y, z = 56, –44, 4; $Z = 4.27$; $p < 0.05$ small volume corrected) when making judgments about trustworthiness compared to age. For illustration, using threshold $p < 0.001$ uncorrected, extent threshold of 5 voxels. (**b**) BOLD signal measure categorized by task and trustworthiness of faces. 'Low,' 'med' and 'high' refer to the least trustworthy third, median third and most trustworthy third of faces calculated in the second model described in Methods. The y-axis represents mean (across subjects) percentage signal change relative to whole brain mean over scanning session for each event type. There is no clear pattern of response to the faces according to trustworthiness. Note that statistical inference is drawn only from the parametric model described in Methods, and not from the illustrative model in (**b**).



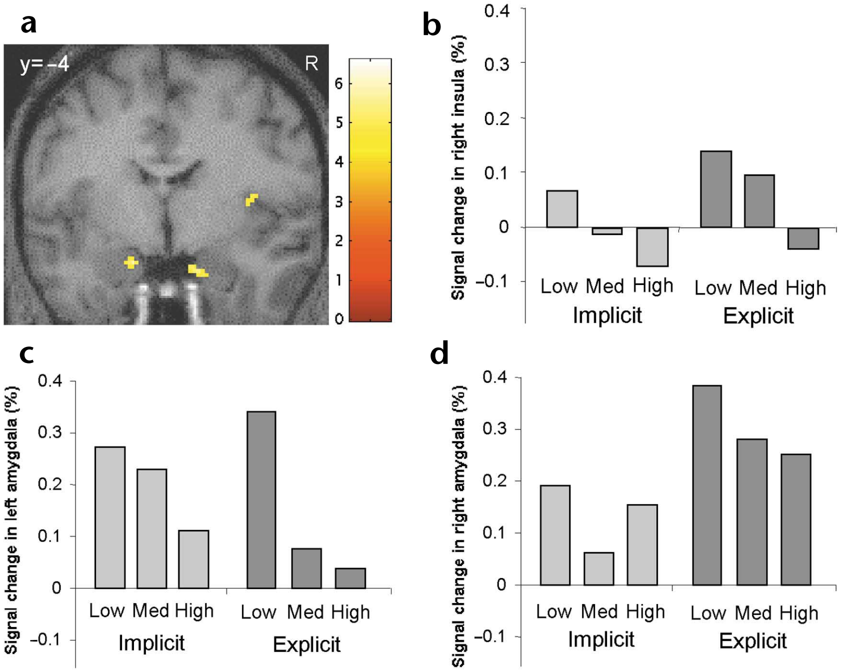


Fig. 3. Main effect of trustworthiness in amygdala and insula. (**a**) Significant increases in BOLD signal to untrustworthy faces in the right and left amygdalae and right insula (right amygdala, –18, 0, –24; Z = 4.29; p < 0.01 corrected; left amygdala, –16, –4, –20; Z = 3.92; p < 0.025 corrected; right insula, 42, –4, 12; Z = 3.48; p < 0.001 uncorrected). (**b**–**d**) Responses to faces as a function of degree of individually rated trustworthiness for right insula (**b**), left amygdala (**c**) and right amygdala (**d**). Note greater responses to less trustworthy faces across all these regions. The y-axis is as in Fig. 2.

Further areas showing increased response to untrustworthy faces included left superior temporal sulcus (–50,–58,10; Z = 4.15) and a region of the right superior middle insula (42, –4, 12; Z = 3.48; **Fig. 3a** and **b**). Additionally, bilateral activation in the fusiform gyrus was evident in this contrast (right, 44, –46, –24, Z = 3.58; left, –48, –48, –24; Z = 3.60; both p < 0.05, corrected for multiple comparisons across a small volume of interest; **Fig.** 4). Table 2 presents regions highlighted by this contrast as well as regions highlighted as more responsive to faces rated as trustworthy.

To ensure that the main effect of untrustworthiness was not driven by a highly significant activation in just one of the tasks alone, a masked conjunction of simple effects of trustworthiness under implicit and explicit task conditions was carried out (Methods). This analysis confirmed that bilateral amygdala, fusiform gyrus and right insula showed significant responses to untrustworthy faces independent of task. Notably, left STS activation was not observed in this contrast, and a *post-hoc* test revealed that the effects in this region were driven principally by trustworthiness judgments under the explicit task.

The contrast pertaining to the interaction of task and trustworthiness demonstrated an area in the lateral orbitofrontal cortex (–28, 42, 10; Z = 3.73, p < 0.0001, uncorrected) responsive to untrustworthy faces in the implicit task and to trustworthy faces in the explicit task. However, this activation failed to survive correction for multiple comparisons across the entire volume of orbitofrontal cortex. No other areas about which we had a prior hypothesis were revealed in this contrast or in the reverse interaction term.

Using an additional model that partialed out effects from facial expression of basic emotions in the stimulus set, we performed a random effects analysis across the 14 subjects. Even under these stringent criteria, right amygdala activation was still evident in this model at both uncorrected (p < 0.001) and small-volume corrected (p < 0.05 corrected for multiple comparisons across bilateral amygdala volume) thresholds (**Fig.** 5). This activation (peak at 22, 2, –18; Z = 4.06) overlapped with that reported in our primary model. At lower thresholds (p < 0.005, uncorrected), there was additional activation in left amygdala.

Discussion

The question addressed in this study was whether the dimension of trustworthiness in faces and the process of making social judgments are associated with distinct patterns of brain activation. By implication, the study is an explicit test of a proposed neurobiological model[11]. The principal findings of activation in amygdala, orbitofrontal cortex and STS are highly consistent with this model. We also extend previous lesion data[8] by showing amygdala activity in response to untrustworthy faces regardless of whether subjects were explicitly making trustworthiness judgments. This finding echoes earlier studies of obligatory threat-related processing in the amygdala[18–21]. In contrast to many imaging studies of facial emotion (for example, ref. 5), the amyg-

Table 1. Cerebral foci of activation in main effect of task.

	Coordinates of peak activation (mm)			Z score
Brain region	x	y	z	
Explicit versus implicit				
Primary visual cortex	2	–98	8	4.49
Right posterior STS*	56	–44	4	4.27
Right superior frontal gyrus	10	14	70	3.99
Left premotor cortex	–48	–2	26	3.84
Left extrastriate cortex	–24	–76	32	3.75
Right cuneus	12	–40	56	3.62
Left primary sensory cortex	–30	–28	72	3.62
Supramarginal gyrus	–62	–40	34	3.50
Right anterior insula	48	34	–6	3.37
Left superior frontal sulcus	–42	12	40	3.34
Left pre-SMA	–6	10	54	3.34
Implicit versus explicit				
Left fusiform gyrus	–36	–36	–18	3.59
Right cuneus	4	–64	12	3.49

All values, p < 0.001 uncorrected. *p < 0.05 corrected for multiple comparisons across a small volume of interest.

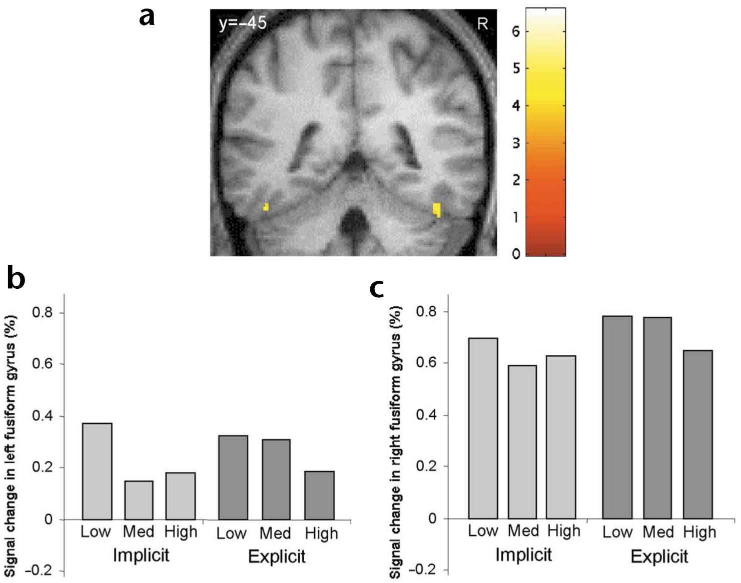


Fig. 4. Main effect of trustworthiness in fusiform gyrus. (**a**) Significant increases in BOLD signal to untrustworthy faces in the fusiform gyrus bilaterally (right, 44, –46, –22; $Z = 3.58$; $p < 0.05$, small volume corrected; left, –48, –48, –24; $Z = 3.60$; $p < 0.05$, small volume corrected). This activation is independent of task in both the left (**b**) and right (**c**) fusiform gyrus. The *y*-axis is as in Fig. 2.

dala response to untrustworthy faces was bilateral, supporting neuropsychological evidence that patients with unilateral amygdala lesions can successfully make trustworthiness judgments[8]. To our knowledge, no previous study has demonstrated an automaticity of amygdala response during complex social judgments.

In addition to the amygdala, the right insula was also activated by faces that subjects considered untrustworthy regardless of task. The insula is activated in a wide variety of functional imaging studies of emotion (for example, refs. 22–25). One suggested role for the insula is the mapping of autonomic changes as they affect the body where such mappings form the basis of 'gut feelings' about emotive stimuli[26,27]. Thus, a possible explanation for the insula activation that we observed is that a consequence of amygdala activation is the generation of autonomically mediated changes in bodily states, which are then re-mapped to the insula.

Differential activation in face-responsive regions of the fusiform gyrus was observed in relation to trustworthiness in face stimuli. Increased activity is found in modality-specific cortical areas in response to stimuli with emotional content relative to non-emotional stimuli (for example, refs. 20, 21, 28–31). Enhanced extrastriate activation in response to emotional stimuli has been attributed to modulatory influences from the amygdala[32], possibly mediated by anatomical back-projections[33]. Indeed, a human lesion study highlights a possible role for the amygdala in enhancing perceptual processing of threat stimuli[34]. We suggest that such processes extend to faces representing potential threat at the social level and that a neural consequence is enhanced fusiform activation.

The right STS showed task-related activation in the explicit judgment condition but no differential activity according to trustworthiness. In other words, the right STS was activated when subjects made explicit judgments about trustworthiness. In this regard, the STS showed activity when subjects were required to make inferences concerning the likely intentionality of others. This region has been implicated in functional imaging studies on biological motion[35] and biological-like motion[36]. More critically, activity in posterior STS and adjacent regions at the temporo-parietal junction is observed when subjects make theory of mind inferences[37–39]. This region is suggested to be involved in intention detection[40,41], rather than biological motion processing *per se*. Intention detection is a critical component in determining whether or not to trust an individual, which may explain the activity in this region in our study.

Evidence from human patients with discrete lesions of orbitofrontal cortex indicate that this region is critical for complex social judgment[9,10]. Unlike the amygdala, this region showed task-dependent activation. When subjects made explicit judgments of trustworthiness, this region responded more strongly to faces deemed trustworthy. By contrast, when judging age, this region showed greater responses to untrustworthy individuals. Other studies have reported similar task-dependent responses in lateral orbitofrontal cortex. For example, responses in a region of lateral

Table 2. Cerebral foci of activation in differential effects of trustworthiness of faces.

	Coordinates of peak activation (mm)			Z score
Brain region	*x*	*y*	*z*	
Untrustworthy versus trustworthy				
Right amygdala*	18	0	–24	4.29
Left superior temporal sulcus	–50	–58	10	4.15
Right intraparietal sulcus	22	–54	48	4.00
Left extrastriate cortex	–34	–90	24	3.94
Left amygdala*	–16	–4	–20	3.92
Right pre-SMA	8	8	62	3.83
Left parahippocampal gyrus	–18	–30	–18	3.81
Right auditory cortex	66	–18	4	3.75
Left inferior temporal gyrus	–60	–14	–30	3.64
Left fusiform gyrus*	–48	–48	–24	3.60
Right fusiform gyrus*	44	–46	–22	3.58
Thalamus	4	–12	14	3.52
Right insula	42	–4	12	3.48
Left superior temporal gyrus	–58	–32	14	3.42
Trustworthy versus untrustworthy				
Left insula	–36	4	–4	3.65
Right dorsolateral prefrontal/ frontopolar cortex	34	52	6	3.23

All values, $p < 0.001$ uncorrected. $^*p < 0.05$ corrected for multiple comparisons across a small volume of interest.

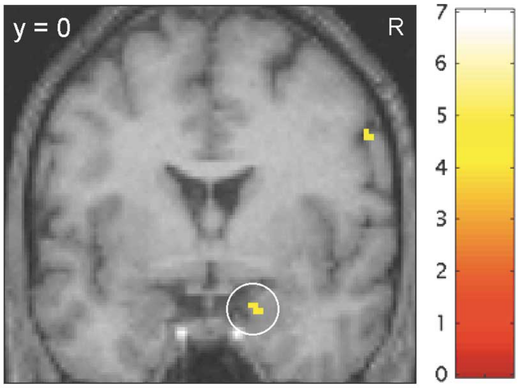


Fig. 5. Main effect of trustworthiness in amygdala independent of facial emotion. Significant increases in BOLD signal in response to untrustworthy faces in right amygdala even when scores for four basic facial emotions are additionally used as parametric covariates in the analysis. This activation is significant at $p < 0.05$, corrected for multiple comparisons across the volume of bilateral amygdala. Activation peak at 18, 2, –22 ($Z = 4.06$), but overlaps with right amygdala activation focus shown in Fig. 2. At lower threshold of $p < 0.005$ uncorrected, activation is evident in left amygdala.

orbitofrontal cortex vary between preference and recognition judgment tasks on the same stimuli[42]. A dissociation between implicit/automatic social judgment and explicit (laboratory-tested) social judgment has also been reported in a patient with orbitofrontal cortex damage[9]. This patient remained able to evaluate social situations under explicit task instructions but was impaired in day-to-day ("automatic"[9]) social judgments. Note that activation in this region did not survive correction for multiple comparisons in our study, and we emphasize effects in this region based on its known involvement in social judgments[9,10].

Several regions of interest in this study (amygdala, orbitofrontal cortex and insula) are activated in processing specific facial expressions. Facial expressions of fear consistently activate the amygdala[5,6,43], whereas facial expressions of disgust activate the anterior insula[6,7]. Additionally, we demonstrate correlations between the trustworthiness scores and scores for facial emotions attributed to our stimulus set (**Fig. 1b–e**). Consequently, one possibility is that differential patterns of activation seen in this study reflect influences from one or more emotional expressions alone. We assessed this possibility by analyzing the fMRI data with additional nuisance covariates pertaining to the degree of emotional expression of each of four basic emotions (anger, fear, happiness, sadness). A significant right amygdala response to untrustworthy faces persisted, even after this secondary analysis accounted for the variance attributed to facial emotion. These results suggest that facial expressions of emotion provide a constituent element in making trustworthiness judgments but that amygdala responses also were independent of these effects. Notably, patients with bilateral amygdala lesions show deficits in making social judgments in the context of maintained ability to use information about emotional expression of face stimuli[8].

It is interesting to speculate how the results of this study might generalize to social judgments about stimuli in other modalities. Patients with bilateral amygdala lesions are able to make accurate trustworthiness judgments based on verbal reports[8]. It is plausible therefore that amygdala involvement in trustworthiness decisions may be modality-specific. This hypothesis could be tested in follow-up experiments involving trustworthiness judgments about vocal stimuli, or scenarios about individuals based on written descriptions. Our prediction would be that superior temporal sulcus activation would remain in these other contexts and that fusiform modulation would be substituted by modality-specific cortical responses, for example, auditory cortex in the case of vocal stimuli.

In conclusion, we present functional brain imaging evidence for a neural substrate of social cognition that conforms to a previously proposed neurobiological model[11]. Our data extends this model by highlighting a dissociation between automatic and intentional engagement within this proposed circuitry. Thus, social judgments about faces reflect a combination of brain responses that are stimulus driven, in the case of the amygdala, and driven by processes relating to inferences concerning the intentionality of others, in the case of STS.

Methods

Subjects. Informed consent to partake in a study approved by the Joint National Hospital for Neurology and Neurosurgery/Institute of Neurology Ethics Committee was obtained from 16 right-handed Caucasian volunteers (8 male, 8 female; age range 18–30 years; mean age 23.3 years). Two subjects (both females) were excluded from the analysis; one revealed psychiatric history after scanning and another provided extreme trustworthiness ratings. (Spearman's rho of correlation of ratings with mean of all other subjects, –0.445; for all remaining subjects, Spearman's rho values were over 0.3.) All remaining subjects were free from psychiatric or neurological history. All subjects except one had completed more than two years of post-16 education, and mean length of post-16 formal education was 4.8 years.

Stimuli. Grayscale frontal images of 120 Caucasian male faces were selected from a larger selection of images following a pilot study outside the scanner. The images were selected to cover a range of trustworthiness scores rated by the subjects in the pilot study ($n = 30$; 13 females, 17 males, ages 17–32, mean age 23.5), but to score as low as possible on ratings of 'happiness' and 'anger'. Gaze direction of all stimuli was directly forward. Stimuli were adjusted to be of approximately equal size and luminance and manipulated such that each face was centered on a gray background in a 400 × 400 pixel image. Of the 120 stimuli used in the imaging study, 60 were high school student photographs and 60 photographs of university students. There was no significant difference in average trustworthiness score between the two groups (Mann-Whitney U test, $p > 0.90$).

Psychological task. The scanning session for each participant was divided into two parts. In one half of the session, 60 faces were presented sequentially, and participants made a judgment, indicating with a push-button response, whether the face was a high school or university student. In the other half of the session, they judged whether the face was trustworthy or untrustworthy. The order of tasks was counterbalanced between participants. At the start of each task, a word appeared on screen informing the subject of the task requirement ("School/Uni" or "Trustworthiness").

Stimuli were presented on a gray background once each in random order, randomly interspersed with 60 null events. Each stimulus was presented for 1 s with an inter-trial interval of 2 s. Between faces, a fixation cross was presented. Null events were of 3 s duration, during which time a fixation cross remained on screen. Stimuli subtended visual angles of approximately 10° vertically and 5° horizontally.

Image acquisition. Subjects were scanned during task performance using a Siemens VISION system (Erlangen, Germany) at 2 Tesla to acquire gradient-echo, echoplanar T2*-weighted images with BOLD (blood oxygenation level dependent) contrast. Each volume comprised 33 × 2.2 mm axial scans with 3-mm in-plane resolution, and volumes were continuously acquired every 2.5 s. Subjects were placed in light head restraint within the scanner to limit head movement during acquisition. Each run began with 5 'dummy' volumes (subsequently discarded) to allow for T1 equilibration effects. Additionally, a T1-weighted structural image was acquired in each subject.

All functional volumes were realigned[44] and slice timing corrected (R. Henson *et al.*, *Neuroimage* **9**, 125, 1999), normalized into a standard space[45] to allow group analysis, and smoothed with an 8-mm FWHM Gaussian kernel to account for residual intersubject differences.

Debriefing. After scanning, participants undertook a self-paced task in which they rated all the faces on a scale of trustworthiness from 1 (highly untrustworthy) to 7 (highly trustworthy). When all 120 faces had been rated, a second task was performed, in which participants named emotions that they perceived in the faces by means of a seven-way forced choice procedure (neutral, happy, sad, angry, disgust, fear, surprise). To assist subjects with this task, we gave them a printed sheet with photographs of one face from the Ekman and Friesen series[46] expressing each of these seven emotions.

Emotion ratings for stimuli. An additional set of 16 subjects (10 males, 6 females; age range 19–34 years; mean age 23.7 years) undertook a task in which they rated the degree of emotional expression within each face on each of four basic emotions (anger, fear, happiness, sadness) in turn. Ratings were from 1 (neutral for this particular emotion) to 7 (highest degree of this particular emotion).

Data analysis. Imaging data were analysed with SPM99 using an event-related model[47]. The experimental design allowed a parametric factorial analysis whereby trustworthiness was a parametric regressor and the task (age or trustworthiness judgment) the second factor.

The presentation of each face was modeled by convolving a delta function at each event onset with a canonical hemodynamic response function (HRF) and its temporal derivative to create regressors of interest. These regressors were then parametrically modulated to model subject-specific trustworthiness judgments: that is, the height of the HRF for stimuli was modulated as a function of the trustworthiness score assigned to that stimulus by the subject. Subject-specific parameter estimates pertaining to each regressor were calculated for each voxel[48]. Contrast images were calculated by applying appropriate linear contrasts to the parameter estimates for the parametric regressor of each event. These contrast images were then entered into a one-sample *t*-test across the 14 subjects (that is, a random effects analysis). In regions about which we had a prior hypothesis, we applied a correction for multiple comparisons across a small volume of interest to the *p*-values in this region[49]. We report predicted regions surviving this correction at $p < 0.05$. Volumes of interest for amygdala, orbitofrontal cortex and STS were defined by drawing a mask around the regions bilaterally on a normalized T1 structural image with reference to an atlas of human neuroanatomy[50] using the software package MRIcro (http://www.psychology.nottingham.ac.uk/staff/cr1/mricro.html). Total volume of the amygdala mask was approximately 10 cm^3, volume of the orbitofrontal mask approximately 50 cm^3, and volume of the STS mask approximately 20 cm^3. In the case of the fusiform gyri, small-volume correction was based upon a sphere of 10 mm radius centered on coordinates derived from a previous study[21]. We report descriptively activations outside regions of interest surviving a threshold of $p < 0.001$ uncorrected with an extent threshold of 5 contiguous voxels.

To ensure that the main effect of trustworthiness did not arise from a highly significant activation in just one of the simple effects (i.e., that activation was task independent), we created a mask from random-effects SPMs for the simple effect of untrustworthiness under both tasks (each thresholded at $p < 0.05$, uncorrected). This was used to mask the main effect of trustworthiness. Activations surviving this masking procedure reflect responses during both implicit and explicit judgments.

For the purposes of illustration, a second model was constructed by dividing the events for each subject into three groups by rank score for individual stimuli (that is, the least trustworthy third of faces as one event type, the median third as a second, and the most trustworthy third as a third). This model is used in Figs. 1–3 to demonstrate the direction of BOLD signal change with respect to trustworthiness score. Note that statistical inferences are drawn solely from the parametric model described above.

The mean ratings of facial emotion derived from a second set of 16 age-matched subjects (see above) were used to construct another model for the data. In this model, subject-specific ratings for trustworthiness were entered as parametric covariates, as before. Additionally, mean ratings for each of the four emotions (anger, fear, happiness, sadness) were entered as nuisance parametric covariates. The parameter estimates for trustworthiness are therefore rendered independent of the effects of the four facial expressions, and variance better explained by the effects of a given facial expression will be attributed to the regressor modeling that facial expression. Contrast images for trustworthiness derived from this model were then entered into a random-effects analysis.

Acknowledgements

This work was supported by a programme grant to R.J.D. from the Wellcome Trust.

Competing interest statement

The authors declare that they have no competing financial interests.

RECEIVED 5 DECEMBER; ACCEPTED 26 DECEMBER 2001

1. Humphrey, N. *Consciousness Regained: Chapters in the Development of Mind* (Oxford Univ. Press, New York, 1983).
2. Adolphs, R., Tranel, D., Damasio, H. & Damasio, A. Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature* **372**, 669–672 (1994).
3. Hornak, J., Rolls, E. T. & Wade, D. Face and voice expression identification in patients with emotional and behavioural changes following ventral frontal lobe damage. *Neuropsychologia* **34**, 247–261 (1996).
4. Adolphs, R., Damasio, H., Tranel, D., Cooper, G. & Damasio, A. R. A role for somatosensory cortices in the visual recognition of emotion as revealed by three-dimensional lesion mapping. *J. Neurosci.* **20**, 2683–2690 (2000).
5. Morris, J. S. *et al.* A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature* **383**, 812–815 (1996).
6. Phillips, M. L. *et al.* A specific neural substrate for perceiving facial expressions of disgust. *Nature* **389**, 495–498 (1997).
7. Sprengelmeyer, R., Rausch, M., Eysel, U. T. & Przuntek, H. Neural structures associated with recognition of facial expressions of basic emotions. *Proc. R. Soc. Lond. B Biol. Sci.* **265**, 1927–1931 (1998).
8. Adolphs, R., Tranel, D. & Damasio, A. R. The human amygdala in social judgment. *Nature* **393**, 470–474 (1998).
9. Eslinger, P. J. & Damasio, A. R. Severe disturbance of higher cognition after bilateral frontal lobe ablation: patient EVR. *Neurology* **35**, 1731–1741 (1985).
10. Damasio, A. R. *Descartes' Error: Emotion, Reason and the Human Brain* (Putnam, New York, 1994).
11. Brothers, L. The social brain: a project for integrating primate behavior and neurophysiology in a new domain. *Concepts Neurosci.* **1**, 27–51 (1990).
12. Adolphs, R. Social cognition and the human brain. *Trends. Cogn. Sci.* **3**, 469–479 (1999).
13. Adolphs, R. The neurobiology of social cognition. *Curr. Opin. Neurobiol.* **11**, 231–239 (2001).
14. Allison, T., Puce, A. & McCarthy, G. Social perception from visual cues: role of the STS region. *Trends Cogn. Sci.* **4**, 267–278 (2000).
15. Dolan, R. J. *et al.* Neural activation during covert processing of positive emotional facial expressions. *Neuroimage* **4**, 194–200 (1996).
16. Blair, R. J., Morris, J. S., Frith, C. D., Perrett, D. I. & Dolan, R. J. Dissociable neural responses to facial expressions of sadness and anger. *Brain* **122**, 883–893 (1999).
17. Lane, R. D., Chau, P. M. & Dolan, R. J. Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures. *Neuropsychologia* **37**, 989–997 (1999).
18. Morris, J. S., Ohman, A. & Dolan, R. J. Conscious and unconscious emotional learning in the human amygdala. *Nature* **393**, 467–470 (1998).
19. Whalen, P. J. *et al.* Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *J. Neurosci.* **18**, 411–418 (1998).
20. Strange, B. A., Henson, R. N., Friston, K. J. & Dolan, R. J. Brain mechanisms for detecting perceptual, semantic, and emotional deviance. *Neuroimage* **12**, 425–433 (2000).
21. Vuilleumier, P., Armony, J. L., Driver, J. & Dolan, R. J. Effects of attention and emotion on face processing in the human brain. An event-related fMRI study. *Neuron* **30**, 829–841 (2001).
22. Buechel, C., Morris, J., Dolan, R. J. & Friston, K. J. Brain systems mediating aversive conditioning: an event-related fMRI study. *Neuron* **20**, 947–957 (1998).
23. Buechel, C., Dolan, R. J., Armony, J. L. & Friston, K. J. Amygdala-hippocampal involvement in human aversive trace conditioning revealed through event-related functional magnetic resonance imaging. *J. Neurosci.* **19**, 10869–10876 (1999).
24. Casey, K. L. Forebrain mechanisms of nociception and pain: analysis through imaging. *Proc. Natl. Acad. Sci. USA* **96**, 7668–7674 (1999).
25. Critchley, H. D., Mathias, C. J. & Dolan, R. J. Neural activity in the human brain relating to uncertainty and arousal during anticipation. *Neuron* **29**, 537–545 (2001).

26. Damasio, A. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness* (Harcourt Brace, New York, 1999).
27. Critchley, H. D., Mathias, C. J. & Dolan, R. J. Neuroanatomical basis for first- and second-order representations of bodily states. *Nat. Neurosci.* **4**, 207–212 (2001).
28. Breiter, H. C. *et al.* Response and habituation of the human amygdala during visual processing of facial expression. *Neuron* **17**, 875–887 (1996).
29. Isenberg, N. *et al.* Linguistic threat activates the human amygdala. *Proc. Natl. Acad. Sci. USA* **96**, 10456–10459 (1999).
30. Morris, J. S., Buchel, C. & Dolan, R. J. Parallel neural responses in amygdala subregions and sensory cortex during implicit fear conditioning. *Neuroimage* **13**, 1044–1052 (2001).
31. Dolan, R. J., Morris, J. S. & de Gelder, B. Crossmodal binding of fear in voice and face. *Proc. Natl. Acad. Sci. USA* **98**, 10006–10010 (2001).
32. Morris, J. S. *et al.* A neuromodulatory role for the human amygdala in processing emotional facial expressions. *Brain* **121**, 47–57 (1998).
33. Amaral, D. G. & Price, J. L. Amygdalo-cortical projections in the monkey (*Macaca fascicularis*). *J. Comp. Neurol.* **230**, 465–496 (1984).
34. Anderson, A. K. & Phelps, E. A. Lesions of the human amygdala impair enhanced perception of emotionally salient events. *Nature* **411**, 305–309 (2001).
35. Bonda, E., Petrides, M., Ostry, D. & Evans, A. Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *J. Neurosci.* **16**, 3737–3744 (1996).
36. Castelli, F., Happe, F., Frith, U. & Frith, C. Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage* **12**, 314–325 (2000).
37. Fletcher, P. C. *et al.* Other minds in the brain: a functional imaging study of 'theory of mind' in story comprehension. *Cognition* **57**, 109–128 (1995).
38. Gallagher, H. L. *et al.* Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia* **38**, 11–21 (2000).
39. Brunet, E., Sarfati, Y., Hardy-Bayle, M. C. & Decety, J. A PET investigation of the attribution of intentions with a nonverbal task. *Neuroimage* **11**, 157–166 (2000).
40. Frith, C. D. & Frith, U. Interacting minds—a biological basis. *Science* **286**, 1692–1695 (1999).
41. Frith, C. & Frith, U. in *Understanding Other Minds: Perspectives From Developmental Cognitive Neuroscience* (eds. Baron-Cohen, S., Tager-Flusberg, H. & Cohen, D. J.) 334–356 (Oxford Univ. Press, New York, 2000).
42. Elliott, R. & Dolan, R. J. Neural response during preference and memory judgments for subliminally presented stimuli: a functional neuroimaging study. *J. Neurosci.* **18**, 4697–4704 (1998).
43. Phillips, M. L. *et al.* Neural responses to facial and vocal expressions of fear and disgust. *Proc. R. Soc. Lond. B Biol. Sci.* **265**, 1809–1817 (1998).
44. Friston, K. *et al.* Spatial registration and normalization of images. *Hum. Brain Mapp.* **2**, 165–189 (1995).
45. Talairach, J. & Tournoux, P. *Co-planar Stereotaxic Atlas of the Human Brain* (Theime, Stuttgart, Germany, 1988).
46. Ekman, P. & Friesen, W. V. *Pictures of Facial Affect* (Consulting Psychologists Press, Palo Alto, California, 1975).
47. Josephs, O., Turner, R. & Friston, K. Event-related fMRI. *Hum. Brain Mapp.* **5**, 243–248 (1997).
48. Friston, K. *et al.* Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* **2**, 189–210 (1995).
49. Worsley, K. *et al.* A unified statistical approach for determining significant signals in images of cerebral activation. *Hum. Brain Mapp.* **4**, 58–73 (1996).
50. Duvernoy, H. M. *The Human Brain* (Springer, Vienna, 1999).