MIT Libraries | DSpace@MIT

MIT Open Access Articles

A Common Neural Code for Perceived and Inferred Emotion

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

Citation: Skerry, A. E., and R. Saxe. "A Common Neural Code for Perceived and Inferred Emotion." Journal of Neuroscience 34, no. 48 (November 26, 2014): 15997–16008.

As Published: http://dx.doi.org/10.1523/jneurosci.1676-14.2014

Publisher: Society for Neuroscience

Persistent URL: http://hdl.handle.net/1721.1/97243

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of Use: Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.

Massachusetts Institute of Technology

Behavioral/Cognitive

A Common Neural Code for Perceived and Inferred Emotion

Amy E. Skerry[1] and Rebecca Saxe[2]

[1]Department of Psychology, Harvard University, Cambridge, Massachusetts 02138, and [2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Although the emotions of other people can often be perceived from overt reactions (e.g., facial or vocal expressions), they can also be inferred from situational information in the absence of observable expressions. How does the human brain make use of these diverse forms of evidence to generate a common representation of a target's emotional state? In the present research, we identify neural patterns that correspond to emotions inferred from contextual information and find that these patterns generalize across different cues from which an emotion can be attributed. Specifically, we use functional neuroimaging to measure neural responses to dynamic facial expressions with positive and negative valence and to short animations in which the valence of a character's emotion could be identified only from the situation. Using multivoxel pattern analysis, we test for regions that contain information about the target's emotional state, identifying representations specific to a single stimulus type and representations that generalize across stimulus types. In regions of medial prefrontal cortex (MPFC), a classifier trained to discriminate emotional valence for one stimulus (e.g., animated situations) could successfully discriminate valence for the remaining stimulus (e.g., facial expressions), indicating a representation of valence that abstracts away from perceptual features and generalizes across different forms of evidence. Moreover, in a subregion of MPFC, this neural representation generalized to trials involving subjectively experienced emotional events, suggesting partial overlap in neural responses to attributed and experienced emotions. These data provide a step toward understanding how the brain transforms stimulus-bound inputs into abstract representations of emotion.

Key words: **abstraction; concepts; emotion attribution; multimodal; social cognition; theory of mind**

Introduction

To recognize someone's emotion, we can rely on facial expression, tone of voice, and even body posture. Perceiving emotions from these overt expressions poses a version of the "invariance problem" faced across perceptual domains (Ullman, 1998; DiCarlo et al., 2012): we recognize emotions despite variation both within modality (e.g., sad face across viewpoint and identity) and across modalities (e.g., sadness from facial and vocal expressions). Emotion recognition may therefore rely on bottom-up extraction of invariants within a hierarchy of increasingly complex feature-detectors (Tanaka, 1993). However, we can also infer emotions in the absence of overt expressions by reasoning about the situation a person encounters (Ortony, 1990; Zaki et al., 2008; Scherer and Meuleman, 2013). To do so, we rely on abstract causal principles (e.g., social rejection causes sadness) rather than direct perceptual cues. Ultimately, the brain must integrate these diverse sources of information into a common code that supports empathic responses and flexible emotion-based inference.

Received April 25, 2014; revised Sept. 18, 2014; accepted Sept. 24, 2014.

Author contributions: A.E.S. and R.S. designed research; A.E.S. and R.S. performed research; A.E.S. and R.S. analyzed data; A.E.S. and R.S. wrote the paper.

This work was supported by National Science Foundation Graduate Research Fellowship (A.E.S.) and NIH Grant 1R01 MH096914-01A1 (R.S.). We thank Laura Schulz, Nancy Kanwisher, Michael Cohen, Dorit Kliemann, Stefano Anzellotti, and Jorie Koster-Hale for helpful comments and discussion.

The authors declare no competing financial interests.

Correspondence should be addressed to Amy E. Skerry, William James Hall, 33 Kirkland Street, Cambridge, MA 02138. E-mail: amy.skerry@gmail.com.

DOI:10.1523/JNEUROSCI.1676-14.2014

Copyright © 2014 the authors 0270-6474/14/3315997-12$15.00/0

What neural mechanisms underlie these different aspects of emotion recognition? Previous neuroimaging studies have revealed regions containing information about emotions in overt expressions: different facial expressions, for example, elicit distinct patterns of neural activity in the superior temporal sulcus and fusiform gyrus (Said et al., 2010a,b; Harry et al., 2013; see also Pitcher, 2014). In these studies, emotional stimuli were presented in a single modality, leaving it unclear the precise dimensions represented in these regions. Given that facial expressions can be distinguished based on features specific to the visual modality (e.g., mouth motion, eyebrow deflection, eye aperture; Ekman and Rosenberg, 1997; Oosterhof and Todorov, 2009), face-responsive visual regions could distinguish emotional expressions based on such lower-level features.

To represent what is in common across sad faces and voices, the brain may also compute multimodal representations. In a recent study (Peelen et al., 2010), subjects were presented with overt facial, bodily, and vocal expressions: in posterior temporal cortex (lpSTC) and middle medial prefrontal cortex (MMPFC), the pattern of response across different modalities was more similar for the same emotion than for different emotions. Thus, emotional stimuli sharing no low-level perceptual features seem to be represented similarly in these regions.

However, we not only recognize emotions from canonical perceptual cues, but also infer emotions from causal context alone. We identify emotions in the absence of familiar expressions, even for situations we have never observed or experienced. In the present study, we test for neural representations of emotional valence that generalize across both overt facial expressions

and emotions inferred from the situation a character is in. We first identify neural patterns that contain information about emotional valence for each type of stimulus. We then test whether these neural patterns generalize across the two stimulus types, the signature of a common code integrating these very different types of emotional information. Finally, we investigate whether attributing emotional experiences to others and experiencing one's own emotions recruit a common neural representation by testing whether these same neural patterns generalize to emotional events experienced by participants themselves.

Materials and Methods

Summary

In Experiment 1, we used functional magnetic resonance imaging (fMRI) to measure blood oxygen level-dependent (BOLD) responses to emotional facial expressions and to animations depicting a character in an emotion-eliciting situation. While emotion-specific representations could, in principle, take the form of a uniform response across voxels in a region (detectable with univariate analyses), prior research has yielded little evidence for consistent and selective associations between discrete brain regions and specific emotions (Fusar-Poli et al., 2009; Lindquist et al., 2012). Thus, the present research uses multivariate analyses that exploit reliable signal across distributed patterns of voxels to uncover neural representations at a spatial scale smaller than that of entire regions (Haxby et al., 2001; Kamitani and Tong, 2005; Kriegeskorte et al., 2006; Norman et al., 2006). With this approach, we test for representations of emotional valence that are specific to a particular type of stimulus (facial expressions or causal situations) and representations that generalize across the two stimulus types. To identify stimulus-independent representations, we trained a pattern classification algorithm to discriminate emotional valence for one stimulus type (e.g., dynamic facial expressions) and tested its ability to discriminate valence for the remaining type (e.g., animations depicting causal situations). Thus, for each region of interest (ROI), we test whether there is a reliable neural pattern that supports classifying emotions when trained and tested on facial expressions, when trained and tested on situations, and when requiring generalization across facial expressions and situations.

We then test whether attributing emotions to others engages neural mechanisms involved in the first-person experience of emotion. Previous research has implicated MPFC not only in emotion attribution, but also in subjective experience of emotional or rewarding outcomes (Lin et al., 2012; Clithero and Rangel, 2013; Winecoff et al., 2013; Chikazoe et al., 2014). However, the relationship between experienced reward and emotion attribution remains poorly understood. In Experiment 2, we measured BOLD responses to positive and negative situations for another individual (replicating Experiment 1) and to trials in which subjects themselves experienced positive and negative outcomes (winning and losing money). Again, we test whether there is a reliable neural pattern that supports classifying the valence of events when trained and tested on third-party situations, when trained and tested on first-person rewards, and when requiring generalization across third-person and first-person experiences.

Regions of interest

Based on prior literature (Peelen et al., 2010), our regions of interest for abstract, conceptual representations of emotion were the pSTC and MMPFC. We localized in individual subjects a middle MPFC ROI comparable with that of Peelen et al. (2010), using a standard social versus nonsocial contrast (Saxe and Kanwisher, 2003; Dodell-Feder et al., 2011; see below). Because pSTC could not be identified by standard localizer tasks, we identified bilateral group ROIs based on the peak coordinate from Peelen et al. (2010). Our primary analyses target these three ROIs, accounting for multiple comparisons with a corrected $\alpha = 0.05/3$ (0.017).

In addition to the MMPFC region identified by Peelen et al. (2010), adjacent regions of dorsal and ventral MPFC have been strongly implicated in studies of emotion and affective value (Amodio and Frith, 2006; Hynes et al., 2006; Völlm et al., 2006; Etkin et al. 2011). Moreover, the MPFC is part of a larger set of regions [the posterior cingulate/precuneus (PC), bilateral temporal parietal junction (rTPJ and lTPJ), and right anterior temporal lobe (rATL)] that are reliably recruited when reasoning about others' mental states (Saxe and Kanwisher, 2003; Mitchell, 2009), including emotional states (Zaki et al., 2010; Bruneau et al., 2012; Spunt and Lieberman, 2012). This set of six regions [dorsal MPFC (DMPFC), ventral MPFC (VMPFC), rTPJ, lTPJ, PC, and rATL, in addition to MMPFC described above) was identified in individual subjects using the social versus nonsocial contrast (described below). We test these remaining regions for representations of both perceived and inferred emotions [with $\alpha = 0.05/6$ (0.008) to correct for comparisons across these six ROIs].

To test for modality-specific representations, we localized regions that might contain information specific to overt facial expressions: the right middle superior temporal sulcus (rmSTS), hypothesized to code for facial motion parameters (Pelphrey et al., 2005; Calder et al., 2007; Carlin et al., 2011), and face-selective patches in right occipitotemporal cortex thought to code for identity-relevant face features [occipital face area (rOFA) and fusiform face area (rFFA); Kanwisher and Yovel, 2006]. For this analysis, we again correct for multiple comparisons using $\alpha = 0.017$ (0.05/3).

Finally, in Experiment 2, we examined how the mechanisms involved in third-person attribution of emotional states relate to mechanisms involved in processing first-person subjective value. To do so, we identified a region of orbitofrontal cortex (OFC/VMPFC) that has been previously implicated in processing reward/emotional value (Kable and Glimcher, 2007; Plassmann et al., 2007; Chib et al., 2009; Winecoff et al., 2013; Chikazoe et al., 2014). We used a mask derived from two recent meta-analyses (Bartra et al., 2013; Clithero and Rangel, 2013) to investigate neural responses in an anatomical region of OFC/VMPFC in which neural responses have been shown to consistently correlate with reward value across reward types and decision contexts (anatomical mask available at http://www.rnl.caltech.edu/resources/index.html). Note that this mask is only partially overlapping with the search space used to identify VMPFC responses to theory of mind (in Experiment 1).

Participants

Twenty-one right-handed participants (20–43 years; $M_{age} = 26.84$; 14 male) were recruited for Experiment 1. Sixteen right-handed participants (19–40 years; $M_{age} = 27.88$; seven male) were recruited for Experiment 2. All participants had normal or corrected-to-normal vision and no history of neurological or psychiatric disorders and gave written, informed consent in accordance with the requirements of the MIT institutional review board.

fMRI tasks and stimuli

In Experiment 1, each subject participated in several behavioral tasks as well as three fMRI tasks: an Emotion Attribution task and two tasks used to localize regions involved in theory of mind and face perception. Subjects in Experiment 2 completed only the Emotion Attribution task and the theory of mind localizer.

Emotion Attribution task. In the Emotion Attribution task (Fig. 1), subjects viewed brief video clips designed to elicit the attribution of an emotional state to a target (Fig. 1 depicts static photos similar to video clips used in the study). The task consisted of video clips of people expressing a positive (happy/smiling) or negative (sad/frowning) emotion (expressions condition) and brief animations in which a simple geometric character experienced an event that would elicit positive or negative emotion (situations condition). In the situations condition, no emotion was expressed, but the character's emotional state could be inferred based on the character's goals and the event outcome. To ensure consistent attributions of emotional valence, independent subjects on Amazon Mechanical Turk ($n = 16$ per item) rated the stimuli from 1 to 7 (negative to positive valence): $M(SEM)_{pos\text{-}faces} = 5.597(0.077)$; $M(SEM)_{neg\text{-}faces} = 2.694(0.084)$; $M(SEM)_{pos\text{-}situations} = 5.401(0.068)$; $M(SEM)_{neg\text{-}situations} = 2.695(0.058)$. Each stimulus type was further divided into two subcategories: "male" and "female" for facial expression clips and "social" and "nonsocial" for situation clips. In the nonsocial condition, the character demonstrated an instrumental goal and achieved or failed to achieve it

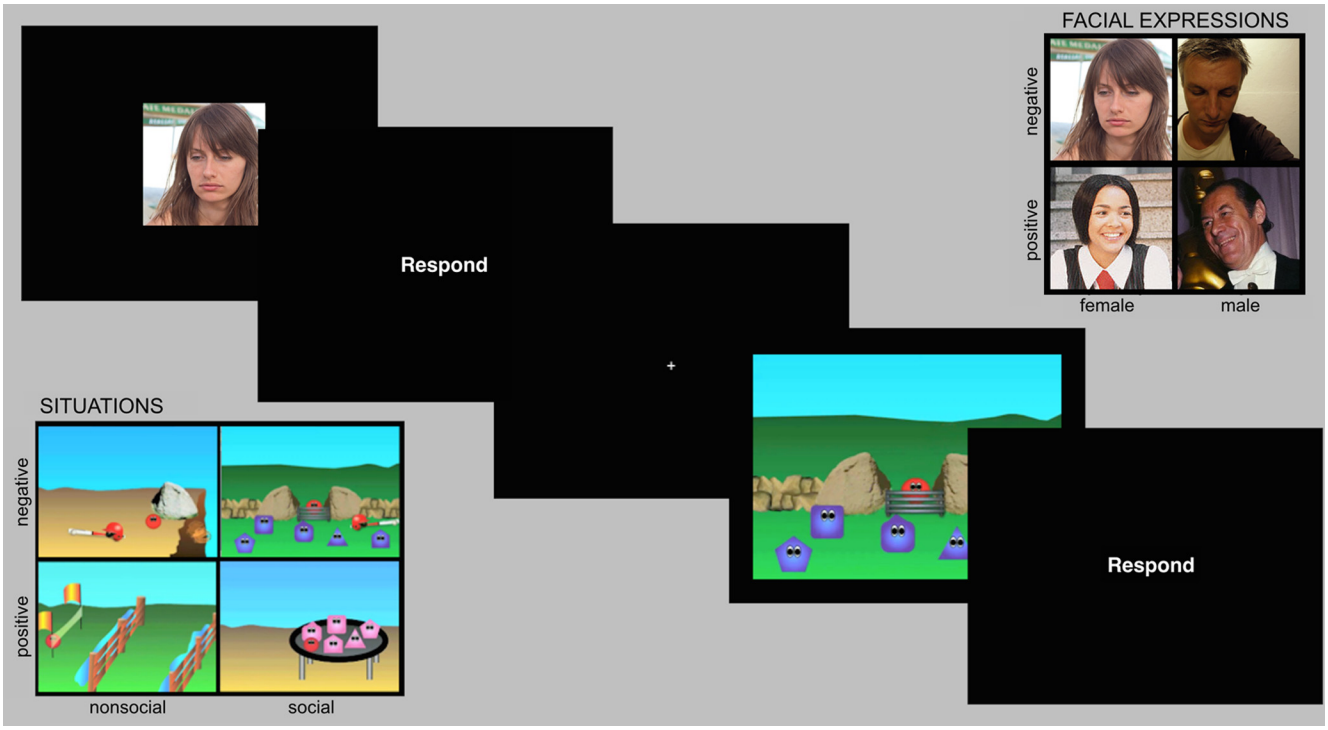

Figure 1. Task structure for Emotion Attribution task. Events consisted of a 4 s clip and a 2 s response. Stimuli included two stimulus types (situation stimuli and facial expression stimuli) and two valence categories (positive and negative valence).

(e.g., attempted to climb a hill and succeeded or tumbled to the bottom); in the social condition, there were multiple agents who acted prosocially or antisocially to the target character (e.g., included or excluded the target from their group). This yielded a total of eight stimulus conditions (male positive, male negative, female positive, female negative, social positive, social negative, nonsocial positive, nonsocial negative). Because the face stimuli involved a close-perspective view on a single entity, these stimuli were presented at 7.8 × 7.4° visual angle, whereas the context animations were presented at 16.7 × 12.5°. We used dynamic, naturalistic facial expressions from movies, which are relatively uncontrolled compared with artificial stimuli (e.g., face morphs). However, our main interest is in representations that generalize to animations in the situations condition; low-level visual confounds that generalize across the two perceptually distinct stimulus sets are, therefore, highly unlikely. An advantage of these stimuli in the present design is that they achieve an unusual a balance between external validity (Zaki and Ochsner, 2009; Spunt and Lieberman, 2012) and experimental control.

The experiment consisted of eight runs (9.43 min/run), each containing 6 stimuli in each of the eight conditions, for a total of 48 stimuli per condition. Each condition contained 24 semantically distinct events, each of which was presented twice over the course of the experiment with superficial transformations (the background scene for context animations and a minor luminance change for facial expressions), and the left–right orientation varied across the two presentations. Each clip was presented at fixation for 4 s, followed by a 1750 ms window during which subjects made a behavioral response and a 250 ms blank screen. Subjects were instructed to press a button to indicate the intensity of the character's emotion in each event (1 to 4, neutral to extreme), which focused subjects' attention on the character's emotional state but ensured that motor responses (intensity) were orthogonal to the discrimination of interest (valence). The clips were presented in a jittered, event-related design, and a central fixation cross was presented between trials with a variable interstimulus interval of 0–14 s. Optseq2 (http://surfer.nmr.mgh.harvard.edu/optseq/) was used to create efficient stimulus presentation schedules with a first-order counterbalancing constraint such that each condition preceded each other with approximately equal probability across the experiment. The assignment of conditions to positions within this sequence was randomized across participants. The order of individual stimulus clips for a given condition was chosen pseudorandomly for each participant, with the constraint that repetitions of each stimulus occurred in the same even–odd folds as the first presentation (e.g., an event first presented in run 2 would be repeated in run 6, and an event presented in run 3 would be repeated in run 7).

In Experiment 2, subjects completed a modified and abbreviated version of this task (four runs). On 50% of trials, subjects viewed nonsocial situation stimuli from Experiment 1 (96 total trials); on the remaining trials, subjects were presented with positive and negative events in which they either gained or lost money from a postscan bonus (reward condition; Fig. 2). On each reward trial, subjects viewed a cycle of 20 rapidly presented random monetary values (2 s total), followed by the reward outcome for the trial, shown in green (2 s). Negative values ranged from −$0.20 to −$1.00, and positive values ranged from +$0.20 to +$2.00; this asymmetry allowed subjects to have net gain for their bonus and accounted for the fact that losses are experienced more strongly than comparable gains (Tversky and Kahneman, 1991). The experimental design and behavioral task were identical to Experiment 1, except that subjects were asked to rate the character's emotional intensity on the situation trials and their own emotional intensity on the reward trials.

Theory of mind localizer. Subjects were presented with short textual scenarios that required inferences about mental state representations (Belief condition) or physical representations such as a map, photo, or painting (Photo condition; Dodell-Feder et al., 2011; stimuli are available at http://saxelab.mit.edu/superloc.php). These two types of scenarios were similar in their meta-representational demands and logical complexity, but only the scenarios in the Belief condition required building a representation of another person's thoughts and beliefs. Scenarios were displayed for 10 s, followed immediately by a true or false question (4 s) about either the representation (Belief or Photo) or the reality of the situation. Each run (4.53 min) consisted of 10 trials separated by 12 s interstimulus intervals, and 12 s blocks of fixation were included at the beginning and end of each run. One to two runs were presented to each participant. The order of stimulus type (Belief or Photo) and correct answer (True or False) were counterbalanced within and across runs.

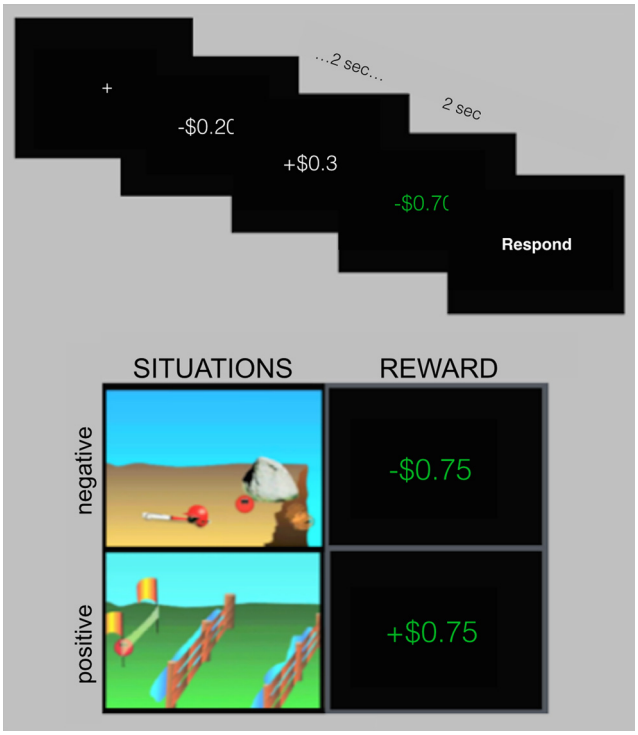

Figure 2. Task structure for Experiment 2. Events consisted of a 4 s trial and 2 s response. Stimuli included two stimulus types (situation stimuli and reward stimuli) and two valence categories (positive and negative valence). Reward trials involved 2 s of rapid cycling through random values, followed by 2 s during which the reward outcome was displayed.

Face perception localizer. Subjects viewed two conditions designed to identify face-selective regions: dynamic faces (video clips of human children's faces) and dynamic objects (video clips of objects in motion; from Pitcher et al., 2011). For each of these conditions, there were a total of 30 clips (3 s each, separated by 333 ms of blank screen), and six clips were presented in each block. This localizer also included two other conditions, biological motion and structure from motion, which were not of interest for the present analyses. All conditions were presented as 20 s blocks followed by 2 s of rest, and 12 s blocks of fixation were included at the beginning and end of each run, as well as once in the middle of the run. Each condition was presented twice per run, and subjects received two runs lasting 5 min each, with condition order counterbalanced within and across runs and across participants. To maintain attention, subjects were required to complete a one-back task during viewing. Two of 21 subjects did not complete this localizer because of insufficient scan time.

Behavioral tasks. The Autism-Spectrum Quotient (Baron-Cohen et al., 2001) and the Interpersonal Reactivity Index (Davis, 1983) were completed via on-line Qualtrics surveys. Participants also completed an Empathic Accuracy task based on the study by Zaki et al. (2008) and the verbal reasoning, matrices, and riddles components of the KBIT2 (Kaufman, 1990).

Acquisition

Data were acquired on a 3T Siemens Tim Trio scanner in the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, using a Siemens 32-channel phased array head coil. We collected a high-resolution (1 mm isotropic) T1-weighted MPRAGE anatomical scan, followed by functional images acquired with a gradient-echo EPI sequence sensitive to BOLD contrast [repetition time (TR), 2 s; echo time, 30 ms; flip angle, 90°; voxel size, 3 × 3 × 3 mm; matrix 64 × 64; 32 axial slices]. Slices were aligned with the anterior/posterior commissure and provided whole-brain coverage (excluding the cerebellum).

Analysis

Pilot data. In addition to the 21 subjects reported, 8 independent pilot subjects were analyzed to fix the parameters of the analyses reported below (e.g., size of smoothing kernel, type of classifier, method for feature selection). A general concern with fMRI analyses, and with the application of machine learning techniques to fMRI data in particular, is that the space of possible and reasonable analyses is large and can yield qualitatively different results. Analysis decisions should be made independent of the comparisons or tests of interest; otherwise, one risks overfitting the analysis to the data (Simmons et al., 2011). One way to optimize an analysis stream without such overfitting is to separate subjects into an exploratory or pilot set and a validation or test set. Thus, the analysis stream reported here was selected based on the parameters that appeared to yield the most sensitive analysis of eight pilot subjects.

Preprocessing. MRI data were preprocessed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/), FreeSurfer (http://surfer.nmr.mgh.harvard.edu/), and in-house code. FreeSurfer's skull-stripping software was used for brain extraction. SPM was used to motion correct each subject's data via rigid rotation and translation about the six orthogonal axes of motion, to register the functional data to the subject's high-resolution anatomical image, and to normalize the data onto a common brain space (Montreal Neurological Institute). In addition to the smoothing imposed by normalization, functional images were smoothed using a Gaussian filter (FWHM, 5 mm).

Defining regions of interest. To define individual ROIs, we used hypothesis spaces derived from random-effects analyses of previous studies [theory of mind (Dufour et al., 2013): bilateral TPJ, rATL, PC, subregions of MPFC (DMPFC, MMPFC, VMPFC); face perception (Julian et al., 2012): rmSTS, rFFA, rOFA], combined with individual subject activations for the localizer tasks. The theory of mind task was modeled as a 14 s boxcar (the full length of the story and question period, shifted by 1 TR to account for lag in reading, comprehension, and processing of comprehended text) convolved with a standard hemodynamic response function (HRF). A general linear model was implemented in SPM8 to estimate β values for Belief trials and Photo trials. We conducted high-pass filtering at 128 Hz, normalized the global mean signal, and included nuisance covariates to remove effects of run. The face perception task was modeled as a 22 s boxcar, and β values were similarly estimated for each of condition (dynamic faces, dynamic objects, biological motion, structure from motion). For each subject, we used a one-sample t test implemented in SPM8 to generate a map of t values for the relevant contrast (Belief > Photo for the theory of mind ROIs, faces > objects for the face perception ROIs), and for each ROI, we identified the peak t value within the hypothesis space. An individual subject's ROI was defined as the cluster of contiguous suprathreshold voxels (minimum $k = 10$) within a 9 mm sphere surrounding this peak. If no cluster was found at $p < 0.001$, we repeated this procedure at $p < 0.01$ and $p < 0.05$. We masked each ROI by its hypothesis space (defined to be mutually exclusive) such that there was no overlap in the voxels contained in each functionally defined ROI. An ROI for a given subject was required to have at least 20 voxels to be included in multivariate analyses. For the pSTC region (Peelen et al., 2010), we generated a group ROI defined as a 9 mm sphere around the peak coordinate from that study, as well as an analogous ROI for the right hemisphere.

Multivariate analyses. Multivoxel pattern analysis (MVPA) was conducted using an in-house code developed in Python using the publicly available PyMVPA toolbox (http://www.pymvpa.org/; Fig. 3). We conducted MVPA within ROIs that were functionally defined based on individual subject localizer scans. High-pass filtering (128 Hz) was conducted on each run, and linear detrending was performed across the whole time course. A time point was excluded if it was a global intensity outlier (>3 SD above the mean intensity) or corresponded to a large movement (>2 mm scan to scan). The data were temporally compressed to generate one voxel-wise summary for each individual trial, and these single trial summaries were used for both training and testing. Individual trial patterns were calculated by averaging the preprocessed bold images for the 6 s duration of the trial, offset by 4 s to account for HRF lag. Rest time points were removed, and the trial summaries were concatenated into one experimental vector in which each value was a trial's average response. The pattern for each trial was then z-scored relative to the mean across all trial responses in that voxel.

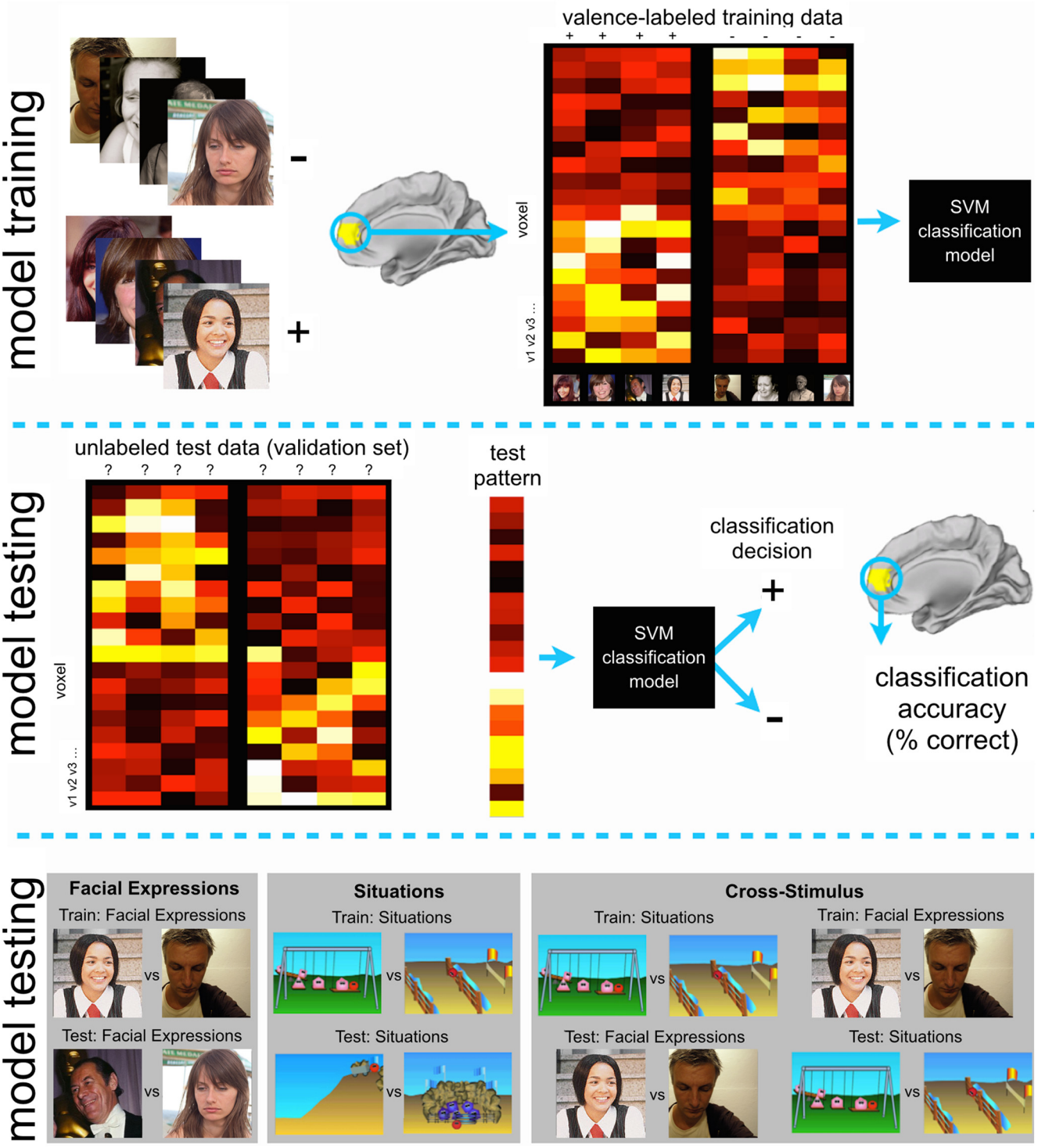

Figure 3. MVPA analysis procedure. Top, Valence-labeled voxel patterns (from a single ROI) used to train a linear support vector machine (SVM). Middle, Learned voxel weights used to predict valence of unlabeled test data (voxel patterns not used for training). Bottom, Cross-validation schemes for testing for stimulus-specific and stimulus-independent emotion representations.

Given the high dimensionality of fMRI data and the relatively small number of training examples available, feature selection is often useful to extract voxels likely to be informative for classification (Mitchell et al., 2004; De Martino et al., 2008; Pereira et al., 2009). Within each ROI, we conducted voxel-wise ANOVAs to identify voxels that were modulated by the task (based on the F statistic for task vs rest contrast). This univariate selection procedure tends to eliminate high-variance, noisy voxels (Mitchell et al., 2004). Because this selection procedure is orthogonal to all of the classifications reported here, it could be performed once over the whole dataset without constituting peeking, meaning that the same voxels could be used as features in each cross-validation fold. The top 80 most active voxels within the ROI were used for classification (selecting a fixed number of voxels also helps to minimize differences in the number of voxels across regions and subjects).

The data were classified using a support vector machine implemented with libSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/; Chang and Lin, 2011). This classifier uses condition-labeled training data to learn a weight for each voxel, and subsequent stimuli (validation data not used

for model training) can then be assigned to one of two classes based on a weighted linear combination of the response in each voxel. In a support vector machine, the linear decision function can be thought of as a hyperplane dividing the multidimensional voxel space into two classes, and voxel weights are learned so as to maximize the distance between the hyperplane and the closest observed example. We conducted binary classification with a linear kernel using a fixed regularization parameter ($C = 1$) to control the tradeoff between margin size and training error. We restricted ourselves to linearly decodable signal under the assumption that a linear kernel implements a plausible readout mechanism for downstream neurons (Seung and Sompolinsky, 1993; Hung et al., 2005; Shamir and Sompolinsky, 2006). Given that the brain likely implements nonlinear transformations, linear separability within a population can be thought of as a conservative but reasonable estimate of the information available for explicit readout (DiCarlo and Cox, 2007).

For each classification, the data were partitioned into multiple cross-validation folds where the classifier was trained iteratively on all folds but one and tested on the remaining fold. Classification accuracy was then averaged across folds to yield a single classification accuracy for each subject in the ROI. A one-sample *t* test was then performed over these individual accuracies, comparing with chance classification of 0.50 (all *t* tests on classification accuracies were one-tailed). Whereas parametric tests are not always appropriate for assessing the significance of classification accuracies (Stelzer et al., 2013), the assumptions of these tests are met in the present case: the accuracy values are independent samples from separate subjects (rather than individual folds trained on overlapping data), and the classification accuracies were found to be normally distributed around the mean accuracy. For within-stimulus analyses (classifying within facial expressions and within situation stimuli), cross-validation was performed across runs (i.e., iteratively train on seven runs, test on the remaining eighth). For cross-stimulus analyses, the folds for cross-validation were based on stimulus type. To ensure complete independence between training and test data, folds for the cross-stimulus analysis were also divided based on even versus odd runs (e.g., train on even run facial expressions, test on odd run situations).

Whole-brain searchlight classification. The searchlight procedure was identical to the ROI-based procedure except that the classifier was applied to voxels within searchlight spheres rather than individually localized ROIs. For each voxel in a gray matter mask, we defined a sphere containing all voxels within a three-voxel radius of the center voxel. The searchlight size (123 voxels) was selected to approximately match the size of the regions in which effects were found with the ROI analysis, and we again conducted an ANOVA to select the 80 most active voxels in the sphere. Classification was then performed on each cross-validation fold, and the average classification accuracy for each sphere was assigned to its central voxel, yielding a single accuracy image for each subject for a given discrimination. We then conducted a one-sample *t* test over subjects' accuracy maps, comparing accuracy in each voxel to chance (0.5). This yielded a group *t*-map, which was assessed at a $p < 0.05$, FWE corrected (based on SPM's implementation of Gaussian random fields).

Whole-brain random-effects analysis (univariate). We also conducted a whole-brain random effects analysis to identify voxels in which the univariate response differentiated positive and negative valence for faces and for situations. The conjunction of these two contrasts would identify voxels in which the magnitude of response was related to the valence for both stimulus types.

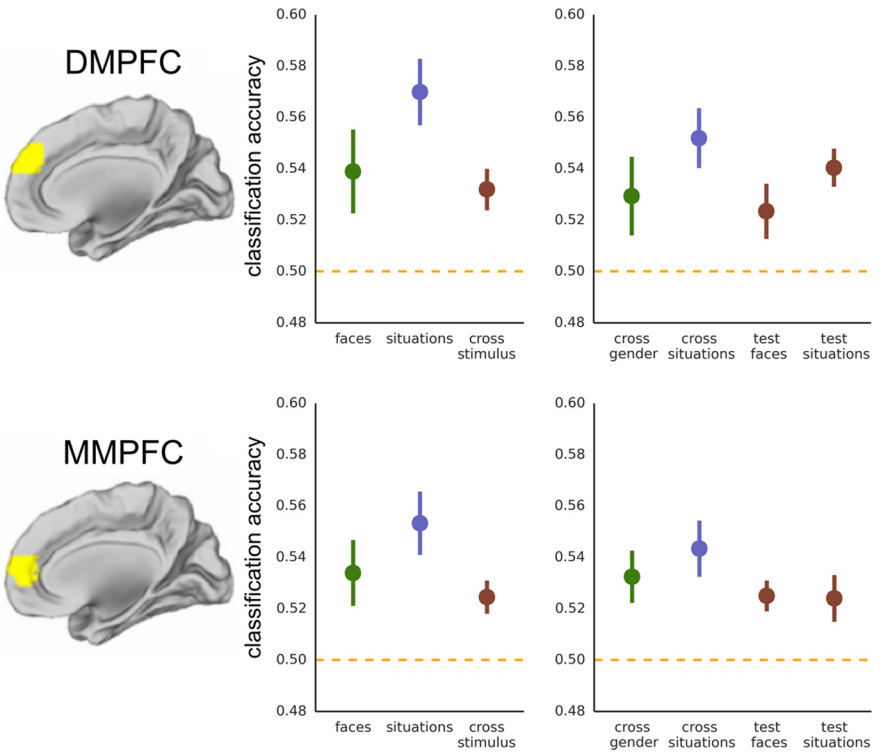

Figure 4. DMPFC/MMPFC: Experiment 1. Classification accuracy for facial expressions (green), for situation stimuli (blue), and when training and testing across stimulus types (red). Cross-stimulus accuracies are the average of accuracies for train facial expression/test situation and train situation/test facial expression. Chance equals 0.50.

Results

Experiment 1

Regions of interest

Using the contrast of Belief > Photo, we identified seven ROIs (rTPJ, lTPJ, rATL, PC, DMPFC, MMPFC, VMPFC) in each of the 21 subjects, and using the contrast of faces > objects, we identified right lateralized face regions OFA, FFA, and mSTS in 18 subjects (of 19 subjects who completed this localizer).

Multivariate results

Multimodal regions (pSTC and MMPFC). For classification of emotional valence for facial expressions, we replicated the results of Peelen et al. (2010) with above-chance classification in MMPFC [M(SEM) = 0.534(0.013), $t_{(18)} = 2.65$, $p = 0.008$; Fig. 4] and lpSTC [M(SEM) = 0.525(0.010), $t_{(20)} = 2.61$, $p = 0.008$; Fig. 5]. Classification in right posterior superior temporal cortex (rpSTC) did not reach significance at a corrected (0.05/3) threshold [M(SEM) = 0.516(0.007), $t_{(20)} = 2.23$, $p = 0.019$]. Note that although the magnitude of these effects is small, these results reflect classification of single-event trials, which are strongly influenced by measurement noise. Small but significant classification accuracies are common for single-trial, within-category distinctions (Anzellotti et al., 2013; Harry et al., 2013).

The key question for the present research is whether these regions contain neural codes specific to overt expressions or whether they also represent the valence of inferred emotional states. When classifying valence for situation stimuli, we again found above-chance classification accuracy in MMPFC [M(SEM) = 0.553(0.012), $t_{(18)} = 4.31$, $p > 0.001$]. We then tested for stimulus-independent representations by training on one kind

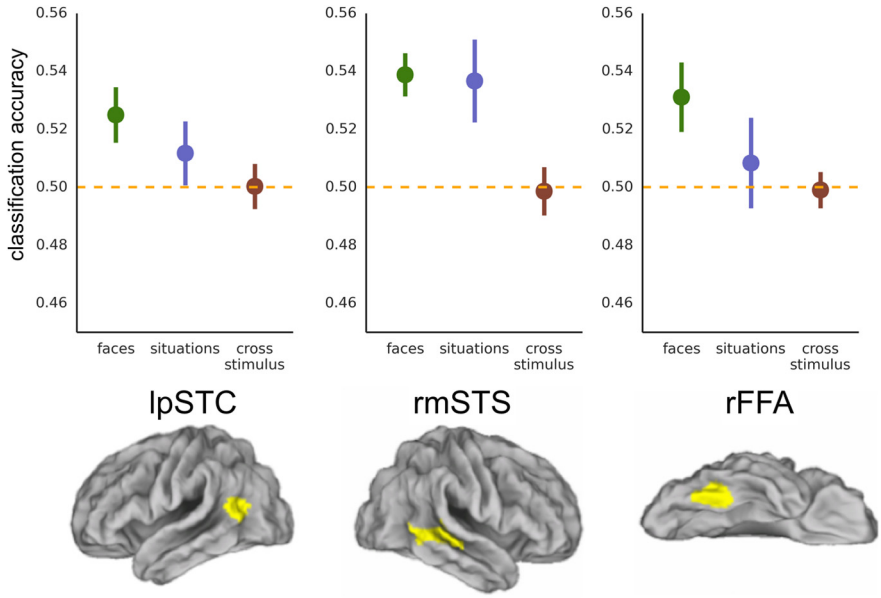

Figure 5. Classification accuracy for facial expressions (green), for situation stimuli (blue), and when training and testing across stimulus types (red). Cross-stimulus accuracies are the average of accuracies for train facial expression/test situation and train situation/test facial expression. Chance equals 0.50.

of stimulus and testing on the other. Consistent with the existence of an abstract valence code, MMPFC supported above-chance valence classification across both stimulus types [M(SEM) = 0.524(0.007), $t_{(18)} = 3.77$, $p = 0.001$]. In contrast, lpSTC did not perform above chance when classifying the valence of situation stimuli [M(SEM) = 0.512(0.011), $t_{(20)} = 1.06$, $p = 0.152$], nor when requiring generalization across stimulus type [M(SEM) = 0.500(0.008), $t_{(20)} = 0.04$, $p = 0.486$]. To directly compare accuracy in lpSTC when classifying within facial expression stimuli and when generalizing across stimulus types, we conducted a paired sample t test (one-tailed) comparing classification accuracy for faces to accuracy for cross-stimulus classification: classification accuracy was significantly higher for faces compared with cross-stimulus classification ($M = 0.525$, $M = 0.500$, $t_{(20)} = 2.00$, $p = 0.029$).

Theory of mind regions. We performed these same analyses in six remaining theory of mind regions (at a corrected $\alpha = 0.05/6, 0.008$). In DMPFC (Fig. 4), we observed results very comparable with those observed in MMPFC: above-chance classification of facial emotion [M(SEM) = 0.539(0.016), $t_{(18)} = 2.39$, $p = 0.014$], of emotion from situations [M(SEM) = 0.570(0.013), $t_{(18)} = 5.38$, $p < 0.001$], and when generalizing across stimulus types [M(SEM) = 0.532(0.008), $t_{(18)} = 3.95$, $p < 0.001$]. VMPFC did not perform above chance at a corrected threshold ($p < 0.008$) when classifying facial expressions [M(SEM) = 0.525(0.009), $t_{(17)} = 2.62$, $p = 0.009$] or situation stimuli [M(SEM) = 0.524(0.012), $t_{(17)} = 1.98$, $p = 0.032$]; however, cross-stimulus decoding was above chance [M(SEM) = 0.527(0.007), $t_{(17)} = 3.79$, $p = 0.001$].

None of the other theory of mind regions classified above threshold when distinguishing positive and negative facial expressions [rTPJ: M(SEM) = 0.501(0.010), $t_{(20)} = 0.06$, $p = 0.478$; lTPJ: M(SEM) = 0.521(0.012), $t_{(20)} = 1.85$, $p = 0.040$; rATL: M(SEM) = 0.525(0.012), $t_{(20)} = 2.05$, $p = 0.027$; PC: M(SEM) = 0.514(0.011), $t_{(20)} = 1.32$, $p = 0.102$], when distinguishing positive and negative situations [rTPJ: M(SEM) = 0.528(0.014), $t_{(20)} = 2.04$, $p = 0.027$; lTPJ: M(SEM) = 0.515(0.009), $t_{(20)} = 1.57$, $p = 0.066$; rATL: M(SEM) = 0.510(0.012), $t_{(20)} = 0.80$, $p = 0.216$; PC: M(SEM) = 0.523(0.012), $t_{(20)} = 1.84$, $p = 0.040$], or when generalizing across stimulus types [rTPJ: M(SEM) = 0.503(0.007), $t_{(20)} = 0.45$, $p = 0.330$; lTPJ: M(SEM) = 0.509(0.007), $t_{(20)} = 1.38$, $p = 0.092$; rATL: M(SEM) = 0.510(0.006), $t_{(20)} = 1.85$, $p = 0.039$; PC: M(SEM) = 0.495(0.008), $t_{(20)} = -0.60$, $p = 0.724$].

Face-selective cortex. For valence in facial expressions, we also performed a secondary analysis in face-selective regions rOFA, rFFA, and rmSTS (at a corrected threshold of 0.05/3; Fig. 5). We replicated previous reports (Said et al., 2010a,b; Furl et al., 2012; Harry et al., 2013) with classification accuracies significantly above chance in rmSTS [M(SEM) = 0.539(0.007), $t_{(14)} = 5.20$, $p < 0.001$] and in rFFA [M(SEM) = 0.531(0.012), $t_{(14)} = 2.59$, $p = 0.011$]; classification in rOFA did not survive correction for multiple comparisons [M(SEM) = 0.529(0.016), $t_{(13)} = 1.87$, $p = 0.042$]. For the situation stimuli, the rFFA failed to classify valence when it was inferred from context [rFFA: M(SEM) = 0.508(0.016), $t_{(14)} = 0.54$, $p = 0.300$]. In the rmSTS, on the other hand, there was reliable information about situation stimuli in addition to the face stimuli [M(SEM) = 0.537(0.014), $t_{(14)} = 2.57$, $p = 0.011$]. However, neither region supported above-chance cross-stimulus classification [rFFA: M(SEM) = 0.499(0.006), $t_{(14)} = -0.16$, $p = 0.563$; rmSTS: M(SEM) = 0.499(0.008), $t_{(14)} = -0.17$, $p = 0.565$], and classification accuracy was reliably higher (one-tailed test) when training and testing on faces compared with when requiring generalization across stimulus types in rmSTS ($M = 0.539$, $M = 0.499$, $t_{(14)} = 4.52$, $p < 0.001$) and in rFFA ($M = 0.531$, $M = 0.499$, $t_{(14)} = 2.26$, $p = 0.020$).

Follow-up analyses

Given successful valence decoding in dorsal and middle MPFC, we conducted several follow-up analyses to examine the scope and generality of these effects. For facial expressions, we performed cross-validation across the orthogonal dimension of face gender. Both regions of MPFC performed above chance [DMPFC: M(SEM) = 0.529(0.015), $t_{(18)} = 1.92$, $p = 0.035$; MMPFC: M(SEM) = 0.532(0.010), $t_{(18)} = 3.20$, $p = 0.003$], indicating that the valence-specific voxel patterns generalize across two face sets that differed at the level of exemplars, identity, and gender. We also tested for generalization across face sets in the remaining regions that supported decoding of facial expressions (rmSTS, rFFA, lpSTC). The neural patterns generalized across the male and female face sets in rmSTS [M(SEM) = 0.524(0.012), $t_{(14)} = 2.02$, $p = 0.032$] but not in rFFA [M(SEM) = 0.512(0.012), $t_{(14)} = 1.00$, $p = 0.167$] or lpSTC [M(SEM) = 0.509(0.009), $t_{(20)} = 1.05$, $p = 0.154$].

For situation stimuli, both regions of MPFC were able to classify valence across the orthogonal dimension: social versus nonsocial situations [DMPFC: M(SEM) = 0.552(0.012), $t_{(18)} = 4.44$, $p < 0.001$; MMPFC: M(SEM) = 0.543(0.011), $t_{(18)} = 3.97$, $p < 0.001$]. Finally, to test for possible asymmetry in the cross-stimulus classification, we separated the cross-stimulus analysis into training on faces/testing on situations and training on situations/testing on faces. We observed above-chance classification for both train/test partitions in both DMPFC [testing on faces:

Table 1. Whole brain, Experiment 1: Searchlight results ($p < 0.05$, FWE corrected)

Stimulus	Number of voxels	Peak *t*	*x*	*y*	*z*	Region
Situations	52	11.80	4	46	38	DMPFC
		8.24	6	50	28	
	9	9.49	−8	54	26	DMPFC
	28	9.21	4	58	14	MMPFC
		9.02	4	56	22	
	1	7.98	16	60	24	MMPFC
	1	7.86	0	50	36	DMPFC
	4	7.82	0	54	30	DMPFC
	1	7.55	−8	54	18	MMPFC
	2	7.43	8	56	20	MMPFC
	1	7.40	−2	54	36	DMPFC
	1	7.30	−28	−78	32	L OCC/TEMP
Faces	8	8.77	−30	−88	−4	L MID OCC GYRUS
	2	8.48	38	−92	8	R MID OCC GYRUS
	3	8.16	2	52	20	MMPFC
	1	7.88	6	52	22	MMPFC
	2	7.60	8	56	20	MMPFC
	1	7.52	28	−82	32	R SUP OCC
Cross-stimulus	42	10.91	−2	50	34	DMPFC
		9.28	0	48	24	
		7.28	8	56	20	
	1	8.93	8	56	10	MMPFC
	1	7.34	12	66	10	MMPFC

L OCC/TEMP, Left occipital/temporal; L MID OCC GYRUS, left middle occipital gyrus, R MID OCC GYRUS, right middle occipital gyrus; R SUP OCC, right superior occipital.

M(SEM) = 0.523(0.011), $t_{(18)} = 2.18$, $p = 0.021$; testing on situations: M(SEM) = 0.540(0.007), $t_{(18)} = 5.47$, $p < 0.001$] and MMPFC [testing on faces: M(SEM) = 0.525(0.006), $t_{(18)} = 4.13$, $p < 0.001$; testing on situations: M(SEM) = 0.524(0.009), $t_{(18)} = 2.64$, $p = 0.008$].

In summary, it appears that dorsal and middle subregions of MPFC contain reliable information about the emotional valence of a stimulus when the emotion must be inferred from the situation and that the neural code in this region is highly abstract, generalizing across diverse cues from which an emotion can be identified. In contrast, although both rFFA and the region of superior temporal cortex identified by Peelen et al. (2010) contain information about the valence of facial expressions, the neural codes in those regions do not appear generalized to valence representations formed on the basis of contextual information. Interestingly, the rmSTS appears to contain information about valence in faces and situations but does not form a common code that integrates across stimulus type.

Whole-brain analyses

To test for any remaining regions that may contain information about the emotional valence of these stimuli, we conducted a searchlight procedure, revealing striking consistency with the ROI analysis (Table 1; Fig. 6). Only DMPFC and MMPFC exhibited above-chance classification for faces and contexts, and when generalizing across these two stimulus types. In addition, for classification of facial expressions alone, we observed clusters in occipital cortex. Clusters in the other ROIs emerged at a more liberal threshold (rOFA and rmSTS at $p < 0.001$ uncorrected; rFFA, rpSTC, and lpSTC at $p < 0.01$). In contrast, whole-brain analyses of the univariate response revealed no regions in which the mean response distinguished between positive and negative facial expressions or between positive and negative contexts (at $p < 0.05$, FWE correction based on Gaussian random fields).

Experiment 2

The results of Experiment 1 suggest that DMPFC and MMPFC contain abstract, stimulus-independent information about emotional valence of perceived and inferred emotions. How is this region related to the regions of MPFC typically implicated in processing value and/or subjective experience? For Experiment 2, we first used a group anatomical mask (Bartra et al., 2013; Clithero and Rangel, 2013) to identify a region of OFC/VMPFC previously implicated in reward/value processing. Consistent with previous reports (Kable and Glimcher, 2007; Chib et al., 2009), this region showed an overall magnitude effect for positive > negative rewards ($t_{(15)} = 3.20$, $p = 0.006$; Fig. 7) and could classify positive versus negative reward trials reliably above chance [M(SEM) = 0.542(0.020), $t_{(15)} = 2.09$, $p = 0.027$]. Interestingly, this canonical reward region did not reliably distinguish positive and negative situations for others [M(SEM) = 0.521(0.018), $t_{(15)} = 1.15$, $p = 0.135$], and there was no evidence for a common valence code generalizing across self and other [M(SEM) = 0.512(0.014), $t_{(15)} = 0.80$, $p = 0.219$]. Classification accuracies were significantly higher when discriminating self-reward values compared with when generalizing across reward and situation trials ($M = 0.542$, $M = 0.512$, $t_{(15)} = 1.90$, $p = 0.038$, one-tailed).

What about the regions implicated in abstract valence representation in Experiment 1? By decoding valence within the situation stimuli, we replicate the finding of Experiment 1 that DMPFC and MMPFC contain information about the emotion attributed to a target even when that emotion must be inferred from context [DMPFC: M(SEM) = 0.543(0.021), $t_{(15)} = 2.04$, $p = 0.030$; MMPFC: M(SEM) = 0.536(0.019), $t_{(15)} = 1.95$, $p = 0.035$; Fig. 8]. Do we observe these same neural patterns on trials in which subjects evaluate their own subjectively experienced emotions? In MMPFC, we observed above-chance valence classification for reward trials [M(SEM) = 0.539(0.018), $t_{(15)} = 2.17$, $p = 0.023$] in addition to situation trials. Moreover, neural patterns generalized across positive/negative situations and positive/negative outcomes for the self [M(SEM) = 0.526(0.010), $t_{(15)} = 2.60$, $p = 0.010$]. In dorsal MPFC, in contrast, we observed similar classification of the valence of reward outcomes [M(SEM) = 0.544(0.025), $t_{(15)} = 1.74$, $p = 0.051$], but this region failed to classify above chance when generalizing across self and other [M(SEM) = 0.514(0.013), $t_{(15)} = 1.07$, $p = 0.150$].

Discussion

Are there neural representations of emotions that generalize across diverse sources of evidence, including overt emotional expressions and emotions inferred from context alone? In the present study, we identified regions in which voxel-wise response patterns contained information about the emotional valence of facial expressions and a smaller number of regions that distinguished the valence of emotion-eliciting situations. Our results, together with existing literature (Peelen et al., 2010), provide candidate neural substrates for three levels of representation: modality-specific representations bound to perceptual invariants in the input, intermediate multimodal representations that generalize across canonical perceptual schemas, and conceptual representations that are fully invariant to the information used to identify emotions.

Conceptual representations

In DMPFC/MMPFC, we decoded emotional valence from facial expressions and from animations depicting emotion-eliciting situations. Like other domains of high-level cognition, emotion knowledge is theory like (Carey, 1985; Gopnik and Wellman, 1992), requiring abstract concepts (e.g., of goals, expectations) to be integrated in a coherent, causal manner. The present results

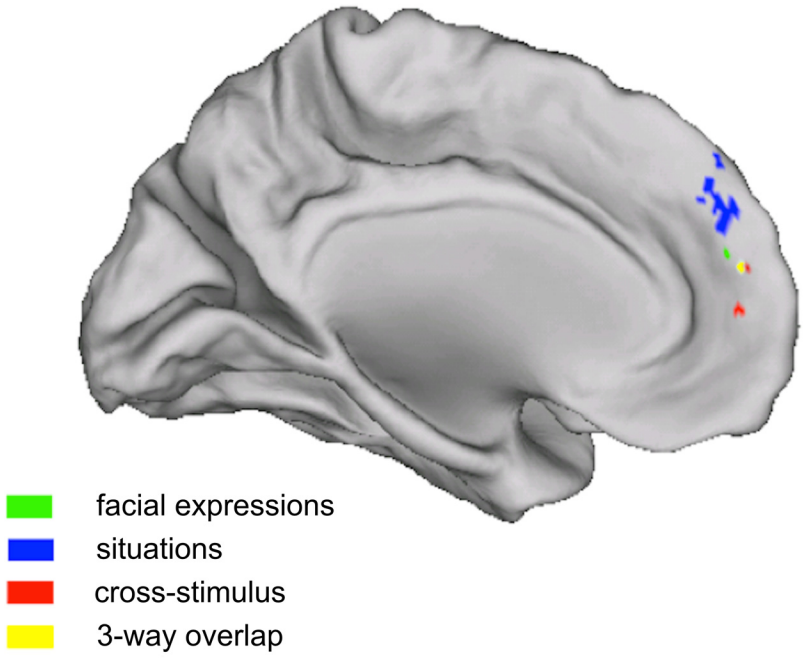

Figure 6. Whole brain: Experiment 1. Classification in whole-brain searchlight (sphere radius, 3 voxels). $p < 0.05$ (FWE corrected using Gaussian random fields).

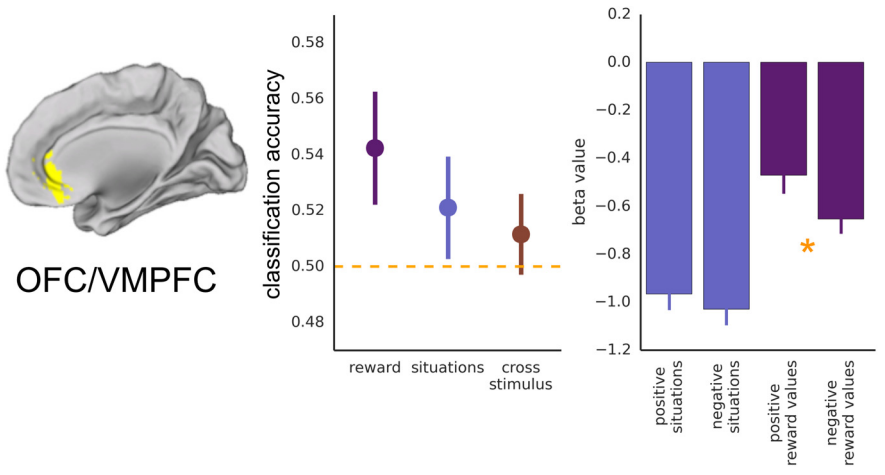

Figure 7. OFC/VMPFC. Results from anatomical OFC/VMPFC reward ROI (Bartra et al., 2013; Clithero and Rangel, 2013). Left, Classification accuracy for reward outcomes (purple), for situation stimuli (blue), and when training and testing across stimulus types (red). Chance equals 0.50. Right, Mean β values in the ROI for each stimulus condition, asterisk indicates significant difference ($p < 0.05$).

suggest that valence representations in DMPFC/MMPFC are elicited by such inferential processes. We could classify valence when training on faces and testing on situations (and vice versa), replicating the finding that emotion representations in MMPFC generalize across perceptually dissimilar stimuli (Peelen et al., 2010). Moreover, our results demonstrate an even stronger form of generalization: perceived emotions and emotions inferred through generative, theory-like processes activate similar neural patterns in DMPFC/MMPFC, indicating a mechanism beyond mere association of co-occurring perceptual schemas. Thus, the MPFC may contain a common neural code that integrates diverse perceptual and inferential processes to form abstract representations of emotions.

Previous research leaves open the question of whether activity in MPFC reflects mechanisms specific to emotion attribution or mechanisms involved in value or valence processing more generally. In Experiment 2, we found evidence for both kinds of representations. First, we found that the region of OFC/VMPFC implicated in reward processing (Clithero and Rangel, 2013; anatomical ROI from Bartra et al., 2013) does not contain information about the valence of attributed emotions. Second, we found no evidence for a shared representation of experienced and attributed emotion in dorsal MPFC. Finally, in MMPFC, we observed neural patterns that generalized across attributed and experienced emotional events. One interpretation of this result is that attributing positive or rewarding experiences to others depends on general purpose reward representations that code value in social and nonsocial contexts (Chib et al., 2009; Lin et al., 2012, Ruff and Fehr, 2014). Alternatively, neural responses in MMPFC could reflect the participant's own empathic reaction to the depicted experiences (e.g., witnessing someone achieve a goal elicits positive emotions in participants). If so, the participant's empathic reaction might be causally involved in the process of attributing emotions to others (consistent with "simulation theory"; Goldman and Sripada, 2005; Niedenthal, 2007) or might be a downstream consequence of attribution. Previous results do indicate a causal role for MPFC in emotion perception and attribution: damage to MPFC is associated with deficits in emotion recognition (Shamay-Tsoory et al., 2003, 2009), and direct disruption of MPFC via transcranial magnetic stimulation has been shown to impair recognition of facial expressions (Harmer et al., 2001; see also Mattavelli et al., 2011). Moreover, the degree to which MPFC is recruited during an emotion attribution task predicts individual differences in the accuracy of emotion judgments (Zaki et al., 2009a,b). Future research should continue to distinguish the specific contents of attributed emotions from the emotional response of the participant. For example, can patterns in MPFC be used to classify the attribution of more specific emotions that are unlikely to be shared by the observer (e.g., loneliness vs regret)?

Modality-specific representations

In face-selective regions (rFFA and rmSTS), we found that neural patterns could distinguish positive and negative facial expressions, replicating previous reports of emotion-specific neural representations in these regions (Fox et al., 2009; Said et al., 2010a,b; Xu and Biederman, 2010; Furl et al., 2012; Harry et al., 2013). Neural populations could distinguish facial expressions by responding to relatively low-level parameters that differ across expressions, by extracting mid-level invari-

ants (e.g., eye motion, mouth configuration) that generalize across within-modality transformations (e.g., lighting, position), or by computing explicit representations of facial emotion that integrate multiple facial parameters. The present study used naturalistic stimuli that varied in lighting conditions, face direction, and face position and found reliable generalization across male and female face sets in rmSTS. Thus, it is possible that these neural patterns distinguish facial expressions based on representations invariant to certain low-level transformations (Anzellotti et al., 2013). Future research should investigate this possibility by systematically testing the generalization properties of neural responses to emotional expressions across variation in low-level dimensions (e.g., face direction) and higher-level dimensions (e.g., generalization from sad eyes to a sad mouth). Interestingly, the rmSTS also contained information about emotional valence in situation stimuli, but the neural patterns did not generalize across these distinct sources of evidence, suggesting two independent valence codes in this region.

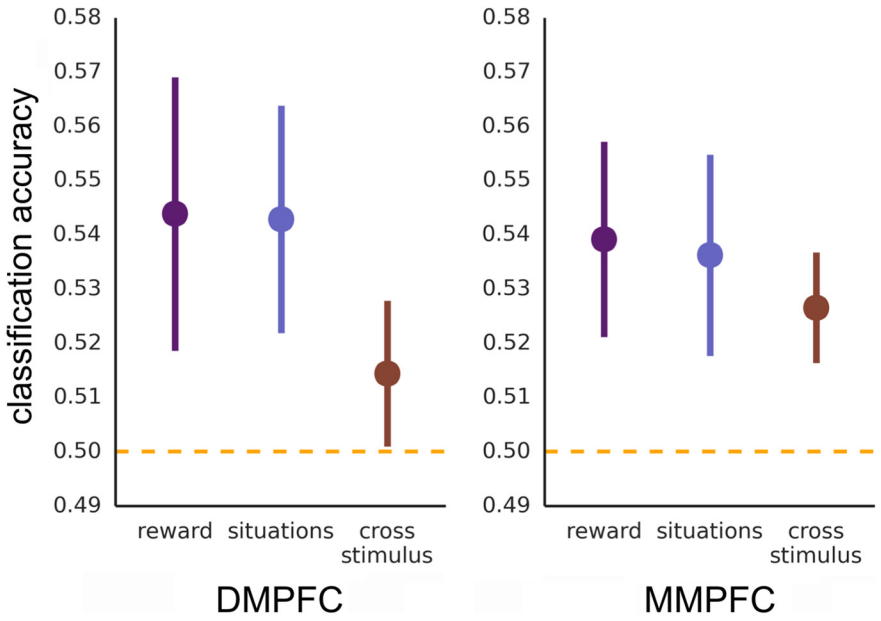

Figure 8. MPFC: Experiment 2. Classification accuracy for reward outcomes (purple), for situation stimuli (blue), and when training and testing across stimulus types (red). Cross-stimulus accuracies are the average of accuracies for train reward/test situation and train situation/test reward. Chance equals 0.50.

Multimodal representations

We also replicate the finding that pSTC contains information about the emotional valence of facial expressions (Peelen et al., 2010). However, unlike DMPFC/MMPFC, we find no evidence for representations of emotions inferred from situations. Interestingly, Peelen et al. (2010) found that the pSTC could decode emotional expressions across modalities (faces, bodies, voices), suggesting that this region may support an intermediate representation that is neither fully conceptual nor tied to specific perceptual parameters. For example, pSTC could be involved in pooling over associated perceptual schemas, leading to representations that generalize across diverse sensory inputs but do not extend to more abstract, inference-based representations. This interpretation would be consistent with the region's proposed role in cross-modal integration (Kreifelts et al., 2009; Stevenson and James, 2009). Thus, the present findings reveal a novel functional division within the set of regions (pSTC and MMPFC) previously implicated in multimodal emotion representation (Peelen et al., 2010).

Open questions

While these data provide important constraints on the levels of representation associated with different regions, important questions remain open. First, do the regions identified here contain information about more fine-grained emotional distinctions beyond valence? Previous studies have successfully decoded a larger space of perceived emotions in MMPFC, STS, and FFA (Peelen et al., 2010; Said et al., 2010a,b; Harry et al., 2013). For emotions inferred from context, the neural representation of more fine-grained emotional distinctions (e.g., inferring sadness vs fear) will be a key question for future research.

This study also leaves open the role of other regions (e.g., amygdala, insula, inferior frontal gyrus) that have previously been associated with emotion perception and experience (Shamay-Tsoory et al., 2009; Singer et al., 2009; Pessoa and Adolphs, 2010). What is the precise content of emotion representations in these regions, and do they contribute to identifying specific emotional states in others? With the searchlight procedure, we found little evidence for representations of emotional valence outside the a priori ROIs. However, whole-brain analyses are less sensitive than ROI analyses, and although multivariate analyses alleviate some of the spatial constraints of univariate methods, they still tend to rely on relatively low-frequency information (Op de Beeck, 2010; Freeman et al., 2011), meaning that MVPA provides a lower bound on the information available in a given region (Kriegeskorte and Kievit, 2013). Neurophysiological studies (Gothard et al., 2007; Hadj-Bouziane et al., 2012) may help to elucidate the full set of regions contributing to emotion attribution.

Relatedly, how does information in these different regions interact during the process of attribution? A tempting speculation is that the regions described here make up a hierarchy of information flow (Adolphs, 2002; Ethofer et al., 2006; e.g., modality-specific, face-selective cortex ⇔ multimodal pSTC ⇔ conceptual MPFC). However, additional connectivity or causal information (Friston et al., 2003; Bestmann et al., 2008) would be required to confirm such an account and to directly map different representational content onto discrete stages.

Finally, these findings are complementary to previous investigations of semantic representations [e.g., object categories (Devereux et al., 2013; Fairhall and Caramazza, 2013)], which have identified modality-specific representations (e.g., in visual cortex) and representations that generalize across modalities (e.g., across words and pictures in left middle temporal gyrus). The present findings highlight a distinction between representations that are multimodal and those that are based on theory-like causal inferences. Does this distinction apply to other domains,

and can it help to clarify the neural organization of abstract knowledge more broadly?

General conclusions

The challenge of emotion recognition demands neural processes for exploiting different sources of evidence for others' emotions, as well as a common code for integrating this information to support emotion-based inference. Here, we demonstrate successful decoding of valence for emotional states that must be inferred from context as well as emotions directly perceived from overt expressions. By testing the scope and generality of the responses in different regions, we provide important constraints on possible computational roles of these regions and begin to elucidate the series of representations that make up the processing stream for emotional perception, attribution, and empathy. Thus, the present research provides a step toward understanding how the brain transforms stimulus-bound inputs into abstract representations of emotions.

References

Adolphs R (2002) Neural systems for recognizing emotion. Curr Opin Neurobiol 12:169–177. CrossRef Medline

Amodio DM, Frith CD (2006) Meeting of minds: the medial frontal cortex and social cognition. Nat Rev Neurosci 7:268–277. CrossRef Medline

Anzellotti S, Fairhall SL, Caramazza A (2013) Decoding representations of face identity that are tolerant to rotation. Cereb Cortex 24:1988–1995. CrossRef Medline

Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E (2001) The Autism-Spectrum Quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. J Autism Dev Disord 31:5–17. CrossRef Medline

Bartra O, McGuire JT, Kable JW (2013) The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. Neuroimage 76:412–427. CrossRef Medline

Bestmann S, Ruff CC, Blankenburg F, Weiskopf N, Driver J, Rothwell JC (2008) Mapping causal interregional influences with concurrent TMS–fMRI. Exp Brain Res 191:383–402. CrossRef Medline

Bruneau EG, Pluta A, Saxe R (2012) Distinct roles of the "shared pain" and "theory of mind" networks in processing others' emotional suffering. Neuropsychologia 50:219–231. CrossRef Medline

Calder AJ, Beaver JD, Winston JS, Dolan RJ, Jenkins R, Eger E, Henson RN (2007) Separate coding of different gaze directions in the superior temporal sulcus and inferior parietal lobule. Curr Biol 17:20–25. CrossRef Medline

Carey S (1985) Conceptual change in childhood. Cambridge, MA: MIT.

Carlin JD, Calder AJ, Kriegeskorte N, Nili H, Rowe JB (2011) A head view-invariant representation of gaze direction in anterior superior temporal sulcus. Curr Biol 21:1817–1821. CrossRef Medline

Chang C, Lin C (2011) LIBSVM: a library for support vector machines. ACM Trans Intellig Sys Tech 2:1–27.

Chib VS, Rangel A, Shimojo S, O'Doherty JP (2009) Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. J Neurosci 29:12315–12320. CrossRef Medline

Chikazoe J, Lee DH, Kriegeskorte N, Anderson AK (2014) Population coding of affect across stimuli, modalities and individuals. Nat Neurosci 17:1114–1122. CrossRef Medline

Clithero JA, Rangel A (2013) Informatic parcellation of the network involved in the computation of subjective value. Soc Cogn Affect Neurosci 9:1289–1302. CrossRef Medline

Davis MH (1983) Measuring individual differences in empathy: evidence for a multidimensional approach. J Pers Soc Psychol 44:113–126. CrossRef

De Martino F, Valente G, Staeren N, Ashburner J, Goebel R, Formisano E (2008) Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. Neuroimage 43:44–58. CrossRef Medline

Devereux BJ, Clarke A, Marouchos A, Tyler LK (2013) Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. J Neurosci 33:18906–18916. CrossRef Medline

DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. Trends Cogn Sci 11:333–341. CrossRef Medline

DiCarlo JJ, Zoccolan D, Rust NC (2012) How does the brain solve visual object recognition? Neuron 73:415–434. CrossRef Medline

Dodell-Feder D, Koster-Hale J, Bedny M, Saxe R (2011) fMRI item analysis in a theory of mind task. Neuroimage 55:705–712. CrossRef Medline

Dufour N, Redcay E, Young L, Mavros PL, Moran JM, Triantafyllou C, Gabrieli JD, Saxe R (2013) Similar brain activation during false belief tasks in a large sample of adults with and without autism. PLoS One 8:e75468. CrossRef Medline

Ekman P, Rosenberg EL (1997) What the face reveals: basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford: Oxford UP.

Ethofer T, Anders S, Erb M, Herbert C, Wiethoff S, Kissler J, Grodd W, Wildgruber D (2006) Cerebral pathways in processing of affective prosody: a dynamic causal modeling study. Neuroimage 30:580–587. CrossRef Medline

Etkin A, Egner T, Kalisch R (2011) Emotional processing in anterior cingulate and medial prefrontal cortex. Trends Cogn Sci 15:85–93. CrossRef Medline

Fairhall SL, Caramazza A (2013) Brain regions that represent amodal conceptual knowledge. J Neurosci 33:10552–10558. CrossRef Medline

Fox CJ, Moon SY, Iaria G, Barton JJ (2009) The correlates of subjective perception of identity and expression in the face network: an fMRI adaptation study. Neuroimage 44:569–580. CrossRef Medline

Freeman J, Brouwer GJ, Heeger DJ, Merriam EP (2011) Orientation decoding depends on maps, not columns. J Neurosci 31:4792–4804. CrossRef Medline

Friston KJ, Harrison L, Penny W (2003) Dynamic causal modelling. Neuroimage 19:1273–1302. CrossRef Medline

Furl N, Hadj-Bouziane F, Liu N, Averbeck BB, Ungerleider LG (2012) Dynamic and static facial expressions decoded from motion-sensitive areas in the macaque monkey. J Neurosci 32:15952–15962. CrossRef Medline

Fusar-Poli P, Placentino A, Carletti F, Landi P, Allen P, Surguladze S, Benedetti F, Abbamonte M, Gasparotti R, Barale F, Perez J, McGuire P, Politi P (2009) Functional atlas of emotional faces processing: a voxel-based meta-analysis of 105 functional magnetic resonance imaging studies. J Psychiatry Neurosci 34:418–432. Medline

Goldman AI, Sripada CS (2005) Simulationist models of face-based emotion recognition. Cognition 94:193–213. CrossRef Medline

Gopnik A, Wellman HM (1992) Why the child's theory of mind really is a theory. Mind Lang 7:145–171. CrossRef

Gothard KM, Battaglia FP, Erickson CA, Spitler KM, Amaral DG (2007) Neural responses to facial expression and face identity in the monkey amygdala. J Neurophysiol 97:1671–1683. CrossRef Medline

Hadj-Bouziane F, Liu N, Bell AH, Gothard KM, Luh WM, Tootell RB, Murray EA, Ungerleider LG (2012) Amygdala lesions disrupt modulation of functional MRI activity evoked by facial expression in the monkey inferior temporal cortex. Proc Natl Acad Sci U S A 109:E3640–E3648. CrossRef Medline

Harmer CJ, Thilo KV, Rothwell JC, Goodwin GM (2001) Transcranial magnetic stimulation of medial–frontal cortex impairs the processing of angry facial expressions. Nat Neurosci 4:17–18. CrossRef Medline

Harry B, Williams MA, Davis C, Kim J (2013) Emotional expressions evoke a differential response in the fusiform face area. Front Hum Neurosci 7:692. CrossRef Medline

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425–2430. CrossRef Medline

Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005) Fast readout of object identity from macaque inferior temporal cortex. Science 310:863–866. CrossRef Medline

Hynes CA, Baird AA, Grafton ST (2006) Differential role of the orbital frontal lobe in emotional versus cognitive perspective-taking. Neuropsychologia 44:374–383. CrossRef Medline

Julian JB, Fedorenko E, Webster J, Kanwisher N (2012) An algorithmic method for functionally defining regions of interest in the ventral visual pathway. Neuroimage 60:2357–2364. CrossRef Medline

Kable JW, Glimcher PW (2007) The neural correlates of subjective value during intertemporal choice. Nat Neurosci 10:1625–1633. CrossRef Medline

Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. Nat Neurosci 8:679–685. CrossRef Medline

Kanwisher N, Yovel G (2006) The fusiform face area: a cortical region specialized for the perception of faces. Philos Trans R Soc Lond B Biol Sci 361:2109–2128. CrossRef Medline

Kaufman A (1990) KBIT-2: Kaufman Brief Intelligence Test. Bloomington, MN: NCS Pearson.

Kreifelts B, Ethofer T, Shiozawa T, Grodd W, Wildgruber D (2009) Cerebral representation of non-verbal emotional perception: fMRI reveals audiovisual integration area between voice- and face-sensitive regions in the superior temporal sulcus. Neuropsychologia 47:3059–3066. CrossRef Medline

Kriegeskorte N, Kievit RA (2013) Representational geometry: integrating cognition, computation, and the brain. Trends Cogn Sci 17:401–412. CrossRef Medline

Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. Proc Natl Acad Sci U S A 103:3863–3868. CrossRef Medline

Lin A, Adolphs R, Rangel A (2012) Social and monetary reward learning engage overlapping neural substrates. Soc Cogn Affect Neurosci 7:274–281. CrossRef Medline

Lindquist KA, Wager TD, Kober H, Bliss-Moreau E, Barrett LF (2012) The brain basis of emotion: a meta-analytic review. Behav Brain Sci 35:121–143. CrossRef Medline

Mattavelli G, Cattaneo Z, Papagno C (2011) Transcranial magnetic stimulation of medial prefrontal cortex modulates face expressions processing in a priming task. Neuropsychologia 49:992–998. CrossRef Medline

Mitchell JP (2009) Inferences about mental states. Philos Trans R Soc Lond B Biol Sci 364:1309–1316. CrossRef Medline

Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang X, Just M, Newman S (2004) Learning to decode cognitive states from brain images. Mach Learn 57:145–175. CrossRef

Niedenthal PM (2007) Embodying emotion. Science 316:1002–1005. CrossRef Medline

Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci 10:424–430. CrossRef Medline

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

Op de Beeck HP (2010) Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? Neuroimage 49:1943–1948. CrossRef Medline

Ortony A (1990) Reactions to Events I. In: The cognitive structure of emotions. p 228. Cambridge, UK: Cambridge UP.

Peelen MV, Atkinson AP, Vuilleumier P (2010) Supramodal representations of perceived emotions in the human brain. J Neurosci 30:10127–10134. CrossRef Medline

Pelphrey KA, Morris JP, Michelich CR, Allison T, McCarthy G (2005) Functional anatomy of biological motion perception in posterior temporal cortex: an fMRI study of eye, mouth and hand movements. Cereb Cortex 15:1866–1876. CrossRef Medline

Pereira F, Mitchell T, Botvinick M (2009) Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45:S199–S209. CrossRef Medline

Pessoa L, Adolphs R (2010) Emotion processing and the amygdala: from a "low road" to "many roads" of evaluating biological significance. Nat Rev Neurosci 11:773–783. CrossRef Medline

Pitcher D (2014) Facial expression recognition takes longer in the posterior superior temporal sulcus than in the occipital face area. J Neurosci 34:9173–9177. CrossRef Medline

Pitcher D, Dilks DD, Saxe RR, Triantagyllou C, Kanwisher N (2011) Differential selectivity for dynamic versus static information in face selective cortical regions. Neuroimage 56:2356–2363. CrossRef Medline

Plassmann H, O'Doherty J, Rangel A (2007) Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. J Neurosci 27:9984–9988. CrossRef Medline

Ruff CC, Fehr E (2014) The neurobiology of rewards and values in social decision making. Nat Rev Neurosci 15:549–562. CrossRef Medline

Said CP, Moore CD, Engell AD, Todorov A, Haxby JV (2010a) Distributed representations of dynamic facial expressions in the superior temporal sulcus. J Vis 10:11. CrossRef Medline

Said CP, Moore CD, Norman KA, Haxby JV, Todorov A (2010b) Graded representations of emotional expressions in the left superior temporal sulcus. Front Syst Neurosci 4:6. CrossRef Medline

Saxe R, Kanwisher N (2003) People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind." Neuroimage 19:1835–1842. CrossRef

Scherer KR, Meuleman B (2013) Human emotion experiences can be predicted on theoretical grounds: evidence from verbal labeling. PLoS One 8:e58166. CrossRef Medline

Seung HS, Sompolinsky H (1993) Simple models for reading neuronal population codes. Proc Natl Acad Sci U S A 90:10749–10753. CrossRef Medline

Shamay-Tsoory SG, Tomer R, Berger BD, Aharon-Peretz J (2003) Characterization of empathy deficits following prefrontal brain damage: the role of the right ventromedial prefrontal cortex. J Cogn Neurosci 15:324–337. CrossRef Medline

Shamay-Tsoory SG, Aharon-Peretz J, Perry D (2009) Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. Brain 132:617–627. CrossRef Medline

Shamir M, Sompolinsky H (2006) Implications of neuronal diversity on population coding. Neural Comput 18:1951–1986. CrossRef Medline

Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 22:1359–1366. CrossRef Medline

Singer T, Critchley HD, Preuschoff K (2009) A common role of insula in feelings, empathy and uncertainty. Trends Cogn Sci 13:334–340. CrossRef Medline

Spunt RP, Lieberman MD (2012) An integrative model of the neural systems supporting the comprehension of observed emotional behavior. Neuroimage 59:3050–3059. CrossRef Medline

Stelzer J, Chen Y, Turner R (2013) Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. Neuroimage 65:69–82. CrossRef Medline

Stevenson RA, James TW (2009) Audiovisual integration in human superior temporal sulcus: inverse effectiveness and the neural processing of speech and object recognition. Neuroimage 44:1210–1223. CrossRef Medline

Tanaka K (1993) Neuronal mechanisms of object recognition. Science 262:685–688. CrossRef Medline

Tversky A, Kahneman D (1991) Loss aversion in riskless choice: a reference-dependent model. Q J Econ 106:1039–1061. CrossRef

Ullman S (1998) Three-dimensional object recognition based on the combination of views. Cognition 67:21–44. CrossRef Medline

Völlm BA, Taylor AN, Richardson P, Corcoran R, Stirling J, McKie S, Deakin JF, Elliott R (2006) Neuronal correlates of theory of mind and empathy: a functional magnetic resonance imaging study in a nonverbal task. Neuroimage 29:90–98. CrossRef Medline

Winecoff A, Clithero JA, Carter RM, Bergman SR, Wang L, Huettel SA (2013) Ventromedial prefrontal cortex encodes emotional value. J Neurosci 33:11032–11039. CrossRef Medline

Xu X, Biederman I (2010) Loci of the release from fMRI adaptation for changes in facial expression, identity, and viewpoint. J Vis 10:36. CrossRef

Zaki J, Ochsner KN (2009) The need for a cognitive neuroscience of naturalistic social cognition. Ann N Y Acad Sci 1167:16–30. CrossRef

Zaki J, Bolger N, Ochsner K (2008) It takes two: the interpersonal nature of empathic accuracy. Psychol Sci 19:399–404. CrossRef Medline

Zaki J, Weber J, Bolger N, Ochsner K (2009a) The neural bases of empathic accuracy. Proc Natl Acad Sci U S A 106:11382–11387. CrossRef Medline

Zaki J, Bolger N, Ochsner K (2009b) Unpacking the informational bases of empathic accuracy. Emotion 9:478–487. CrossRef Medline

Zaki J, Hennigan K, Weber J, Ochsner KN (2010) Social cognitive conflict resolution: contributions of domain-general and domain-specific neural systems. J Neurosci 30:8481–8488. CrossRef Medline