doi:10.1093/scan/nsq086 SCAN (2011) 6, 572–581

Amygdala and dorsomedial prefrontal cortex responses to appearance-based and behavior-based person impressions

Sean G. Baron,[1] M. I. Gobbini,[2] Andrew D. Engell,[3] and Alexander Todorov[1]

[1]Department of Psychology, Princeton University, Princeton, NJ, 08540, USA, [2]Department of Psychology, University of Bologna, 40127 Bologna, Italy, and [3]Department of Psychology, Yale University, New Haven, CT, 06511, USA

We explored the neural correlates of learning about people when the affective value of both facial appearance and behavioral information is manipulated. Participants were presented with faces that were either rated as high or low on trustworthiness. Subsequently, we paired these faces with positive, negative, or no behavioral information. Prior to forming face–behavior associations, a cluster in the right amygdala responded more strongly to untrustworthy than to trustworthy faces. During learning, a cluster in the dorsomedial prefrontal cortex (dmPFC) responded more strongly to faces paired with behaviors than faces not paired with behaviors. We also observed that the activity in the dmPFC was correlated with behavioral learning performance assessed after scanning. Interestingly, individual differences in the initial amygdala response to face trustworthiness *prior* to learning modulated the relationship between dmPFC activity and learning. This finding suggests that the activity of the amygdala can affect the interaction between dmPFC activity and learning.

Keywords: faces; learning; trustworthiness; PFC; amygdala

INTRODUCTION

People are able to form person impressions after less than 40 ms exposure to a person's face (Bar *et al.*, 2006; Todorov *et al.*, 2009). These judgments can affect interpersonal interactions (Nisbett and Wilson, 1977; McCulloch *et al.*, 2008) and influence important outcomes such as political elections (Ballew and Todorov, 2007; Olivola and Todorov, 2010) or criminal sentencing decisions (Downs and Lyons, 1991; Zebrowitz and McDonald, 1991; Blair *et al.*, 2004). However, people do not live in a static social world where faces flicker in and out of existence as the sole cue driving person impressions. Instead, even brief periods of interaction allow the opportunity for learning from others' behaviors to contribute to more nuanced and accurate appraisals. Behaviors add diagnostic information that can be processed when inferring character traits (Todorov and Uleman, 2002; Uleman *et al.*, 2005; Bliss-Moreau *et al.*, 2008), and as with face-based impression formation, these behavior-based inferences can occur quickly and independent of attentional resources (Todorov and Uleman, 2003).

The goal of the present study was to use functional magnetic resonance imaging (fMRI) to explore which brain regions are involved in learning about people when the affective value of both facial appearance and behavioral descriptions is manipulated. We presented participants with emotionally neutral faces that were rated as highly trustworthy or untrustworthy. These faces were then paired with positive, negative, or no behavioral descriptions. We manipulated the trustworthiness of faces for two reasons. First, we wanted to use a trait that would best extend to generalized face evaluation. Oosterhof and Todorov (2008) showed that face valence accounts for most of the variance in a variety of trait judgments, and that trustworthiness judgments can reliably act as a proxy for valence evaluation. Second, there is a history of research describing the neural correlates of face-based impression formation as it relates to trustworthiness. The amygdala has been consistently identified as one of the regions involved in trustworthiness evaluation (Adolphs *et al.*, 1998; Winston *et al.*, 2002; Engell *et al.*, 2007) and face–valence evaluation more broadly (Todorov and Engell, 2008). Specifically, the amygdala response increases as the perceived trustworthiness of faces decreases (Winston *et al.*, 2002; Engell *et al.*, 2007; Todorov *et al.*, 2008; but see Said *et al.*, 2009 for non-linear responses). To replicate these findings, in the first stage of the experiment, participants were presented with trustworthy- and untrustworthy-looking faces unaccompanied by behavioral information. Consistent with the prior studies, we expected that untrustworthy looking faces would evoke a stronger response in the amygdala than trustworthy looking faces.

However, a critical question for this study was whether the amygdala response would change as a function of behavior learning. In the second stage of the experiment, participants

Received 3 March 2010; Accepted 8 September 2010
Advance Access publication 28 October 2010

This research was supported by National Science Foundation grants 0446846 and 0823749. The authors thank Jean-Baptiste Pochon and Christopher P. Said for their comments and ideas with respect to this work. Funded by NSF grants 0446846 & BCS 0823749.

Correspondence should be addressed to Sean G. Baron, Department of Psychology, Princeton University, Princeton, NJ 08540, USA. E-mail: sbaron@princeton.edu

were presented with the faces from the first stage of the experiment but the faces were paired with positive, negative, or no behavioral descriptions. Previous studies have shown that behavior or trait information has a large effect on person judgments (Todorov and Olson, 2008; Rudoy and Paller, 2009). For example, in a design similar to the current study, Todorov and Olson (2008) presented participants with trustworthy- and untrustworthy-looking faces paired with positive or negative behaviors. Faces associated with positive behaviors were judged more positively than faces associated with negative behaviors independent of the perceived trustworthiness of the face. Interestingly, whereas this behavior learning effect was detectable in a patient with a lesion in the hippocampus, it was not detectable in a patient, whose lesion extended into the amygdala and the temporal pole, suggesting that the latter regions are important for forming affective associations with faces.

In addition to investigating whether the amygdala responses would be modulated by behavioral information, we also sought to detect brain regions that might underlie the learning of this type of information. One region that appears to play a role in person learning is the dorsomedial prefrontal cortex (dmPFC). Mitchell and colleagues (2004, 2005, 2006), have studied the effect of behavioral information on the formation of person impressions. In these studies, participants were presented with faces and social information relevant to the faces, and asked to form person impressions. Across the studies, the dmPFC emerged as the region most reliably responsive to the formation of impressions about others. Interestingly, dmPFC did not appear to be responsive to the formation of impressions related to inanimate objects (Mitchell *et al.*, 2005). This region has also been implicated in the spontaneous retrieval of person knowledge upon presentation of familiar faces (Gobbini *et al.*, 2004; Gobbini and Haxby, 2007; Todorov *et al.*, 2007). At the same time, this region does not respond to faces that are only visually familiar (Gobbini and Haxby, 2006). Given these findings, the dmPFC seems particularly well suited to mediate the learning of person specific information. Consequently, we expected to observe that the dmPFC would respond preferentially to faces presented with behavioral information, and that its activity would be correlated with learning of face–behavior associations.

We were also interested in exploring the relationship between the initial amygdala response to the perceived trustworthiness of faces and the dmPFC response during the learning of face–behavior associations. Specifically, we tested whether the strength of the amygdala response during the initial presentation of faces—when they were not associated with behaviors—would moderate the relationship between the dmPFC activity during learning and the learning observed in the participants' judgments. This hypothesis was based on the findings of a recent behavioral and event-related potential (ERP) study (Rudoy and Paller, 2009). In this study, participants were presented with trustworthy- and untrustworthy-looking faces paired with positive or negative trait words and asked to learn the face–trait associations. Rudoy and Paller found that the effect of face trustworthiness emerged before the effect of learning. More relevant to the current hypothesis, they also found that participants who were strongly influenced by face appearance were less likely to be influenced by learned trait associations. This result, along with prior evidence for functional interactions between the dmPFC and the amygdala in social face judgments (Kim *et al.*, 2004), led us to hypothesize that participants who have a strong differential amygdala response to trustworthy and untrustworthy faces would show weaker learning as a result of a modulation of dmPFC activity.

Finally, we attempted to identify brain regions whose responses to face trustworthiness changed over the course of the experiment as a function of the associated behaviors. Specifically, we looked for regions showing a significant interaction of face trustworthiness, valence of behavioral description, and time (pre-learning of behavioral associations *vs* post-learning of behavioral associations). The demarcation of such regions could provide clues to the neural network(s) involved in the updating of person impressions.

METHODS

Participants

Twenty-four (nine female) volunteers participated in the study. They were between the ages of 19 and 29 (mean = 23). All participants were right-handed, had normal or corrected-to-normal vision, and had no prior history of neurological or psychiatric disease. After providing informed consent, participants took part in both an imaging and behavioral study. All participants were fully debriefed at the completion of the study in accordance with Princeton University's Institutional Review Panel guidelines.

Face stimuli

We used as stimuli a set of 48 face images from the Karolinska Directed Emotional Faces set (Lundqvist *et al.*, 1998). This set of standardized face images consists of male and female amateur actors between 20 and 30 years of age. Each actor wore a gray T-shirt and had no facial hair, jewelry, eyeglasses, or visible make-up. Only frontal headshot images of individuals exhibiting a neutral expression and direct eye-gaze were used. All faces had been previously rated on trustworthiness (Engell *et al.*, 2007). From these ratings, six subsets of eight face images (four highly trustworthy and four highly untrustworthy) were created. Each of these image subsets had an equal number of images of men and women. Within each subset, the mean trustworthiness of trustworthy faces was significantly greater than that of untrustworthy faces.

These subsets were then paired with behavioral descriptions to create six experimental conditions. These conditions included: trustworthy and untrustworthy faces paired with

positive behavioral descriptions (TP and UP), trustworthy and untrustworthy faces paired with negative behavioral descriptions (TN and UN), and trustworthy and untrustworthy faces paired with no behavioral descriptions (TX and UX). To control for potential effects of unique face–behavior pairs, three counter-balanced versions of the study were created. Thus, every face was paired with every type of behavior across participants.

Behavioral descriptions

We chose the most extreme positive (e.g. 'She adopted a homeless child.') and negative (e.g. 'He stole money from the priest.') behaviors (16 of each) from Todorov *et al.* (2007). Each sentence was edited so that their reading length was approximately the same. Proper nouns (names) were replaced by face appropriate gender-specific pronouns.

All stimuli were projected onto a screen located at the rear of the bore of the magnet. Participants were able to view these stimuli via an angled mirror attached to the RF coil placed above their eyes.

FMRI task procedures

Participants were told that they were taking part in a face memory experiment. The experiment consisted of three stages. In the first, pre-learning, stage (the first out of five acquisition time series), participants were presented with the faces only. In the second, learning, stage (the second through fourth time series), participants were shown faces paired with behaviors. In the third, post-learning, stage (the fifth time series), participants were presented with faces only as in the first stage (Figure 1).

During each time series, participants completed the same task; however, series differed in length and stimuli presentation. Series one (pre-learning) and five (post-learning) consisted of 21-sequences of stimuli. Each sequence was composed of an 'observational' period and a 'test' period.

Fig. 1 Basic fMRI paradigm design. (**A** and **C**) In the pre- and post-learning periods participants encountered multiple stimuli sequences consisting of faces. (**B**) During the learning period participants saw multiple stimuli sequences consisting of behavioral description–face pairings. (A–C) In all periods, face stimuli were followed by fixation screens, and each sequence of face stimuli or behavior–face pairings ended with a test face.

The observational period consisted of eight images. The test period immediately followed the observational period, and contained one 'test-image'. Observational period stimuli were taken either from the set of 48 faces described above, or from a set of eight scrambled face images. Across all 21 sequences, these 48 faces and eight scrambled images were presented three times each—none were presented more than once in a sequence. For each observational period between zero and two scrambled images were included. That is, in each of these observational periods, between eight and six face stimuli were presented. All stimuli within the observational periods were presented for 500 ms each followed immediately by a 2500-ms fixation screen. All test periods began with a 2000-ms presentation of a question mark immediately followed by a 500-ms presentation of a test-image. Participants were instructed to indicate whether or not they had seen the 'test-image' during the preceding observational period. Each sequence was separated by a 10-s rest period to allow hemodynamic activation to return to baseline.

Time series two through four consisted of six blocks of faces preceded by behaviors. As in the first and fifth time series, face stimuli were presented for 500 ms and at the end of each observational period participants indicated their recollection of a 'test-image'. However, before each face, either a 4000-ms fixation cross (for TX and UX) or a 4000-ms behavioral description (for TP, TN, UP, UN) was presented. Each behavior–face pair was followed by a 3500-ms fixation screen. These series also differed from series one and five in that each observation period consisted of six rather than eight face-stimuli, and that no scrambled images were ever presented. Each of the 48 face–behavior pairs was presented once per time series. Together, these three time series will subsequently be referred to as the learning period. All time series began and ended with a 16-s presentation of a fixation cross. Furthermore, each run was pseudo-randomized so that each of the six experimental conditions preceded or followed every other condition with approximately the same frequency.

Behavioral task procedures

After finishing the fMRI part of the experiment, participants completed the behavioral part of the experiment. Participants were seated in front of a computer and instructed to rate all 48 faces on their trustworthiness. Faces were presented sequentially in a randomized order. Each face remained on-screen until participants entered their response via keyboard. The response scale ranged from 1 (very untrustworthy) to 9 (very trustworthy).

Image acquisition

The measure of neural activation was blood oxygenation level-dependent (BOLD) signal. We acquired gradient echo planar images (EPI) 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). We achieved near whole brain coverage by using 33 interleaved 3-mm axial slices. Prior to EPI time series, 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 for registration of functional activity to the subject's anatomy and for spatial normalization of data across participants.

Image analysis

All fMRI data were analyzed with Analysis of Functional Neuro-images software (AFNI; Cox, 1996). Prior to analysis the first four echo planar images (EPI) of each time-series were discarded to allow MR signal to reach steady-state equilibrium. Participants' 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.

For each participant, voxel-wise multiple regression was used to generate parameter estimates. Nineteen regressors of interest (for each of the six conditions per the pre-learning, learning and post-learning periods; and one regressor for scrambled faces presented in the pre- and post-learning periods) specific to the 500-ms presentation of either face or scrambled-face stimuli were convolved with a canonical hemodynamic response function and entered into a general linear model. Motion estimates, the presentation of behavioral descriptions, and the 'test-face' sections of each time series were included as regressors of no interest. Each participant's parameter estimate maps were projected into Talairach space (Talairach and Tournoux, 1998) prior to performing any group-level analyses.

For each participant, we used the results of the regression analysis to calculate the average voxel-wise neural response to each of the six experimental conditions for each of the three learning periods of the study. These individual brain maps were then used in analyses of specific functional regions of interest (fROIs) that emerged from the contrasts described below.

To assess the effect of facial trustworthiness on the amygdala during the pre-learning period we used a *t*-test to contrast the parameter estimates of trustworthy and untrustworthy faces. Because we had an a priori prediction concerning the amygdala, the resultant parametric map was thresholded at an uncorrected voxel-wise α-level of 0.001, and any significant clusters within either amygdala were included in subsequent analyses. As described below, we found a cluster of voxels in the right amygdala. We further analyzed the responses in this region in both the learning and post-learning portions of the study.

To examine the effect of learning, we completed a *t*-test on the parameter estimates supplied by the GLM of each

participant to contrast faces with behaviors versus faces without behaviors during the learning period. This parametric map was then thresholded at an uncorrected voxel-wise α-level of 0.001. In order to determine the minimum cluster size for corrected significance of $P < 0.05$ we then completed a Monte Carlo simulation of null-hypothesis data, using the AlphaSim program in AFNI. These simulations resulted in a minimum cluster size of 504 mm^3. The same contrast and cluster-size restrictions were also completed for the post-learning portion of the study. For both the learning and post-learning portions of the study, we also computed contrasts for the effect of the valence of behavior (faces associated with positive behaviors *vs* faces associated with negative behaviors) and face trustworthiness (trustworthy *vs* untrustworthy faces). However, none of the regions observed in these contrasts survived correction for multiple comparison.

Finally, because the events in the pre- and post-learning stages of the experiment were identical and we were interested in the interaction of face trustworthiness, behavioral descriptions, and learning, we completed a whole brain analysis to uncover brain regions where a significant three-way interaction might occur. For this whole-brain analysis, we submitted the parameter estimates for each experimental condition across the relevant time periods for every participant to a 2 (learning period: pre-learning and post-learning) $\times$ 2 (behavior type: positive, negative, null) $\times$ 2 (face trustworthiness: trustworthy and untrustworthy) repeated-measures analysis of variance (ANOVA). As above, we report only regions that reached a minimum cluster size of 504 mm^3 at an uncorrected voxel-wise α-level of 0.001.

RESULTS

Behavioral results

It is important to demonstrate that behavioral learning affected person judgments. To assess behavioral learning, we completed a 2 (face trustworthiness: trustworthy and untrustworthy) $\times$ 3 (behavior type: positive, negative, null) repeated-measures ANOVA. Not surprisingly, trustworthy faces ($M = 5.63$, $SD = 1.19$) were rated as more trustworthy than untrustworthy faces [$M = 4.47$, $SD = 1.15$; $F(1,23) = 141.33$, $P < 0.05$, for the main effect of face trustworthiness]. More importantly, there was also a significant main effect of behavior type [$F(2,46) = 21.62$, $P < 0.05$]. A series of pairwise comparisons, using Bonferroni adjustments for multiple comparisons, further explains this main effect. Faces paired with positive behaviors ($M = 5.70$, $SD = 1.17$) were rated as significantly more trustworthy than faces presented without behaviors ($M = 5.03$, $SD = 1.13$, $P < 0.05$). Furthermore, faces presented without behaviors were rated as significantly more trustworthy than those paired with negative behaviors ($M = 4.42$, $SD = 1.31$, $P < 0.05$). See Figure 2.

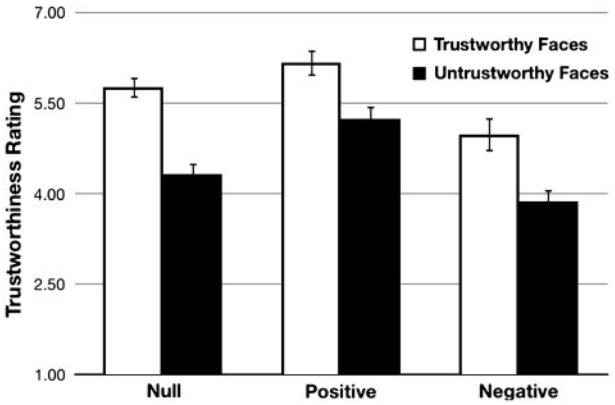

Fig. 2 Average trustworthiness ratings of faces as a function of face trustworthiness and valence of behaviors associated with the faces. There is a significant main effect of face trustworthiness (trustworthy faces > untrustworthy faces). There is also a significant main effect of behavior type [faces + positive behaviors > faces + no behaviors (null) > faces + negative behaviors]. The response scale ranged from 1 (very untrustworthy) to 9 (very trustworthy). Error bars represent standard error of the mean.

fMRI results

Amygdala response to trustworthiness and behaviors for each learning period

We were first interested in replicating results found in previous studies (Winston *et al.*, 2002; Engell *et al.*, 2007; Todorov *et al.*, 2008) showing a negative linear relationship between increasing facial trustworthiness and amygdala response. From the group-level contrast (*t*-test) for trustworthy versus untrustworthy faces from the pre-learning stage, we identified a 108 mm^3 cluster of voxels originating in the right amygdala that responded significantly more to untrustworthy than trustworthy faces (Figure 3A).[1] Surprisingly, given that at this point in the experiment none of the faces had been paired with behaviors, this cluster's response to trustworthy faces, which were to be presented without behavioral descriptions (TX) during the learning period, was less than its response to any other face–behavior combination (Figure 3B). This is also surprising because every trustworthy face was paired with every behavior type across participants. We conducted analyses to test whether this pattern was caused by one of the three counter-balancing versions of the experiment but failed to find such evidence [$F(2, 46) = 1.15$, $P = 0.35$, for the comparison of TX faces in the three versions]. One potential explanation is in terms of the order of the conditions—this order was the same across versions, although the specific faces changed within version. Regardless, in this pre-learning period all conditions showed an effect of face trustworthiness in the predicted direction (trustworthy faces eliciting significantly less activation than untrustworthy faces).

[1]No brain regions met criterion for significance after correction for multiple comparison. The reported amygdala is included because of our *a priori* hypothesis concerning the amygdala response to facial trustworthiness.

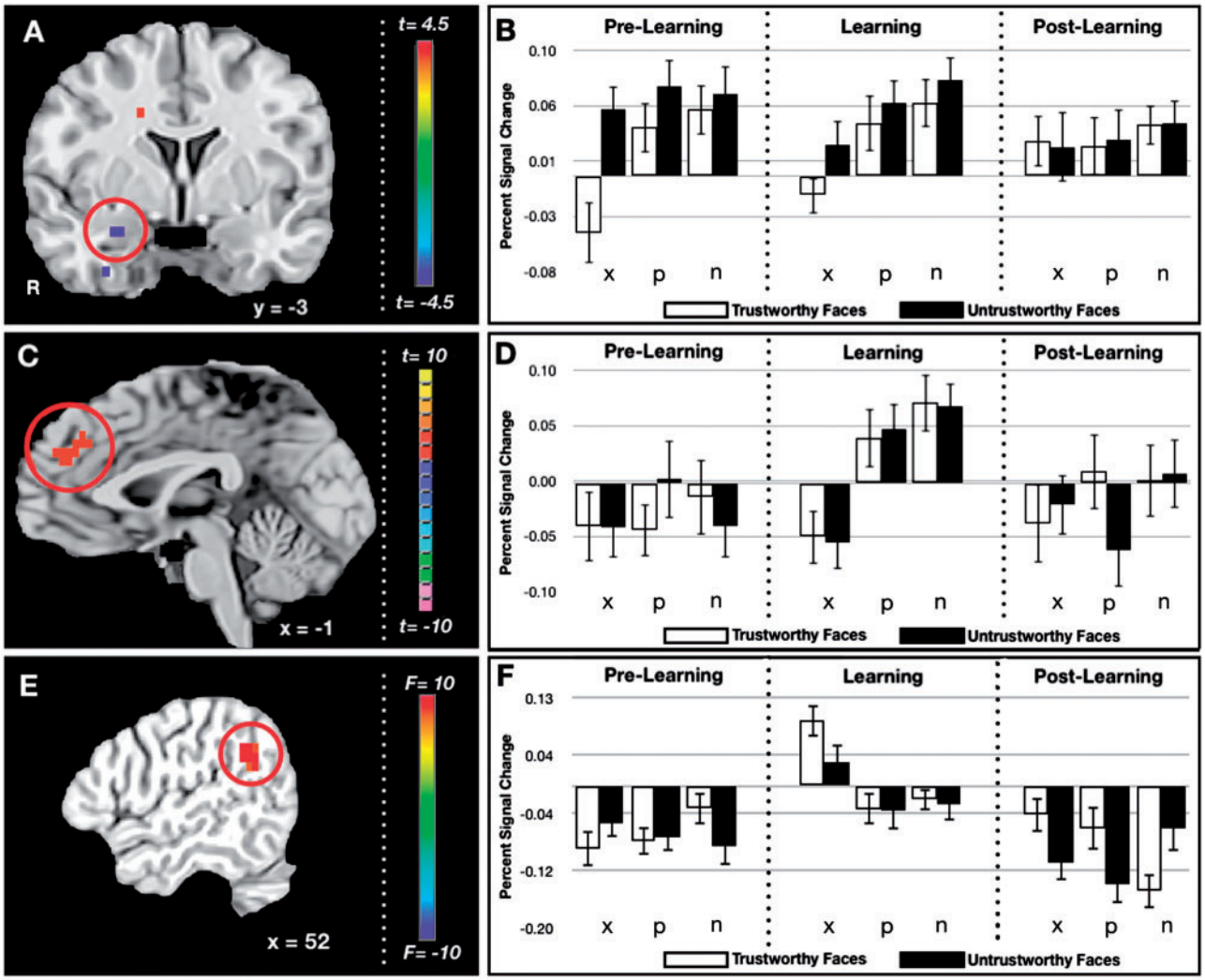

Fig. 3 (**A**) Region of right amygdala responding significantly (maximum, $t_{23} = 4.06$; $P < 0.001$, uncorrected) more for untrustworthy than trustworthy faces. The statistical maps show the results of a *t*-test performed on the coefficients of trustworthy and untrustworthy face regressors during the pre-learning period on individual data. (**B**) Response of the right amygdala during each phase of the experiment. (**C**) Region of left dorsal medial prefrontal cortex responding significantly (maximum, $t_{23} = 4.94$; $P < 0.05$, corrected for multiple comparisons) more for faces presented with behaviors than faces presented without behaviors. The statistical maps show the results of a *t*-test performed on the coefficients for regressors of faces associated with behaviors and faces without behaviors during the learning period on individual data. (**D**) Response of the left dorsal medial prefrontal cortex fROI during each phase of the experiment. (**E**) Region of right temporoparietal cortex (rTPJ) responding significantly (maximum, $F = 12.65$; $P < 0.05$, corrected for multiple comparisons) to the interaction of face trustworthiness, behavioral description, and learning period. The statistical maps show the result of a whole brain ANOVA performed on the coefficients for regressors of faces paired with each behavioral description type in the pre- and post- learning stages. (**F**) Response of the rTPJ fROI during each phase of the experiment. For panels B, D, and F: '*x*' indicates faces paired with no behavior; '*p*' indicates faces paired with positive behaviors; '*n*' indicates faces paired with negative behaviors. Error bars represent standard error of the mean.

We conducted additional analyses within this fROI for the learning stage of the experiment. Specifically, we submitted the average parameter estimate of all voxels within the amygdala fROI to a 2 (face trustworthiness: trustworthy and untrustworthy) × 3 (behavior type: positive, negative, null) repeated-measures ANOVA. During learning, the response to untrustworthy faces ($M = 0.053$, $SD = 0.049$) remained higher than the response to trustworthy faces [$M = 0.029$, $SD = 0.047$; $F(1, 23) = 4.50$, $P < 0.045$]. The analysis also revealed a significant main effect of behavior [$F(2, 46) = 4.33$, $P < 0.019$; $F < 1$ for the interaction). As shown in Figure 3B (middle plot), the response to faces associated with behaviors ($M = 0.059$, $SD = 0.059$) was stronger than the response to faces not associated with behaviors [$M = 0.005$, $SD = 0.069$; $t(23) = 2.52$, $P < 0.019$]. Although the response to faces associated with negative behaviors ($M = 0.067$, $SD = 0.072$) was, on average, greater than the response to faces associated with positive behaviors ($M = 0.050$, $SD = 0.078$), this difference was not significant ($t < 1$).

Because the events in the pre- and post-learning periods were identical, we submitted the data to 2 (learning: pre *vs* post) × 2 (face trustworthiness: trustworthy and untrustworthy) × 3 (behavior type: positive, negative, null)

repeated-measures ANOVA. We were specifically interested in whether the main effect of the face would be qualified by learning and behavior. This analysis revealed a main effect of face trustworthiness [$F(1, 23) = 11.50$, $P < 0.003$], which was qualified by an interaction with learning [$F(1, 23) = 4.81$, $P < 0.039$]. This interaction indicated that whereas the difference between trustworthy and untrustworthy faces was significant in the pre-learning stage of the experiment, it was not significant in the post-learning stage of the experiment (Figure 3B). Unfortunately, we could not attribute this effect to learning of specific behavioral associations, because the pattern was the same for faces that were not associated with behavioral information ($P = 0.17$ for the three-way interaction).

Brain regions responding to faces associated with behavioral information

Throughout the learning period there were three fROI that responded more to faces associated with behaviors than to faces not associated with behaviors: the left inferior frontal gyrus, left dmPFC and left parahippocampal gyrus/amygdala (Table 1). While each of these fROI responded preferentially to faces preceded by behavioral information, it is not clear whether they were involved in the actual behavioral learning reported above.

Table 1 Regions responding significantly more to faces presented with behaviors than faces presented without behaviors (during the learning period)

Region	Center of mass (*x*, *y*, *z*)	Volume (mm^3)	Peak *t*-value
Left inferior frontal gyrus	−43.5, 44.6, 2	1800	5.95
Left dmPFC	−3.8, 41, 31.2	900	4.94
Left parahippocampal gyrus/amygdala	−18.8, −8.8, −10.6	756	5.22

Center of mass coordinates referenced using the Talairach coordinate system.

Although each of these regions responded significantly more to faces associated with behaviors than faces not associated with behaviors and did not show differential responses to faces as a function of the valence of behaviors, we tested whether individual differences in the latter responses correlate with behavioral learning. To do this we calculated and then correlated behavioral learning effects (LE) and fROI-specific LE. The behavioral LE was calculated by subtracting the mean trustworthiness ratings of faces paired with negative behaviors from the mean trustworthiness ratings of faces paired with positive behaviors. These rating were collected during the behavioral portion of the study. Similarly, the fROI-specific LE was calculated by subtracting the mean parameter estimates for faces paired with negative behaviors from the mean parameter estimates for faces paired with positive behaviors. These fROI-specific LEs were calculated from the learning period data. Each participant thus contributed one behavioral LE value and three fROI-specific learning values (one for each fROI). As shown in Figure 4A, the behavioral LE and the fROI-specific LE were positively correlated within the dmPFC and this correlation was significant [$r(24) = 0.44$, $P < 0.032$]. The correlations between the behavioral LE and fROI-specific LEs in the left inferior frontal gyrus [$r(24) = 0.18$, $P = 0.40$] and left parahippocampal gyrus/amygdala [$r(24) = 0.0001$, $P = 0.99$] were not significant.

Exploring the relationships between behavioral learning, dmPFC and amygdala

The results of the correlation analysis suggest that the dmPFC plays a role in learning associations between faces and behaviors. We were interested whether the initial amygdala response to the perceived trustworthiness of faces would moderate the relationship between learning and dmPFC. To test this hypothesis, for each participant, we calculated the amygdala response to face-trustworthiness by subtracting the mean response of the right amygdala fROI to trustworthy faces from its mean response to untrustworthy faces during the pre-learning period. During this period, the faces were

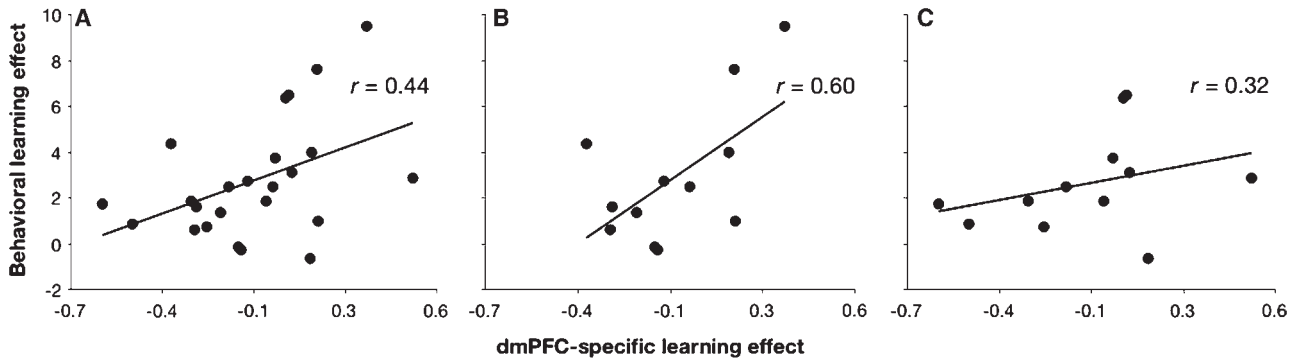

Fig. 4 (**A**–**C**) Scatter-plots of the dmPFC-specific learning and the behavioral learning effects. Each point represents a participant. (A) Plot including all 24 participants. (B and C) Plots of the 12 participants with the weakest (B) and strongest (C) initial amygdala response to face trustworthiness. The stronger the initial amygdala response to face trustworthiness, the weaker the relationship between the dmPFC-specific and behavioral learning effects.

free of behavioral associations and this response should reflect the initial assessment of the face. We then regressed the behavioral LE on the initial amygdala effect, the dmPFC-specific LE, and the interaction of the dmPFC and the amygdala. Consistent with the analysis described above, the dmPFC-specific LE significantly predicted the behavioral LE [standardized $b = 0.43$, $t(20) = 2.41$, $P < 0.05$]. More important, the interaction term was also significant [standardized $b = 0.46$, $t(20) = 2.28$, $P < 0.05$], indicating that the effect of the dmPFC on learning depended on the initial amygdala's response to faces. Specifically, the strength of the initial response of the amygdala affected the relationship between the dmPFC and learning. To illustrate this interaction, Figure 4B plots the relationship between the dmPFC LE and the behavior LE for participants with the weakest amygdala response to face trustworthiness and Figure 4C plots this relationship for participants with the strongest amygdala response. As shown in Figure 4, the stronger an individual's initial response to untrustworthy versus trustworthy faces in the amygdala, the weaker the relationship between the dmPFC activity and actual learning performance.

Additional whole-brain analysis

In addition to the aforementioned analyses, we were interested in determining whether there were any significant three-way interactions of learning, face trustworthiness, and behavioral description type across the entire brain. Consequently, we completed a whole-brain 2 (learning period: pre- and post-learning) × 3 (behavior type: positive, negative, null) × 2 (face trustworthiness: trustworthy and untrustworthy) repeated-measures ANOVA. From this analysis, an 828-mm^3 cluster of voxels in the right temporoparietal junction (rTPJ, Talairach coordinates for center of mass: $x = 55.2$, $y = -48.5$, $z = 24.8$; peak $F = 12.65$) survived corrections for multiple comparisons. See Figure 3E.

To understand this three-way interaction, we analyzed the average parameter estimate of all voxels within this fROI separately for the pre- and post-learning stages of the experiment. Although there was a significant interaction of behaviors and face trustworthiness in the pre-learning period [$F(2,46) = 3.19$, $P = 0.05$], this interaction was quantitatively weaker than the interaction in the post-learning period [$F(2,46) = 11.78$, $P < 0.0005$]. Given that behaviors had yet to be presented during pre-learning complicating the interpretation of the interaction in the pre-learning period, we focus on the post-learning interaction. For trustworthy faces presented during the post-learning period, those associated with negative behaviors evoked the weakest activation. In contrast, the weakest activation for untrustworthy faces resulted from those associated with positive or no behaviors. Specifically, pairwise comparisons (Bonferroni corrected) showed that trustworthy faces paired with negative behaviors ($M = -0.148$, $SD = 0.12$) caused significantly more deactivation than either trustworthy faces paired with positive ($M = -0.039$, $SD = 0.13$, $P < 0.01$) or no ($M = -0.058$, $SD = 0.15$, $P < 0.0005$) behavioral information. Although untrustworthy faces paired with negative behaviors ($M = -0.059$, $SD = 0.16$) evoked less deactivation than faces paired with either positive ($M = -0.14$, $SD = 0.13$, $P = 0.17$) or no ($M = -0.11$, $SD = 0.12$, $P = 0.819$) behavioral information, the pairwise comparisons did not reach significance.

A possible interpretation of the pattern of responses in the rTPJ during the post-learning period is in terms of face–behavior congruency (e.g. TP and UN = congruent; TN and UP = incongruent) (Figure 3F). A 2 (face–behavior congruency: congruent and incongruent) × 2 (face trustworthiness: trustworthy and untrustworthy) × 2 (learning period: pre- and post-learning) repeated-measures ANOVA confirmed that there was a significant interaction of congruency and time [$F(1,23) = 19.57$, $P < 0.0005$]. The effect of congruence was not significant in the pre-learning period [$F(1,23) = 2.35$, $P = 0.14$], but was significant in the post-learning period [$F(1,23) = 16.45$, $P < 0.0005$]. Incongruent face–behavior pairs ($M = -0.14$, $SD = 0.12$) caused significantly more deactivation that congruent pairs ($M = -0.059$, $SD = 0.15$).

DISCUSSION

In the current study, we examined the neural correlates of person impressions resulting from the integration of face- and behavior-based information. To do this we manipulated the perceived trustworthiness of faces that were subsequently paired with valenced behavioral information. Consistent with prior findings (Todorov and Olson, 2008; Rudoy and Paller, 2009), participants' judgments were affected not only by the trustworthiness of the faces but also by the learned face–behavior associations (Figure 2). Faces that were associated with positive behaviors were judged as more trustworthy than faces that were associated with negative behaviors.

Prior to learning these associations, the characteristic negative correlation between the trustworthiness of faces and amygdala activity (Winston *et al.*, 2002; Engell *et al.*, 2007) was observed in the right amygdala (Figure 3B). Increased activity to untrustworthy faces was not observed in the left amygdala. However, there is reason to believe that the left amygdala may have non-monotonic response properties with respect to face trustworthiness (Said *et al.*, 2008; Todorov *et al.*, 2008). In this case, the left amygdala would not show significantly different responses to trustworthy and untrustworthy faces.

We further explored the responses in the right amygdala fROI during the learning and post-learning periods. During the learning period, this region continued to show a stronger response to untrustworthy than trustworthy faces. In addition, this region responded more strongly to faces associated with behaviors than to faces not associated with behaviors. This finding replicates a prior study in which the authors observed a stronger response to faces associated with

valenced behaviors than to faces with no prior associations (Somerville *et al.*, 2006). In the post-learning period, we did not observe any significant effects, although the hemodynamic response to faces associated with negative behaviors was higher than the response to faces associated with positive behaviors. It is not clear why we did not observe significant effects in the post-learning period. Two possible reasons are amygdala habituation to the face stimuli and lack of statistical power given the relatively small number of trials per experimental condition. These reasons could also explain why we did not see any interaction of face trustworthiness, behavioral description type, and learning period in the right amygdala.

During the learning period, several regions showed stronger responses to faces that were preceded by behaviors than to faces that were not. These included the dmPFC, left inferior frontal gyrus and left parahippocampal gyrus/amygdala. The emergence of the dmPFC as a region responding more to faces associated with behaviors was unsurprising given previous literature showing similar dmPFC response properties (Kim *et al.*, 2004; Mitchell *et al.*, 2004, 2005, 2006; Todorov *et al.*, 2007). In a more specific test of each of the aforementioned regions' relationship to learning, we correlated behavioral LE with each ROI-specific LE. This analysis showed that dmPFC activity was significantly correlated with actual learning (Figure 4A). This result suggests the dmPFC is involved in behavior-based impression formation.

We evaluated this relationship between the dmPFC and learning as a function of the initial amygdala response to face trustworthiness. We observed that the amygdala response moderated the relationship between the dmPFC and behavioral learning. Specifically, a highly differentiated response to the trustworthiness of faces (trustworthy < untrustworthy) in the right amygdala during the pre-learning period led to a weaker effect of the dmPFC on learning during the learning period (Figure 4B and C). This result corroborates the findings of Kim *et al.* (2004)—who suggested that the dmPFC works as a 'convergence' zone for face and behavioral information that then interacts with the amygdala. In their study, participants were given behavioral cues that indicated the valence of faces displaying ambiguous emotional expression (Kim *et al.*, 2004). The authors found that the dmPFC was functionally connected with regions, including the amygdala, whose activity correlated with behavioral-cued valence (Kim *et al.*, 2004). The dmPFC did not emerge in a previous study that lacked behavioral cues for the ambiguously valenced faces (Kim *et al.*, 2003), suggesting that while the dmPFC correlates with amygdala activity, it only does so when processing relevant contextual behavioral information.

We believe our results suggest such an interactive relationship between the dmPFC and the amygdala. When an observer has a very strong initial reaction to the perceptually-based trustworthiness of a face, they are more likely to have a weaker response to the behavioral information later tied to that face. This weaker reaction is manifested both in a weaker dmPFC-specific LE during integration of behavioral information and poorer demonstrated learning in later testing. In a colloquial sense, the initial strong amygdala response may set a 'first impression' and act to buffer subsequent behavioral information from changing the initially formed impression. Indeed, there are behavioral results showing that the strength of the effect of face trustworthiness negatively correlates with the effect of behavioral learning on person judgments (Rudoy and Paller, 2009). The interaction between the amygdala and the dmPFC could describe a potential neural mechanism for such an effect.

In addition to the amygdala and dmPFC results, we observed a significant three-way interaction of learning period, behavioral description type, and face trustworthiness in the rTPJ. In the context of social neuroscience research, this region is most consistently identified in tasks requiring attribution of mental states to other people (Saxe and Kanwisher, 2003; Zaitchik *et al.*, 2010). The attribution of mental states involves integration of inferences about goals, motives and situational knowledge. In this respect, it is possible that the rTPJ could track and process behavioral information related to specific people over time. However, the specific response pattern to the different types of face–behavior pairs is atypical given previous literature. In a study manipulating the congruence of target individuals' background and their mental states, Saxe and Wexler (2005) found significantly *greater* rTPJ response to incongruent pairings. In our study, we found that incongruent face–behavioral description pairs actually evoked significantly *less* activation than congruent pairs. Given that the tasks and stimuli are quite different in the two experiments, it is not clear how to explain this apparent inconsistency. It is evident that further research is needed to determine the role of the rTPJ in updating of person representations.

The neural processes underlying impression formation are complex, and more regions than those discussed here are undoubtedly involved. Our most interpretable results revolve around the interactive relationship between the right amygdala and the dmPFC. Here, the initial amygdala response to face trustworthiness interacts with the response of the dmPFC during later impression formation to seemingly facilitate or inhibit behavioral learning. It is important to note that learning occurred even in participants with strong initial amygdala response and weak dmPFC-specific LE. Clearly, further research is needed to fully characterize the relationship between the dmPFC and amygdala as it relates to the processing of social information. For example, studying the effects of directionality (primacy of behavioral learning or appearance information) and consistency of information (multiple and varied person descriptions) on the dmPFC–amygdala interaction would be important for elucidating the nature of this interaction. Despite these standing concerns, our results join a growing set of studies finding an important relationship between dmPFC and the amygdala.

This relationship appears integral for our ability to move beyond initial impressions of a person and develop a more nuanced understanding of them based on their actions.

Conflict of Interest

None declared.

REFERENCES

Adolphs, R., Tranel, D., Damasio, A.R. (1998). The human amygdala in social judgment. *Nature*, *393*, 470–74.

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*, 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.

Bliss-Moreau, E., Barrett, L.F., Wright, C.I. (2008). Individual differences in learning the affective value of others under minimal conditions. *Emotion*, *8*, 479–93.

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

Downs, A.C., Lyons, P.M. (1991). Natural observations of the links between attractiveness and initial legal judgments. *Personality and Social Psychology Bulletin*, *17*(5), 541–7.

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.

Fuchs, R.A., Evans, K.A., Leford, C.C., et al. (2005). The role of the dorsomedial prefrontal cortex, basolateral amygdala, and dorsal hippocampus in contextual reinstatement of cocaine seeking in rats. *Neuropsychopharmacology*, *30*, 296–309.

Gobbini, M.I., Haxby, J.V. (2006). Neural response to the visual familiarity of faces. *Brain Research Bulletin*, *71*, 76–82.

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.

Kim, H., Somerville, L.H., Johnstone, T., Polis, S., Alexander, A.L., Whalen, P.J. (2003). Inverse amygdala and medial prefrontal cortex response to surprised faces. *NeuroReport*, *14*(18), 2317–22.

Kim, H., Somerville, L.H., Johnstone, T., et al. (2004). Contextual modulation of amygdala responsivity to surprised faces. *Journal of Cognitive Neuroscience*, *16*, 1730–45.

Lundqvist, D., Flykt, A., Ohman, A. (1998). *The Karolinska Directed Emotional Faces*. Stockholm: Psychology Section, Department of Clinical Neuroscience, Karolinska Institute.

McCulloch, K.C., Ferguson, M.J., Kawada, C.C.K., Bargh, J.A. (2008). Taking a closer look: on the operation of nonconscious impression formation. *Journal of Experimental Social Psychology*, *44*, 614–23.

McLaughlin, J., See, R.E. (2002). Selective inactivation of the dorsomedial prefrontal cortex and the basolateral amygdala attenuates conditioned-cue reinstatement of extinguished cocaine-seeking behavior in rats. *Psychopharmacology*, *168*(1–2), 57–65.

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. (2004). Encoding-specific effects of social cognition on the neural correlates of subsequent memory. *Journal of Neuroscience*, *24*, 4912–7.

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.

Nisbett, R.E., Wilson, T.D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, *35*, 250–6.

Ochsner, K.N., Gross, J.J. (2005). The cognitive control of emotion. *Trends in Cognitive Sciences*, *9*(5), 242–9.

Olivola, C.Y., Todorov, A. (2010). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior*, *34*, 83–110.

Oosterhof, N.N., Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion*, *9*, 128–133.

Rudoy, J.D., Paller, K.A. (2009). Who can you trust? Behavioral and neural differences between perceptual and memory-based influences. *Frontiers in Human Neuroscience*, *3*, 1–6.

Said, C.P., Baron, S.G., Todorov, A. (2009). Nonlinear amygdala response to face trustworthiness: Contributions of high and low spatial frequency information. *Journal of Cognitive Neuroscience*, *21*, 519–28.

Saxe, R., Kanwisher, N. (2003). People thinking about people: fMRI investigations of theory of mind. *NeuroImage*, *9*(4), 1835–42.

Saxe, R., Wexler, A. (2005). Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia*, *43*(10), 1391–9.

Somerville, L.H., Wig, G.S., Whalen, P.J., Kelley, W.M. (2006). Dissociable medial temporal lobe contributions to social memory. *Journal of Cognitive Neuroscience*, *18*, 1253–65.

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

Todorov, A., Baron, S., Oosterhof, N.N. (2008). Evaluating face trustworthiness: a model based approach. *Social, Cognitive, and Affective Neuroscience*, *3*, 119–27.

Todorov, A., Engell, A. (2008). The role of the amygdala in implicit evaluation of emotionally neutral faces. *Social, Cognitive, and Affective Neuroscience*, *3*, 303–12.

Todorov, A., Gobbini, M.I., Evans, K.K., Haxby, J.V. (2007). Spontaneous retrieval of affective person knowledge in face perception. *Neuropsychologia*, *45*, 163–73.

Todorov, A., Pakrashi, M., Oosterhof, N.N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, *27*, 813–33.

Todorov, A., Olson, I. (2008). Robust learning of affective trait associations with faces when the hippocampus is damaged, but not when the amygdala and temporal pole are damaged. *Social, Cognitive, and Affective Neuroscience*, *3*, 195–203.

Todorov, A., Uleman, J.S. (2002). Spontaneous trait inferences are bound to actors' faces: evidence from a false recognition paradigm. *Journal of Personality and Social Psychology*, *83*, 1051–65.

Todorov, A., Uleman, J.S. (2003). The efficiency of binding spontaneous trait inferences to actors' faces. *Journal of Experimental Social Psychology*, *39*, 549–62.

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.

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.

Zaitchik, D., Walker, C., Miller, S., LaViolette, P., Feczko, E., Dickerson, B.C. (2010). Mental state attribution and the temporoparietal junction: An fMRI study comparing belief, emotion, and perception. *Neuropsychologia*, *48*, 2528–2536.

Zebrowitz, L.A., McDonald, S. (1991). The impact of litigants' babyfacedness and attractiveness on adjudications in small claims courts. *Law and Human Behavior*, *15*, 603–23.