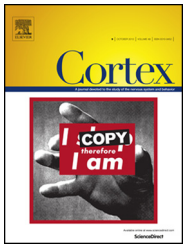
CORTEX 49 (2013) 2484–2493

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex



Research report

Decoding the representation of learned social roles in the human brain

Evelyn Eger [a,b,c,*], *Laura Moretti* [d], *Stanislas Dehaene* [a,b,c,e] *and Angela Sirigu* [d]

[a] *Institut National de la Santé et de la Recherche Médicale, Unit 992, Gif/Yvette, France*
[b] *Commissariat à l'Energie Atomique, NeuroSpin, Gif/Yvette, France*
[c] *Université Paris-Sud, Orsay, France*
[d] *Centre National de la Recherche Scientifique, Unit 5229, Bron, France*
[e] *Collège de France, Paris, France*

ARTICLE INFO

Article history:
Received 7 May 2012
Reviewed 20 August 2012
Revised 9 November 2012
Accepted 6 February 2013
Action editor Alan Sanfey
Published online 20 February 2013

Keywords:
Functional magnetic resonance imaging
Multivariate decoding
Face processing
Social cognition
Learning

ABSTRACT

Humans as social beings are profoundly affected by exclusion. Short experiences with people differing in their degree of prosocial behaviour can induce reliable preferences for including partners, but the neural mechanisms of this learning remain unclear. Here, we asked participants to play a short social interaction game based on "cyber-ball" where one fictive partner included and another excluded the subject, thus defining social roles (includer – "good", excluder – "bad"). We then used multivariate pattern recognition on high-resolution functional magnetic resonance imaging (fMRI) data acquired before and after this game to test whether neural responses to the partners' and neutral control faces during a perceptual task reflect their learned social valence. Support vector classification scores revealed a learning-related increase in neural discrimination of social status in anterior insula and anterior cingulate regions, which was mainly driven by includer faces becoming distinguishable from excluder and control faces. Thus, face-evoked responses in anterior insula and anterior cingulate cortex contain fine-grained information shaped by prior social interactions that allow for categorisation of faces according to their learned social status. These lasting traces of social experience in cortical areas important for emotional and social processing could provide a substrate of how social inclusion shapes future behaviour and promotes cooperative interactions between individuals.

1. Introduction

The need for social affiliation is central to normal human existence. Therefore, the act of social exclusion (or ostracism) practised across human societies and cultures and even by some non-human primates, is usually perceived as a powerful and emotionally distressful signal, even under very artificial circumstances such as a computer game (Williams et al., 2000). Being totally or partially socially excluded can lead the affected individual to try to conform or re-establish social links with the group (Williams et al., 2000), and to develop emotional preferences for partners with a higher tendency to include them (Andari et al., 2010). Functional imaging studies recording brain activity during experiences of social exclusion

* *Corresponding author*. INSERM, U. 992, CEA/Neurospin, Batiment 145, Point Courrier 156, 91191 Gif/Yvette, France.
E-mail address: evelyn.eger@gmail.com (E. Eger).
0010-9452/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cortex.2013.02.008

have revealed that the "social pain" experienced under these conditions seems to share neural substrates with physical pain, i.e., activations in right anterior insula, anterior cingulate, and lateral prefrontal cortex (Eisenberger et al., 2003; Masten et al., 2009; Sebastian et al., 2011). Intracranial electrophysiological recordings have also shown effects on theta power in insula and subgenual anterior cingulate cortex during the experience of exclusion (Cristofori et al., 2012). However, all of these studies were restricted to measuring activity directly while subjects experienced the exclusion situation. The underlying neural mechanisms of how people develop preference or aversion for partners following social interactions still remain largely unclear. Here, we explore the brain correlates of such learned social categorisations, using a modified version of the cyber ball game involving fictive partners with different profiles (includer/excluder). Previous behavioural studies have shown that in such a situation, normal volunteers quickly start to favour partners that included rather than excluded them, as measured by the number of ball tosses sent to those partners and various behavioural ratings such as trust and preference (Andari et al., 2010).

We tested the hypothesis that activity in the brain structures mediating the emotional reaction during exclusion would also carry information reflecting the learnt social categories when the involved partners are subsequently encountered in a context lacking explicit social or affective connotations. We measured with fMRI brain responses to different facial identities before and after a very brief social interaction (ball tossing game) with different partners. The analysis methods we used were based on pattern recognition (support vector classification) which predict from the subject's brain activity the face's social meaning as previously experienced in the ball tossing game (includer = "good"; excluder = "bad"; control face = "neutral"). This approach, often applied in other domains of imaging neuroscience for example to study perceptual representations (Haynes and Rees, 2006; Norman et al., 2006), is applied here, to our knowledge for the first time, to decode learned social attributes. Multivariate pattern recognition methods allow to differentiate between experimental conditions on the basis of the information present in the full pattern of activity across voxels and are therefore sensitive to distributed effects that may remain undetected by conventional mass-univariate mapping procedures testing only for circumscribed activation in- or decreases. In the present context, this approach permits to detect subtle changes in distributed activity patterns in regions of interest (ROIs) relevant for social emotion that arise with learning and discriminate between faces as a function of their learned social category.

2. Methods

2.1. Participants and fMRI acquisition

15 healthy young volunteers (11 male and 4 female aged 24.1 ± 3.8 years) were included in this study which had been approved by the regional ethics committee of Hôpital de Bicêtre, France. Functional images were acquired on a 3 T magnetic resonance (MR) system (Siemens Tim Trio) with 12-channel head coil as T2* weighted echo-planar image (EPI) volumes with 1.5 mm in-plane resolution. 33 transverse slices covering occipital, temporal and frontal lobes up to (ventral parts of) anterior cingulate cortex were obtained in interleaved order (repetition time (TR) 3 sec, field of view (FOV) 192 mm, echo time (TE) 30 msec, flip angle 78°, slice thickness 2 mm).

2.2. Stimuli and paradigm

The experiment consisted of a pre-learning and a post-learning phase during which fMRI data were recorded, and which were separated by the learning phase (social interaction game) without fMRI acquisition, see Fig. 1A. Stimuli were back-projected onto a screen located at a distance of 1 m from the subjects eyes at the end of the scanner bore and viewed via a mirror attached to the head coil. In the two rounds of the cyber ball game (one with male, and one with female partners), two face images were displayed subtending 3.5 × 4.5° visual angle (VA) at ~7.5° in the left and right visual field alongside cartoon characters representing the other players (see Fig. 1B). The partner defined as excluder ("bad"), after a short unbiased period, very rarely (20% of tosses) sent the ball to the subject whereas the includer ("good") directed 50% of tosses each to the participant and the other partner. Thus, different from Andari et al. (2010), who in a three-partner version of the game used an over-including (80/20) "good" partner, here we defined the "good" partner by a 50/50 profile since we had behavioural evidence of the effectiveness of this two-partner version with normal subjects and without monetary rewards at the time of the planning of the current study, and we wanted to avoid situations of an overly one-sided exchange between participant and includer. Moreover, in previous work, over-inclusion did not yield a clear advantage over being equally included (Williams et al., 2000).

The pictures of faces used for this studies (three men, three women) were chosen from the NimStim database (Tottenham et al., 2009). Starting with a pool of 20 faces with neutral expression, each of these faces was rated for attractiveness by a group of 10 subjects (not involved in the fMRI experiment) on a seven-point scale ranging from 1 (extremely unattractive) to 7 (extremely attractive). Based on these scores, we selected six faces (three female, three males) whose attractiveness scores were in the average range (between 3 and 4). Assignment of individual face identities to the three experimental conditions ("good" *vs* "bad" faces in the game *vs* "neutral" control faces not appearing in the game) was close to counterbalanced across subjects (fully counterbalanced within 12 subjects). "Good" and "bad" partners were assigned to opposite sides across the two rounds of the games (e.g., "good" = left, "bad" = right in first round, "bad" = left, "good" = right in second round). The choice of an unrelated face that did not appear in the game as "neutral" face condition, while making the design slightly unbalanced, was a compromise motivated by the need to keep the cyber ball game period sufficiently short and maximise time for fMRI acquisitions.

During the pre-learning and post-learning phases preceeding and following the ball tossing game, subjects viewed mini-blocks (four presentations of the same face, 1 sec on, .5 sec off) of the six different facial identities (three male, three

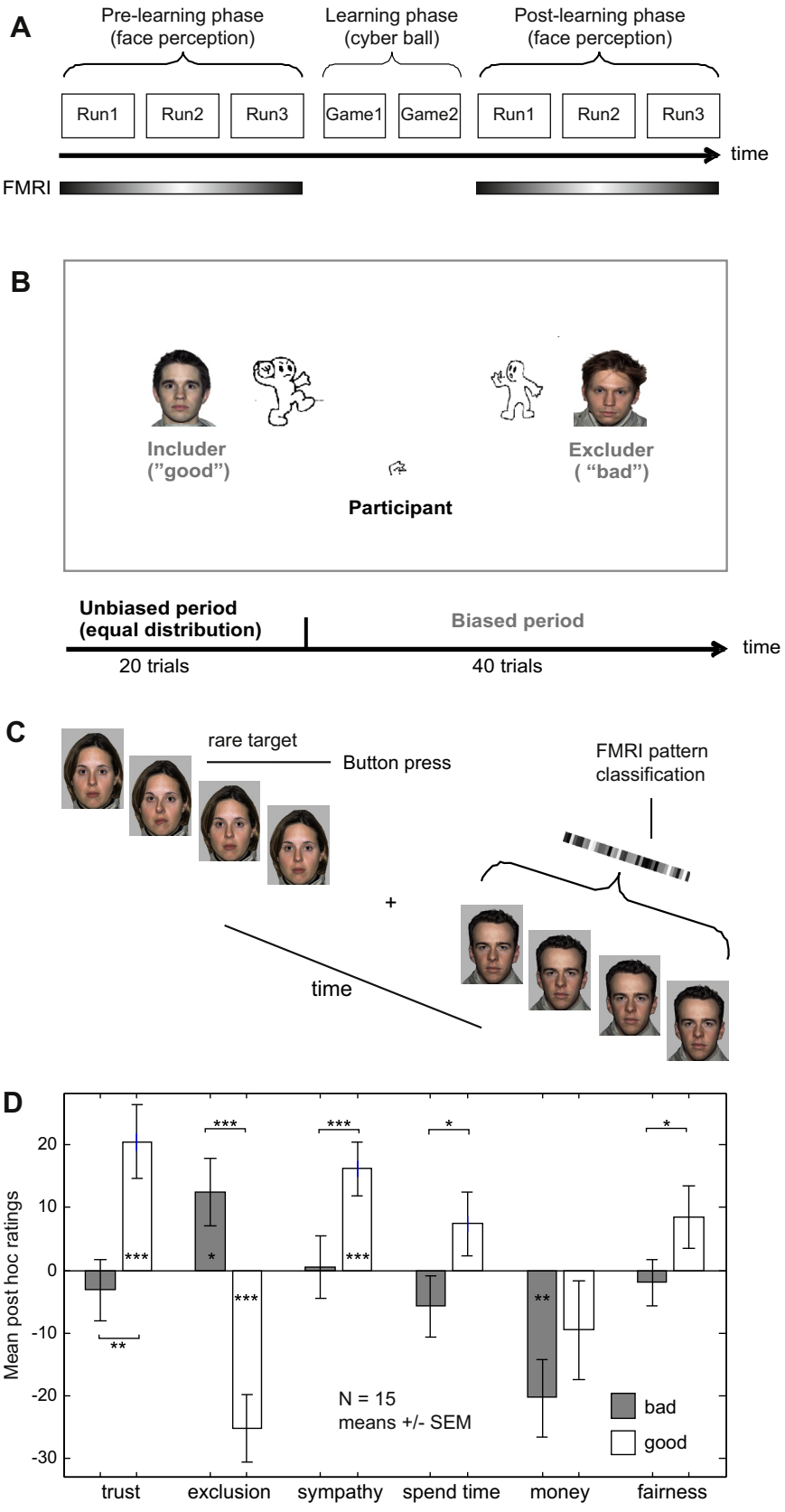


Fig. 1 – Paradigm and behavioural (rating) results: (A) Overall experimental timeline: The learning phase (two rounds of cyber ball game) was preceded by the pre-learning and followed by the post-learning phase with a face perception task. FMRI data (three runs each) were acquired during both pre- and post-learning phases. (B) During the cyber ball game the faces of two virtual partners were displayed next to cartoon characters in the left and right visual field. After an unbiased

female, of which for each gender one "good" and one "bad" partner and one unrelated "neutral" face) presented in randomised order at fixation and separated by baseline periods of 4 sec (see Fig. 1C). Face stimuli subtended approximately 5.5 × 7° VA. The subjects' task which was unrelated to the question of interest consisted in responding by button press to distorted versions of the same faces occurring in 25% of the mini-blocks. Participants completed three runs of ~8 min length each with this task before (pre-learning phase) and after (post-learning phase) the ball tossing game.

Upon completion of the post-learning imaging phase, subjects were re-presented with images of the faces of the four different partners encountered during the games and asked to provide ratings for each partner on a visual analogue scale ranging from −50 (NO) to 50 (YES) where 0 indicated neutrality (neither YES nor NO). The different dimensions rated were defined by the questions: (1) "Would you trust player X?", (2) "Did you feel excluded by player X?", (3) "Do you find player X sympathetic?", (4) "Would you like to spend time with player X?", (5) "Would you entrust money to player X?", and (6) "Do you find player X fair?". All participants in addition completed the "Need to belong scale" (Macdonald and Leary, 2005). This scale measures a person's desire to create or maintain interpersonal connections. Specifically, the measure assesses the desire to be accepted by others, the tendency to seek opportunities to belong to social groups and to react negatively to rejection or social ostracism. Items are scored on a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree), and items expressing a low need to belong are reverse-scored so that higher scores are a reflection of a greater need to belong.

None of our subjects spontaneously expressed doubts concerning whether they were playing against real opponents, although we did not explicitly debrief them on this issue, relying on previous findings that subjects can feel equally distressed by rejection when knowing that they are playing against a computer as when believing in the reality of the other players (Zadro et al., 2004).

2.3. *Analysis*

Pre-processing of the functional imaging data included motion correction, normalisation to Montreal Neurological Institute (MNI) space, and some slight smoothing (3 mm) for noise reduction using SPM5 (http://www.fil.ion.ucl.ac.uk/spm/software/spm5). A general linear model was set up including separate regressors for each mini-block of face stimuli. The onset of each face was modelled by a delta function convolved with a canonical haemodynamic response function and parameters were estimated after applying a high-pass filter of 128 sec and modelling serial autocorrelations as an auto regressive (AR) (1) process. Excluding those mini-blocks in which one of the pictures was a distorted face (target), this resulted in a total of 216 parameter estimate images which were used for pattern recognition analysis. More specifically, these were corresponding to 18 images per facial identity (2 good, 2 bad, and 2 neutral) and experimental phase (pre- *vs* post-learning), coming from six mini-blocks per facial identity in each of the three runs in each experimental phase. In our main analyses discriminating between valences the images were pooled across the two facial identities with the same valence. The number of images per condition in these comparisons was therefore 36.

ROIs (insula, anterior cingulate, ventrolateral prefrontal cortex, ventromedial prefrontal cortex, amygdala, visual cortex) were defined on the basis of anatomical masks in MNI space derived from Wake Forest University (WFU)-PickAtlas (http://fmri.wfubmc.edu/cms/software). Within each mask, the 1000 voxels most activated versus baseline across all face conditions were selected for each subject (with the exception of the amygdala mask, where only 200 voxels were used due to the smaller volume of the region). For the analysis testing separately effects in the anterior and posterior insula, the original PickAtlas ROI was divided in y-direction into two equally-dimensioned subparts, and within each part 500 voxels were selected based on the same criterion as described above. We used the population average, landmark and surface based (PALS) Atlas (Van Essen, 2005) of Caret 5.51 software (http://www.nitrc.org/projects/caret/) for visualisation of ROIs across subjects.

Pattern classification analysis employed support vector machines (SVM) (Christianini and Shawe-Taylor, 2000) in the form of a linear soft-margin classifier (regularisation parameter C fixed to 1) using the Scikit-Learn software (Pedregosa et al., 2011)(http://scikit-learn.org/stable/). All results correspond to accuracies for leave-one-out with cross-validation (more specifically, one of n patterns of each conditions, with a pattern corresponding to a parameter estimate image for one mini-block of four times the same face, was held out for test at each fold of the cross-validation cycle, while the classifier was trained on the remaining $n-1$ patterns for each condition). The main classification results are based on all the voxels within each ROI as described above without further feature selection. To explore the effect of additional

period of 20 ball tosses, the "bad" player was excluding the subject (80% of balls directed to the other virtual partner, and 20% to the subject), while the "good" player was directing balls with equal frequency to the subject and the other partner. (C) In the fMRI runs carried out before and after the social game (pre-and post-learning phase), subjects viewed mini-blocks of four presentations of 1.5 sec of a given face, for "good", "bad", and "neutral" faces (unrelated faces not participating in the ball tossing game) presented in random order and interleaved with 4 sec of baseline. The task consisted in detection of distorted faces, occurring within 25% of the mini-blocks, which were not considered in the analysis of the functional imaging data. (D) Post-experimental ratings of the two "good" and two "bad" faces on a scale ranging from −50 to 50 revealed significant differential effects of valence on measures of trust, feeling of exclusion, sympathy, fairness, and willingness to spend time. Means ± standard error of the mean (SEM) across subjects for "good" and "bad" face conditions are plotted. The *, **, and * signs correspond to significance at $p < .05$, $p < .01$, and $p < .001$, respectively. On the brackets linking two different bars, they refer to the effect of valence in an analysis of variance (ANOVA) for each question, and when placed directly on the bars to contrasts with respect to 0 (which represented no preference).**

(univariate) pre-selection of voxels on classification performance (reported as supplements), an *F*-test was computed separately for each fold of the cross-validation loop on the training data only, testing for discrimination between the two relevant conditions, e.g., good versus bad, bad versus neutral or good versus neutral. The *n* voxels with the highest *F*-values for each fold were then selected for classification, and this procedure was repeated for increasing numbers of voxels in steps of 100. An equivalent approach was applied in the case of across-subject correlation analysis between classification scores and questionnaire scores for "need to belong".

For the initial analysis testing for learning-related effects of discrimination of social status, images were collapsed across the two faces of the same valence ("bad", "good", "neutral"), thus yielding 36 images per valence condition, and pairwise classification accuracies were obtained for all three possible pairs of these conditions, separately for pre- and post-learning phases, within each subject and ROI. We tested for significant learning effects (changes in classification accuracies due to learning) across subjects with repeated measures ANOVAs including as factors experimental phase (pre-learning *vs* post-learning) and valence comparison ("good" *vs* "bad", "good" *vs* "neutral", "bad" *vs* "neutral"). The reported degrees of freedom are adjusted for non-sphericity using Greenhouse–Geisser correction.

To test for increased confusion of same-valence faces as a result of learning, pairwise classifications were run between all potential pairs of individual faces, including those of same valence. These values constitute the full confusion matrix of the classifier, once again separately for pre-learning and post-learning phase. After subtracting the pre-learning (baseline) from the post-learning confusion results, averages were taken separately across those off-diagonal cells that correspond to confusion of same valence and different valence, respectively, within each subject. The difference between same valence and different valence confusion was tested for statistical significance across subject by a paired *t*-test.

An exploratory whole-brain searchlight analysis was performed in addition which, for reasons of computational efficiency, was based on a correlation approach instead of a classification/cross-validation procedure. Correlations were calculated within the spherical environment (with four voxels or 6 mm radius) of each voxel in the volume between a given pair of conditions (e.g., "good" and "bad" faces, averaged across all blocks of those conditions in each experimental phase) yielding separate correlation maps for pre- and post-learning phases. Correlation values were subsequently Fisher-*z* transformed, and the resulting images entered into a random-effects group analysis (implemented in statistical parametric mapping (SPM)) testing for regions showing significantly lower correlations (more dissimilar patterns) in the post- than the pre-learning phase. Results are reported corrected for family-wise error by Gaussian random fields theory as implemented in SPM.

3. Results

3.1. Behaviour

The social learning paradigm employed consisted of two rounds of ~5 min of a modified version of the "cyber-ball" game (Williams et al., 2000) with two fictive partners (see Fig. 1B). FMRI data were acquired just before and just after playing the game. After completion of the whole experiment, subjects were asked to rate the four (during the two rounds) partners' behaviour along six different socially relevant dimensions (see Methods for the corresponding questions). The results of these ratings indicated significant differential effects between "good" and "bad" partners for the feeling of exclusion, trust, sympathy, fairness and willingness to spend time with that person (see Fig. 1D). These effects all pointed in the expected direction, indicating that participants had learned which partners were "good" or "bad". Analysis of the subjects' ball tosses during the game showed that, in this short version of this paradigm, participants sent balls on average more often to the excluding (~60%) than the including partner, $F(1,14) = 5.9$, $p < .05$.

3.2. FMRI decoding

To understand the brain correlates underlying the learned social categories as revealed by the behavioural ratings, we used a linear classifier based on SVM to test whether the faces' social meaning ("good", "bad", "neutral") could be decoded from the subjects' distributed brain activity. We applied pattern classification to relevant ROIs previously shown to be involved in the experience of social exclusion or emotion more generally such as insula, anterior cingulate cortex, ventrolateral prefrontal cortex, ventromedial prefrontal cortex, and amygdala, which were defined by a combination of anatomical (masks in MNI space) and functional criteria (see Methods for details). In addition visual cortex was tested as a lower-level sensory region which we would expect to discriminate between the different facial identities based on their visual features, but not necessarily to show an effect of social learning. We operationalise social learning effects here as either a main effect of experimental phase (pre- *vs* post-learning) or an interaction of experimental phase with valence comparison, on classification accuracies. Results for visual cortex showed that while the different face conditions could be reliably discriminated with an accuracy of ~60%, this discrimination was not affected by social learning [absence of main effect of experimental phase, $F(1,14) = .2$, or interaction with valence comparison, $F(1.5,21.3) = .4$, see Fig. 2A].[1]

For the insula region, both the main effect of experimental phase, $F(1,14) = 6.3$, $p < .05$, and interaction, $F(1.8,24.5) = 4.3$, $p < .05$, reached significance. As can be seen in Fig. 2B, all possible discrimination tests for faces with different valence were at chance before the game, but "good" and "bad" faces showed enhanced discrimination after the game [significant post- *vs* pre-learning difference for this comparison, $F(1,14) = 6.4$, $p < .05$]. Interestingly, insula classification results also indicated that learning the social valence of faces was mainly reflected in a change in the representation of "good" compared to "neutral" faces [significant post- *vs* pre-learning

[1] The used comparison included lower as well as higher visual areas. When restricting analysis to voxels responsive to faces in a separate localiser run comparing faces to phase-scrambled face images we did confirm the absence of significant effects, with overall lower accuracies.

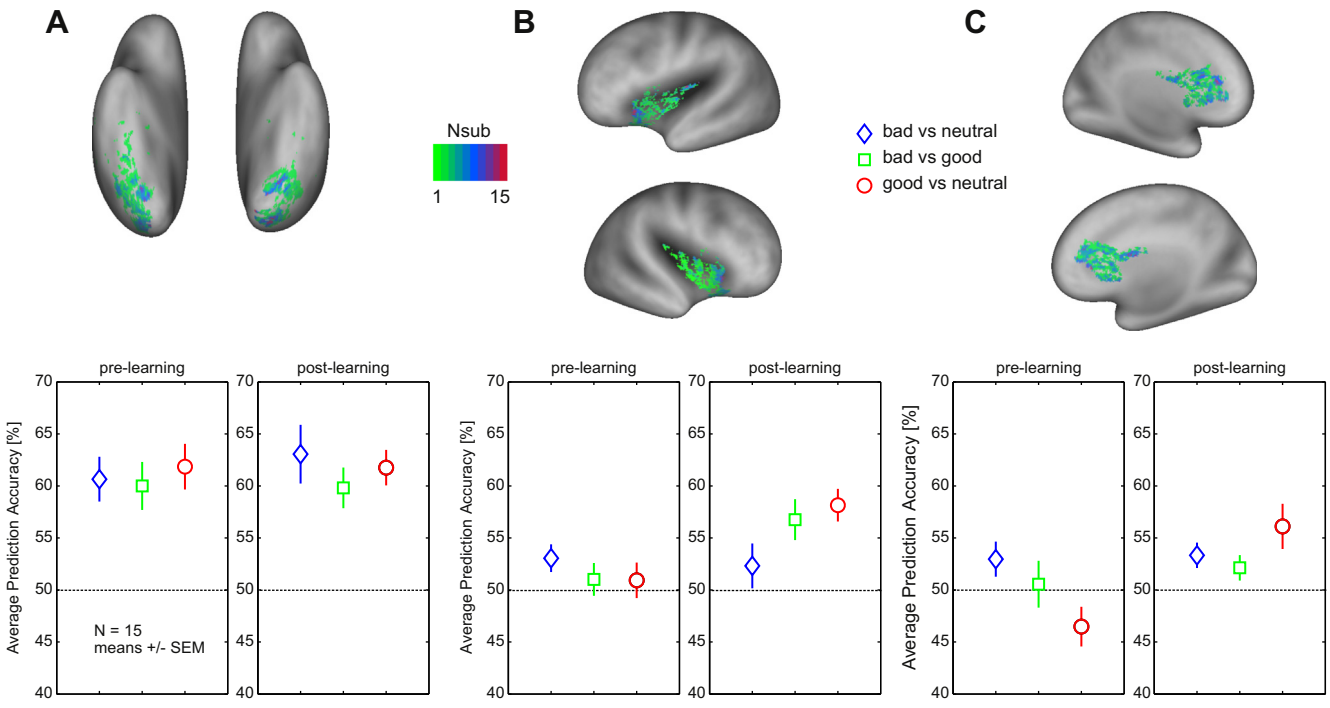


Fig. 2 – FMRI pattern discrimination for faces of different social valence. Results of support vector classification for different ROIs (N = 15, means ± SEM). The 1000 most activated voxels across all face conditions versus baseline within masks of occipito-temporal cortex, insula, and anterior cingulate cortex were used as a ROI on a subject-by-subject basis. The surface mappings (Caret PALS Atlas) illustrate the regions included and the across-subject overlap of voxels (colour coding indicating the number of subjects for which the corresponding voxel was chosen). (A) While visual cortex discriminated between the different faces well above chance, this discrimination was not affected by learning of social valence. Regions showing either an effect of experimental phase or interaction with valence comparison were: Insula (B) which showed close to chance discrimination of faces in the pre-learning phase, while in the post-learning phase faces of different social valence become discriminable [mostly due to "good", but not "bad" faces becoming discriminable from "neutral" (control) faces], and anterior cingulate cortex (C) where a learning effect is also mainly observed for "good" compared to "neutral" faces.

difference, $F(1,14) = 7.9$, $p < .05$], whereas "bad" faces did not become more distinguishable from "neutral" faces [post- *vs* pre-learning difference, $F(1,14) = .1$].

Previous studies have indicated a difference in the involvement of the posterior and anterior insula in emotional processing (Lamm and Singer, 2010). Following this we examined if there were also differences in social learning between these two insula sub-regions. After subdividing our previous insula ROI along the anterior–posterior axis (see Supplemental Fig. 1), we indeed observed a significant main effect of experimental phase, $F(1,14) = 19.0$, $p < .001$, and interaction, $F(1.9,27.1) = 5.1$, $p < .05$, in the anterior sector, while the same main effect and interaction remained non-significant in the posterior one, $F(1,14) = 1.4$ and $F(1.6,22.9) = .4$. The region by experimental phase (pre- *vs* post-learning) interaction was significant [$F(1,29) = 7.78$, $p < .05$]. No significant effect was found between left and right insula [interaction region by experimental phase: $F(1,14) = 1.04$, region by experimental phase by condition: $F(1.7,24.4) = .75$].

The only other region showing either a significant main effect of experimental phase or interaction with valence comparison was the anterior cingulate cortex ROI [significant main effect, $F(1,14) = 6.4$, $p < .05$, tendency for significant interaction, $F(1.5,21.3) = 3.7$, $p = .051$]. As can be seen in Fig. 2C, in this region as in the insula, social learning resulted in enhanced discrimination of "good" from "neutral" faces [significant post- *vs* pre-learning difference, $F(1,14) = 10.8$, $p < .01$].

Critically, the learning effects observed in the insula and anterior cingulate cortex were not driven by changes in the overall blood oxygen dependent (BOLD) signal in these regions, since a similar classification analysis based on the ROI mean signal did not reveal significant effects in either the insula or the anterior cingulate cortex (ACC) [insula: effects of experimental phase: $F(1,14) = .02$, experimental phase × valence comparison: $F(1.8,25.1) = .56$, ACC: effects of experimental phase: $F(1,14) = .94$, experimental phase × valence comparison: $F(2.0,27.7) = .67$]. Furthermore, univariate SPM group analyses testing for the interaction of experimental phase and valence did not reveal any significant voxels surviving correction for either the whole brain or the small volumes determined by the anatomical mask for insula or ACC (see Supplemental Fig. 2).

3.3. *Categorical coding of valence*

Our previous results showed that activation patterns for faces of different learned social roles become discriminable after a

brief learning experience in the anterior insula and the anterior cingulate cortex. But so far it remains unclear in how far this finding represents a true categorical effect of social valence or whether the activation pattern for all faces is modified to some degree. Testing the hypothesis of a categorical effect of valence after learning requires investigating changes in the representation of faces having the same valence (while the previous comparisons collapsed across same-valence faces). To address this point, we hence calculated the full confusion matrix of the classifier by comparing each individual face with each other individual face. In these matrices shown for pre- and post-learning phases for the insula ROI in Fig. 3, the secondary diagonals (marked in red colour) correspond to confusions of faces of the same valence. The average post- versus pre-learning difference in confusion for pairs of same-valence faces was significantly greater than zero [$t(14) = 2.52$, $p < .05$] and greater than for pairs of different-valence faces [$t(14) = 3.6$, $p < .01$]. This result indicates that the similarity of faces with the same social valence increased in the insula as an effect of learning.

3.4. *Inter-individual differences in behaviour and fMRI pattern discrimination*

As described above, behavioural data from ball tosses indicate that subjects did not develop an overall preference for the "good" partner. Instead, on average they sent a higher proportion of balls towards the "bad" partner, contrary to what was observed in a slightly different version of this game (Andari et al., 2010). Nevertheless, we reasoned that individual subjects might show different degrees of preference for the "good" partner as expressed by ball tosses, and that this might account for the strength of neural preferences for the good partner as expressed by fMRI pattern discrimination after learning. Indeed, the across-subject correlation between proportion of ball tosses towards the "good" partner and "good versus "neutral" fMRI pattern discrimination reached significance in the anterior insula ($r = .55$, $p < .05$). We further tested whether the same fMRI comparison might be modulated by individual questionnaire scores for the "need to belong" dimension. Including all voxels in the ROI this correlation ($r = .38$) was not significant. In a more refined analysis considering classification scores with additional feature selection it reached significance: for 400 voxels ($r = .69$, $p < .01$) or 300 voxels ($r = .55$, $p < .05$), see Supplemental Fig. 3, and Methods for details of voxel selection. Thus, pattern classification results in the anterior insula reflect to some extent individual differences in behavioural social measures.

3.5. *Whole-brain multi-voxel pattern analysis (MVPA)*

Finally, to evaluate whether differences in response patterns due to learned social valence existed in other parts of the brain not covered by our ROIs, we performed an exploratory MVPA of the whole brain based on a searchlight-correlation approach (see Methods for details). This analysis is considering effects of a more local nature than our previous ROI analyses, since it used a small spherical neighbourhood of each voxel (6 mm radius). We found significant decreases in pattern similarity between "good" and "bad" faces (post-learning as compared to pre-learning) in the superior temporal sulcus (STS) (corrected for multiple comparisons across the whole brain at the MNI coordinates 50 −60 3 ($t = 9.91$), 56 −44 12 ($t = 8.86$), and 62 −24 −5 ($t = 8.79$), see Supplemental Fig. 4). Results for the left hemisphere, as well as for "good" versus

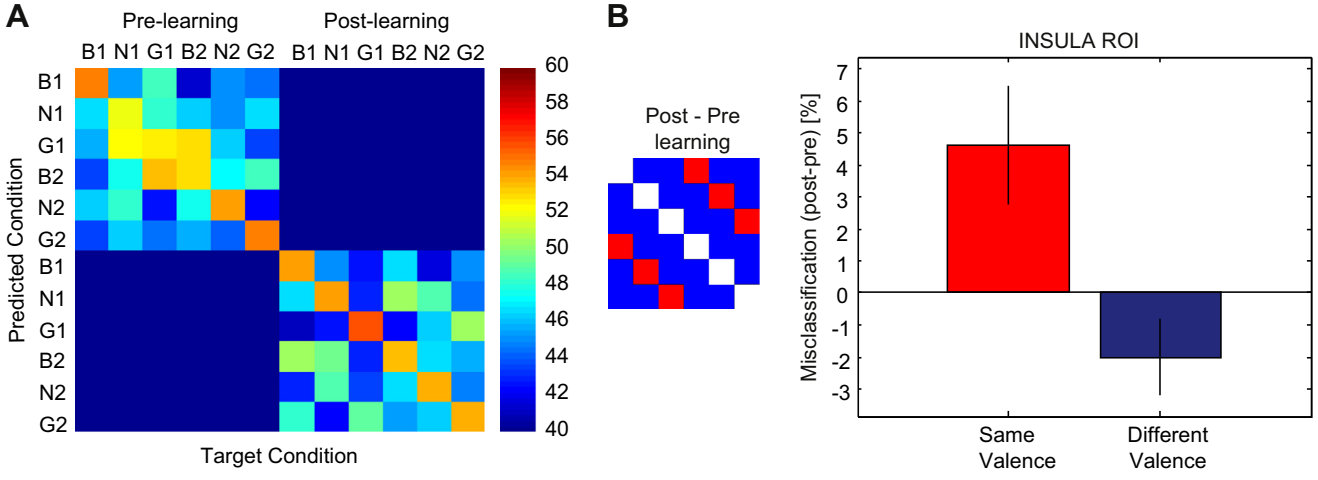


Fig. 3 – Confusion between individual faces and test for a "categorical" effect of social valence (enhanced confusion of same-valence faces) due to learning in the insula ROI. (A) Confusion matrices were created by classifying, for pre- and post-learning phase, between all possible pairs of the six faces and plotting the real condition (*x*-axis) against the predicted condition (*y*-axis). The main diagonal corresponds to the percentage of correct identification for each condition (averaged across all pairwise comparisons). The off-diagonal cells correspond to the percentages of misclassifications of a given condition as one of the other five conditions. The matrices show a slight tendency for secondary diagonals (corresponding to confusions of faces with the same valence) to appear in the post-learning phase. This was quantified by comparing (B) the average post- versus pre-learning difference (learning effect) in confusions for faces of the same valence (read cells) with the corresponding difference for faces of different valence (blue cells). The learning effect on confusion was greater for same than different-valence faces, and significantly larger than 0 for same-valence faces.

"neutral" and "bad" versus "neutral" did not reach a corrected level of significance.

The whole-brain analysis therefore failed to show significant effects equivalent to the ones observed in the insula and ACC in the ROI-based analysis. One reason for this discrepancy could be the use of a different method (correlation instead of SVM classification) which was chosen for reasons of computational efficiency. However, based on previous studies, our experience is that even if whole-brain searchlight analysis is based on the same methods (SVM) as ROI analyses, it does in some cases not show equivalent effects to those observed in individually defined ROIs, if the ROI effects are not very strong in terms of prediction accuracy (as is also the case in the current study). This is likely due to a combination of several factors: the smaller number of voxels included in the searchlight as opposed to the ROI, the inclusion of white matter voxels, and factors applying to whole brain versus ROI analyses in general such as the insufficient alignment of regions across subjects in MNI space, and multiple comparison correction.

4. Discussion

In the current study, we investigated the brain correlates of learned social roles after a brief social interaction with fictive partners expressing different degrees of inclusion. Our results show that subjects discriminated the different social profiles, as confirmed by the ratings obtained for the partners along several socially relevant dimensions. Furthermore, the activity patterns of the includer ("good") and excluder ("bad") faces became more discriminable in the insula and the anterior cingulate cortex. These regions are known for being involved in social emotions and more specifically in processing pain generated by social exclusion (Eisenberger et al., 2003; Sebastian et al., 2011). Crucially, in our study learning-related changes were found when examining the multi-voxel-activity pattern (which allowed for significant read-out of the faces social role from the participants brain activity), but not when considering the mean signal within the same region of interest (for which discrimination accuracy was at chance). While individual faces were discriminated in visual cortex, we did not find any significant effect of social learning in these areas, despite the evidence for emotional modulation of visual cortex in the literature (Vuilleumier and Driver, 2007). One possibility is that visual cortex may have been affected by emotion during, but no longer after the game, and that the learning experience used here was too brief to have an impact on perceptual representations. It is also conceivable that visual cortex does change with learning, but that the dimension used by the classifier for discriminating the different categories may be dissimilar from the dimension that changes with learning. Alternatively, since emotion has been shown to modulate early components of the visually evoked response to faces (Pizzagalli et al., 2002), such early effects also could occur when learning to discriminate social partners but be insufficiently captured by fMRI. Notably, it now has been shown that when using intracranial electrophysiological recordings, activity in regions of the fusiform gyrus is modulated during periods of social exclusion (Cristofori et al., 2012).

Our finding of discrimination of social roles in the insular cortex was specifically related to activity of the anterior sector compared to the posterior one. While the posterior insula contains somatotopic interoceptive representations (Craig, 2002), the anterior insula has been shown to be involved in more complex feeling states and social emotions having both negative and positive valence such as empathy for pain (Kong et al., 2006; Singer et al., 2004b), disgust (Jabbi et al., 2008; Wicker et al., 2003), judgement of trustworthiness (Winston et al., 2002), maternal affiliation (Bartels and Zeki, 2004; Leibenluft et al., 2004), etc. Interestingly, we also found another area, the ACC showing increased discriminative information for partners' social profiles. Anterior insula and ACC have been proposed to form an essential network underlying all subjective feelings (Craig, 2009). Both regions are characterised neuroanatomically by the existence of the so-called von Economo neurons which are typically found in large-brained, highly social mammals, and the presence of these specific cells correlates across species with tests of self-awareness (Craig, 2009). Self-awareness indeed has been suggested as an essential precondition for the development of intentional emotional interactions with conspecifics (Frith and Frith, 2007). Furthermore, degeneration of von Economo neurons in fronto-temporal dementia (FTD) leads to deficits in social-emotional functioning (Seeley, 2010).

While in our study we found that "good" and "bad" faces evoked different activity patterns after learning their role in the cyber ball game, in both anterior insula and anterior cingulate only "good" faces were distinguishable from "neutral" faces (faces that did not participate in the ball tossing game but that were nevertheless familiar through having been presented many times in the pre-learning phase). This preferential learning effect for including partners with a positive valence was unexpected given that the insula responds particularly during the negative experience of total exclusion (Eisenberger et al., 2003; Masten et al., 2009; Sebastian et al., 2011). Nevertheless, two previous studies using conventional mapping methods to investigate learning of faces in different economic paradigms as prisoner's dilemma (Singer et al., 2004a) and ultimatum (Chang and Sanfey, 2009) games showed a similar pattern of results by finding lasting learning effects preferentially for positively valenced, fair partners. However, a theory predicting in which cases the regions in question would respond to negatively as opposed to positively valenced information is currently still lacking.

One possibility, as highlighted by our results, is that the discrimination of "good" (but not "bad") from "neutral" faces may be related to an increased subjective importance or salience of the good partners'/includers' faces after the experience of exclusion. This would be in line with the notion of the insula and the anterior cingulate cortex being part of an emotional salience network (Seeley et al., 2007). The ratings obtained at the end of the experiment clearly support the claim that subjects developed strong preferences for the good partner. They judged the good player as the most trustworthy, the fairest one, the less excluding, the one they found most sympathetic, the one they are willing to spend time with, and the ratings for the good player differed significantly from zero in more cases than for the bad player (as shown in Fig. 1).

Could it therefore be possible that differences in insula activation patterns were modulated by uncertainty/predictability instead of social valence (Preuschoff et al., 2008)? From the point of view of the partners' trial-by-trial behaviour, subjects should learn to some extend to predict the bad partners' behaviour in the game (he/she is likely not going to pass them the ball). On the other hand, the good partner is passing the ball in 50% and is therefore more unpredictable on a trial-by-trial basis. On the other hand, uncertainty about the neutral faces' behaviour should be high as well, and this does not seem to fit with the pattern of discriminations observed. On the other hand, since the rating results mentioned above suggest that the role of the bad player was less strongly perceived than the one of the good player, it is possible that these player's more general role (e.g., he/she is likely to be unfair) is less well predicted than the one of the good players (e.g., he/she is likely to be fair) after the short duration of the game.

In contrast to previous results (Andari et al., 2010) where participants directed a higher proportion of ball tosses towards the "good" partner, in our case the opposite was true. This is likely to be due to differences in task design: The paradigm used by Andari et al. involved three players instead of two (as in our case), where the "good" partner over-included the participant. These differences, in combination with the short duration of the game in our case, might account for the different behavioural results observed. It is also conceivable that subjects sent the ball more often to the "bad" partner in an attempt to make him cooperate, or to explore the limits of his non-responsiveness. Following this scenario, subjects might have shifted their game preference towards the good partner if they were allowed to play more trials.

Two further details of our results investigating inter-subject variability suggest that the insula response is related to how strongly subjects were sensitive to or affected by the different roles of the two partners: we found a significant correlation indicating that the more subjects tossed the ball towards the good partner, the more they discriminated "good" and "neutral" faces in the anterior insula (but not significantly in the ACC) after the game. Along a similar vein, subjects with higher "need to belong" scores showed a stronger difference between activity patterns for "good" and "neutral" faces in this region, potentially consistent with a stronger need for those subjects to form compensatory emotional bonds after experiencing exclusion.

Beyond the regions related to social exclusion and emotion in the primary focus of our study, a more exploratory whole-brain searchlight analysis indicated that after learning faces of different valence were coded by increasingly dissimilar patterns in superior temporal sulcus regions. Apart from their relevance for coding of more low-level properties as biological motion, facial expressions or gaze direction (Haxby et al., 2000; Puce and Perrett, 2003) which were unchanged in our case, these areas have been shown to be involved in tasks requiring so-called theory of mind and have been hypothesised to underlie more specifically the prediction of others' intentions (Frith and Frith, 2007; Saxe et al., 2004). The learning effect leading to greater differentiation of faces with different social roles observed here may therefore reflect the different intentions ascribed to the partners and suggests a fine-scale representation of information in these regions that goes beyond overall modulation by tasks requiring cognitive activities such as mentalising or predictability of intentions.

Altogether, this first application of multivariate decoding to social learning suggest that beyond many applications in more low-level domains of cognitive neuroscience, where decoding may or may not capitalise on columnar layouts of the features in question (Freeman et al., 2011; Haynes and Rees, 2006), and some recent applications to emotion or pain processing (Corradi-Dell'Acqua et al., 2011), this pattern-based analysis approach is also able to detect subtle changes in activity patterns caused by a very short social interaction. Learning in our case mainly affected categorisation of "good" partners, and a preferential creation of positively valenced associations could provide a mechanism for promoting future cooperative interactions and thus have survival value. Still, further studies will have to clarify whether regions expressing learning for "bad" partners are identical with or dissociable from the ones observed here, in cases where the role of the "bad" partner is more salient than it was in our brief learning experience. Another important aim will be to elucidate whether the neural mechanisms we have shown here are also the ones accounting for the lack of social learning in autism (Andari et al., 2010), for which the anterior insula is one important region of interest (Di Martino et al., 2009).

Acknowledgements

We thank Andreas Kleinschmidt for helpful comments on an earlier version of this manuscript. This work was supported by Centre National de le Recherche Scientific (CNRS) and Fondation pour la Recherche Medicale (AS and LM). This experiment was part of a general research program on functional neuroimaging of the human brain which is sponsored by the Atomic Energy Commission (Denis Le Bihan). We thank the NeuroSpin platform staff for their help.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.cortex.2013.02.008.

REFERENCES

Andari E, Duhamel J-R, Zalla T, Herbrecht E, Leboyer M, and Sirigu A. Promoting social behavior with oxytocin in high-functioning autism spectrum disorders. *Proceedings of the National Academy of Sciences of the United States of America*, 107(9): 4389–4394, 2010.

Bartels A and Zeki S. The neural correlates of maternal and romantic love. *NeuroImage*, 21(3): 1155–1166, 2004.

Chang LJ and Sanfey AG. Unforgettable ultimatums? Expectation violations promote enhanced social memory following economic bargaining. *Frontiers in Behavioral Neuroscience*, 3: 36, 2009.

Christianini N and Shawe-Taylor J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press, 2000.

Corradi-Dell'Acqua C, Hofstetter C, and Vuilleumier P. Felt and seen pain evoke the same local patterns of cortical activity in insular and cingulate cortex. *Journal of Neuroscience*, 31(49): 17996–18006, 2011.

Craig AD. How do you feel? Interoception: The sense of the physiological condition of the body. *Nature Reviews Neuroscience*, 3(8): 655–666, 2002.

Craig ADB. How do you feel–now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, 10(1): 59–70, 2009.

Cristofori I, Moretti L, Harquel S, Posada A, Deiana G, Isnard J, et al. Theta signal as the neural signature of social exclusion. *Cerebral Cortex*, 2012. epub ahead of print.

Di Martino A, Ross K, Uddin LQ, Sklar AB, Castellanos FX, and Milham MP. Functional brain correlates of social and nonsocial processes in autism spectrum disorders: An activation likelihood estimation meta-analysis. *Biological Psychiatry*, 65(1): 63–74, 2009.

Eisenberger NI, Lieberman MD, and Williams KD. Does rejection hurt? An FMRI study of social exclusion. *Science*, 302(5643): 290–292, 2003.

Freeman J, Brouwer GJ, Heeger DJ, and Merriam EP. Orientation decoding depends on maps, not columns. *Journal of Neuroscience*, 31(13): 4792–4804, 2011.

Frith CD and Frith U. Social cognition in humans. *Current Biology*, 17(16): R724–R732, 2007.

Haxby, Hoffman, and Gobbini. The distributed human neural system for face perception. *Trends in Cognitive Sciences*, 4(6): 223–233, 2000.

Haynes J-D and Rees G. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7: 523–534, 2006.

Jabbi M, Bastiaansen J, and Keysers C. A common anterior insula representation of disgust observation, experience and imagination shows divergent functional connectivity pathways. *PloS One*, 3(8): e2939, 2008.

Kong J, White NS, Kwong KK, Vangel MG, Rosman IS, Gracely RH, et al. Using fMRI to dissociate sensory encoding from cognitive evaluation of heat pain intensity. *Human Brain Mapping*, 27(9): 715–721, 2006.

Lamm C and Singer T. The role of anterior insular cortex in social emotions. *Brain Structure and Function*, 214(5–6): 579–591, 2010.

Leibenluft E, Gobbini MI, Harrison T, and Haxby JV. Mothers' neural activation in response to pictures of their children and other children. *Biological Psychiatry*, 56(4): 225–232, 2004.

Macdonald G and Leary MR. Why does social exclusion hurt? The relationship between social and physical pain. *Psychological Bulletin*, 131: 202–223, 2005.

Masten CL, Eisenberger NI, Borofsky LA, Pfeifer JH, McNealy K, Mazziotta JC, et al. Neural correlates of social exclusion during adolescence: Understanding the distress of peer rejection. *Social Cognitive and Affective Neuroscience*, 4(2): 143–157, 2009.

Norman KA, Polyn SM, Detre GJ, and Haxby JV. Beyond mind reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10: 424–430, 2006.

Pedregosa F, Varoquaux G, Michel V, Thirion B, Grisel O, Blondel M, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Pizzagalli DA, Lehmann D, Hendrick AM, Regard M, Pascual-Marqui RD, and Davidson RJ. Affective judgments of faces modulate early activity (approximately 160 ms) within the fusiform gyri. *NeuroImage*, 16: 663–677, 2002.

Preuschoff K, Quartz SR, and Bossaerts P. Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience*, 28(11): 2745–2752, 2008.

Puce A and Perrett D. Electrophysiology and brain imaging of biological motion. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 358(1431): 435–445, 2003.

Saxe R, Xiao DK, Kovacs G, Perrett DI, and Kanwisher N. A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia*, 42(11): 1435–1446, 2004.

Sebastian CL, Tan GCY, Roiser JP, Viding E, Dumontheil I, and Blakemore S-J. Developmental influences on the neural bases of responses to social rejection: Implications of social neuroscience for education. *NeuroImage*, 57(3): 686–694, 2011.

Seeley WW. Anterior insula degeneration in frontotemporal dementia. *Brain Structure and Function*, 214(5–6): 465–475, 2010.

Seeley WW, Menon V, Schatzberg AF, Keller J, Glover GH, Kenna H, et al. Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, 27(9): 2349–2356, 2007.

Singer T, Kiebel SJ, Winston JS, Dolan RJ, and Frith CD. Brain responses to the acquired moral status of faces. *Neuron*, 41(4): 653–662, 2004a.

Singer T, Seymour B, O'Doherty J, Kaube H, Dolan RJ, and Frith CD. Empathy for pain involves the affective but not sensory components of pain. *Science*, 303(5661): 1157–1162, 2004b.

Tottenham N, Tanaka JW, Leon AC, McCarry T, Nurse M, Hare TA, et al. The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, 168(3): 242–249, 2009.

Van Essen DC. A population-average, landmark- and surface-based (PALS) atlas of the human cerebral cortex. *NeuroImage*, 28: 635–662, 2005.

Vuilleumier P and Driver J. Modulation of visual processing by attention and emotion: Windows on causal interactions between human brain regions. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362(1481): 837–855, 2007.

Wicker B, Keysers C, Plailly J, Royet JP, Gallese V, and Rizzolatti G. Both of us disgusted in My insula: The common neural basis of seeing and feeling disgust. *Neuron*, 40(3): 655–664, 2003.

Williams KD, Cheung CK, and Choi W. Cyberostracism: Effects of being ignored over the Internet. *Journal of Personal and Social Psychology*, 79(5): 748–762, 2000.

Winston JS, Strange BA, O'Doherty J, and Dolan RJ. Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience*, 5(3): 277–283, 2002.

Zadro L, Williams KD, and Richardson R. How low can you go? Ostracism by a computer is sufficient to lower self-reported levels of belonging, control, selfesteem, and meaningful existence. *Journal of Experimental and Social Psychology*, 40: 560–567, 2004.