Neural Representations of Emotion Are Organized around Abstract Event Features

Neural representations of emotion are organized around abstract event features

Skerry, A.E[1] and Saxe, R.[2]
1. Department of Psychology, Harvard University, Cambridge MA 02138.
2. Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge MA 02139.

Corresponding author:
Amy Skerry
amy.skerry@gmail.com
MIT Building 46-4021
77 Massachusetts Avenue
Cambridge, MA 02139

SUMMARY

Research on emotion attribution has tended to focus on the perception of overt expressions of at most 5 or 6 basic emotions. However, our ability to identify others' emotional states is not limited to perception of these canonical expressions. Instead, we make fine-grained inferences about what others feel based on the situations they encounter, relying on knowledge of the eliciting conditions for different emotions. In the present research, we provide convergent behavioral and neural evidence concerning the representations underlying these concepts. First, we find that patterns of activity in mentalizing regions contain information about subtle emotional distinctions conveyed through verbal descriptions of eliciting situations. Second, we identify a space of abstract situation features that well captures the emotion discriminations subjects make behaviorally, and show that this feature space outperforms competing models in capturing the similarity space of neural patterns in these regions. Together, the data suggest that our knowledge of others' emotions is abstract and high-dimensional, that brain regions selective for mental state reasoning support relatively subtle distinctions between emotion concepts, and that the neural representations in these regions are not reducible to more primitive affective dimensions such as valence and arousal.

INTRODUCTION

The emotional states of others can be identified by a number of cues: we can recognize what someone is feeling based on their facial expressions [1, 2], affective vocalizations [3, 5], or body posture [7, 9, 11]. However, we can also attribute subtle emotions based solely on the situation a person encounters [12], and our vocabulary for attributing these states extends beyond the small set of emotions associated with canonical emotional displays [13–15]. In many cases, surrounding context has been found to modulate or even dominate the perception of emotion from overt expressions [16–19].

While the space of emotional states perceived in faces has been studied extensively [20–22], little is known about how conceptual knowledge of others' emotions is organized, or how that knowledge is encoded in the human brain. What are the relevant features of events that allow us to make fine-grained emotional attributions (e.g. distinguishing when someone will feel angry versus disappointed, or excited versus proud) and what are the dimensions of the space by which neural populations represent emotions? Here, we argue that emotion attribution recruits a rich theory of the causal context of different emotions, and show that dimensions of this intuitive theory underlie emotion representations in brain regions associated with theory of mind.

What neural mechanisms support fine-grained emotion attributions?

As with behavioral research, studies of the neural basis of emotion attribution have generally focused on the perception of affect in facial or vocal expressions [23, 24]. For example, different facial expressions elicit discriminable patterns of activity in regions of the superior temporal sulcus (STS) and the fusiform gyrus [25–28], while emotional prosody can be decoded in secondary auditory cortex [29]. Some have argued that these overt expressions communicate a set of five or six basic emotions [1, 21, 30, 31], while other data suggest that faces and voices support even fewer universal discriminations [32–34]. In addition to regions distinguishing observable displays of emotion, recent research suggests that the medial prefrontal cortex (MPFC) contains representations of emotion that are invariant to perceptual modality [35, 36], generalizing to emotions inferred in the absence of any overt expression [37].

While these studies move beyond modality-specific perceptual representations, they nonetheless focused on relatively coarse distinctions, decoding either valence [37] or 5 basic emotions [36]. Does the MPFC also contain representations that support more fine-grained emotional discriminations? To address this question, we constructed verbal stimuli (see Table 1) describing events or situations that would elicit one of 20 different emotions (validated using 20-AFC behavioral experiment with independent subjects; see Experimental Procedures) and use multi-voxel pattern analysis [38, 39] to test which regions contain information about these subtle emotional distinctions.

As a first step, we train a classifier to distinguish the 20 emotions using distributed patterns of activity across voxels in a region, and test whether the emotion category of a new stimulus can be classified based on the pattern of neural activity it elicits. In addition to whole-brain analyses, we focus on a priori regions of interest [36, 37], the strongest candidates being subregions of MPFC—dorsal medial prefrontal cortex (DMPFC) and middle medial prefrontal cortex (MMPFC). However, the MPFC is part of a larger network of regions involved in reasoning about others' mental states [40, 41]: the posterior cingulate/precuneus (PC), bilateral temporal parietal junction (TPJ) and bilateral anterior temporal lobes (ATL). These remaining regions of the “Theory of Mind” (ToM) network have been associated with causal attribution of emotion [42–44], and thus these remaining regions of the “Theory of Mind” (ToM) network serve as additional candidate regions for fine-grained emotion concepts.

We then use Representational Similarity Analysis (RSA:[39, 45]) to characterize emotion representations in ToM brain regions, and test competing hypotheses about the features that best explain that representational space (Figure 4). RSA complements classification analyses by providing a framework for characterizing the representational structure within a region, and for testing competing models of that structure [45, 46]. While above-chance classification of different emotions would demonstrate that a particular region contains information that can differentiates emotions, classification accuracies alone reveal little about the underlying representations. In RSA, neural population codes are represented in terms of the similarity of the neural patterns elicited by different stimuli or conditions. The representational dissimilarity matrix (RDM) of the conditions in a given region can then be compared to the similarity spaces captured by different models [46, 47]. Importantly, RSA allows for comparison of hypotheses that take different forms (i.e. RDMs can be outputs of formal computational models or derived from quantitative behavioral results) and have different numbers of parameters (the correlation between model and neural RDMs is parameter free, eliminating the risk of over-fitting with more complex models).

Candidate feature spaces for emotion inference

A dominant approach in affective neuroscience has been to represent emotions as points within some low-dimensional space of more basic affective states. One possibility is that the space of emotions is built from a small set of basic emotions (e.g. happiness, sadness, fear, anger, and disgust), each associated with a prototypical facial expression, behavioral profile, and innate neural substrate [1, 48–53]. On this view, the diverse space of human emotion can be understood as combinations of these more basic emotional states [54–56]. For example, a recent study found that although human faces could convey as many as 21 discriminable emotional expressions, these emotions could be decomposed into linear combinations of 6 more primitive expressions [57].

A competing theory of emotional perception is the "circumplex" model, according to which emotions live within a core affective space composed of only two primitive dimensions: valence and arousal [22, 58–61]. Valence and arousal are argued to correspond to two innate systems that are implemented in distinct neural circuits and recruited to varying degrees across different emotions [62–66]. Thus, a second proposal is that neural representations of emotion can be reduced to a linear combination of two these neurophysiological dimensions [67].

These theories provide two competing hypotheses about the features or dimensions that structure neural representations of attributed emotion. Although many have focused on the differences between these two proposals [68, 69], they have much in common. Both approaches aim to reduce emotions to a smaller number of categories or dimensions, which are assumed to be basic affective states rooted in innate neural substrates. However, much of the empirical support for both proposals comes from studies on the perception of emotions from overt expressions [1, 60, 70], and from research on the neural correlates of perceiving or inducing different emotions [63, 65, 71, 72]. While these low-dimensional spaces successfully capture the emotions people perceive in overt expressions, they may be inadequate to account for the full variety of human emotional concepts [4, 12, 13, 73].

Here, we present subjects with the rich causal structure of eliciting situations rather than overt emotional displays. Thus, we hypothesize that the present paradigm will evoke neural representations of emotion that differ both in dimensionality and content from the models that have dominated research on perception of facial expressions.

To test this hypothesis, we consider a third feature space that represents emotions in terms of abstract features of the events that give rise to them. According to appraisal theory, emotional reactions relate systematically to people's interpretations or "appraisals" of the events around them [74–76], and there have been various proposals concerning the specific event appraisals that correspond to different emotions [77–80]. Drawing from this literature[1], we generated a set of 38 abstract event features thought to reliably vary across different emotions concepts (e.g. Did someone cause this situation intentionally or did it occur by accident? Was the person interacting with other people in this situation? Was this situation a familiar event/situation for the person? See Supplemental Experimental Procedures for full list of appraisal features).

A main goal of the present research is to test whether emotion representations in MPFC and other ToM regions can be well-explained by any of these three candidate feature spaces: the "circumplex" space defined by the judgments of valence and arousal for each stimulus, the "basic emotion" space defined by judgments of the extent to which the stimulus elicited each of 6 basic emotions (happy, sad, angry, afraid, disgusted, or surprised), and the 38-dimensional appraisal space. Importantly, the latter space differs from the other two not only in its dimensionality (38 dimensions vs. 6 or 2) but also in its content: rather than reducing the space of emotions to a smaller set of purportedly "basic" affective states, it aims to encode emotions in terms of abstract features of the causal contexts that tend to elicit them. Of course, the hypothesis that neural representations of emotion concepts are best captured by a high-dimensional space of appraisal features is not at odds with the claim that simpler dimensions like valence and arousal contribute to the organization of our emotion knowledge. For example, the 38-dimensional space contains features such as goal consistency and pleasantness that intuitively relate to the dimension of valence. The question, then, is whether the representations in regions like MPFC can be *exhausted* by one of the simpler spaces.

To create RDMs that encode each of these feature spaces, an independent group of subjects rated each verbal stimulus on a large set of features (i.e. 38 abstract event features, as well as ratings of 6 basic emotions, and of valence and arousal). These behavioral ratings were used to form three candidate representational spaces (see Figure 3), where each emotion category is captured as a feature vector within that space (see Figure S3 for examples of appraisal profiles). First, we examine which space best captures the set of emotions subjects attribute behaviorally by testing whether models trained on

[1] We drew on appraisal theory to generate event features because researchers in that tradition have been most explicit about what the relevant conceptual dimensions might be. However, our general approach is also compatible with constructivist theories [4, 6, 8, 10] in which assume that affective primitives like valence and arousal must be combined with abstract conceptual knowledge to differentiate others' emotions.

each of the relevant feature vectors for a subset of the stimuli could reliably classify the emotion label of untrained stimuli (based on stimulus-specific feature values). Do any of these feature spaces provide a sufficient basis to match performances of human subjects in discriminating these 20 emotions? Then, to test which feature space best explains the neural representation of these stimuli, we compute the similarity of conditions within each proposed feature space and compare the RDMs of candidate models to neural RDMs derived from patterns of activity across voxels in a particular region. Thus, we can test whether regions implicated in emotion discrimination are better characterized by a model containing only dimensions such as valence and arousal (the circumplex space) or 6 basic emotions, or by a representational space defined in terms of abstract event variables.

RESULTS

Classification:

In the scanner, subjects read 200 stimuli describing situations that would cause a particular emotion (see Experimental Procedures; example stimuli provided in Table 1). Subjects were instructed to consider how the target would feel in the situation, and rate the intensity of the experience for the target. To confirm that these stimuli elicit reliable fine-grained emotional attributions, a group of subjects on Amazon Mechanical Turk (MTurk: https://www.mturk.com) were asked to choose which of 20 emotion labels best described the emotion of the character in each stimulus. These subjects performed well above chance (relative to the emotion the stimulus was intended to elicit), classifying the stimuli with 65%[2] accuracy (chance=5%; see Figure S2A for behavioral confusion matrix). That is, subjects attribute consistent emotions from these stimuli, providing a benchmark with which to compare different models and brain regions.

To identify regions in which neural patterns contain information about emotions, we first replicate the finding that MPFC contains modality-independent emotion representations by testing whether neural patterns in MFPC can distinguish the valence of these verbal stimuli. We localized MPFC and other regions selective for ToM using individual subject localizers (see supplemental results). We selected a

[2] Note that because the set of emotions included subtle discriminations (apprehensive vs. terrified), and stimulus events could evoke a combination of different emotions, either simultaneously, or over the course of the vignette, we should not expect to obtain 100% agreement across subjects.

subset of emotion conditions that most closely align with the positive and negative conditions used in previous work (Skerry and Saxe, 2014), and tested whether neural patterns in MPFC would support above-chance classification of these conditions. Replicating prior work, classification of valence was reliably above chance in both DMPFC (M(SEM)=0.610(0.028), t(19)=3.889, p<0.001) and MMPFC (M(SEM)=0.603(0.019), t(19)=5.530, p<0.001).

We then investigate whether these or other regions contain information about the full set of 20 emotions. We conducted a whole brain-searchlight to find regions in which the local neighborhood could classify the 20 emotions above chance. The set of regions that could reliably decode the 20 emotions was largely restricted to regions of the ToM network (particularly DMPFC, RTPJ, LTPJ; see Figure 1B and Supplemental Data Table S2). The searchlight analysis exhibits striking overlap with the set of regions recruited for theory of mind (Figure 1B shows overlap between the searchlight (FWE p<.05, k>25) and the random effects analysis of the localizer task, shown at p<.001 uncorrected), and justifies our continued focus on these a priori ROIs.

Consistent with the searchlight results, we were able to classify emotions above chance (1/20 emotions, 5%) based on neural patterns in all individually localized ToM regions: DMPFC: M(SEM)=0.093(0.005), t(19)=9.018, p<0.001; MMPFC: 0.094(0.006), t(19)=7.043, p<0.001; VMPFC: 0.080(0.006), t(17)=5.156, p<0.001; RTPJ: 0.092(0.005), t(21)=8.205, p<0.001; LTPJ: 0.075(0.005), t(21)=4.744, p<0.001;PC: 0.079(0.006), t(21)=4.749, p<0.001;RSTS: 0.082(0.006), t(20)=5.380, p<0.001 (Figure 1A).

Moreover, in the judgments provided by subjects on MTurk there are reliable differences across the emotion categories in the extent to which subjects make consistent judgments (e.g. that subjects are very reliable in their attribution of “terrified” but less consistent in labeling an event as “joyful”; one-way ANOVA: F(19,180)=4.99, p<0.0001. see Figure 2A). These differences serve as another signature with which to compare neural representations. Thus, we also computed separate accuracies for each emotion category in each ROI, and correlated these with the behavioral emotion accuracies. In all ROIs, the accuracy of neural classifications for different emotions was significantly correlated with the accuracy levels observed in the emotion judgments of the behavioral raters on MTurk (Figure 2a): DMPFC: r(18)=0.70, p=0.001; MMPFC: r(18)=0.53, p=0.017; VMPFC: r(18)=0.47, p=0.036; RTPJ: r(18)=0.55, p=0.012; LTPJ: r(18)=0.71, p<0.001; PC: r(18)=0.46, p=0.042; RSTS: r(18)=0.65, p=0.002 (see Fig 2b for scatterplot in DMPFC). Thus, the reliable across-emotion accuracy differences observed behaviorally are paralleled in the emotion-specific accuracies of these neural populations (see Fig S2B for neural confusion matrices).

RSA:

Representational similarity analyses were then used to test specific hypotheses about the structure of the representations in these regions. We generated three competing feature spaces using independent behavioral ratings (Figure 3A) and tested which feature space could best capture the neural representation of the 20 emotions. We first analyzed the behavioral data alone, assessing the extent to which emotion categories could be reliably classified based on feature vectors in each of these candidate spaces. Do any of these feature spaces provide a stimulus representation sufficient to match the performance of human subjects in discriminating these 20 emotions (65%)? We found that although all three feature spaces were well above chance level of 5%, the appraisal feature space outperformed the other lower-dimensional spaces (57%, compared to behavioral benchmark of 65%; see Figure 3B). Using a paired samples t-test across individual items, we found that the abstract appraisal space performed reliably better than the circumplex space ($t(199)=8.288$, $p<0.001$) and the basic emotion space ($t(199)=2.176$, $p=0.031$).

These feature spaces were then compared to neural RDMs in each region to identify the space that best accounts for the similarity of the conditions in their neural patterns. In addition to the 3 spaces described above, we tested a model in which condition similarity is defined in terms of similarity of word-frequency vectors (see Experimental Procedures), following the approach of previous attempts to characterize neural semantic representations in terms of word frequencies or word co-occurrences [81–83]. Does the 38 dimensional appraisal space, which represents emotions in terms of a set of abstract intermediate variables, outperform a raw word-level representation of the stimuli? We also tested three control spaces capturing possible lower-level dimensions of the stimuli—reading ease, syntactic complexity, and rated intensity (confounded with motor response).

For each region, we correlated the RDMs for the competing feature spaces to neural RDMs from individual ROIs (distances of the 20 emotions in their voxel-wise patterns). In the two MPFC subregions, the similarity of conditions in their voxel level patterns was positively correlated with their similarity in the space of 38 appraisal dimensions (at the group-level—DMPFC: kendall's tau = 0.28; MMPFC: kendall's tau= 0.21). Moreover, correlations with individual subject neural RDMs (Figure 5) revealed a reliable relationship between the neural and model RDMs (DMPFC: M(SEM) kendall's tau=0.08(0.02), z(19)=3.32 p<0.001; MMPFC: 0.06(0.02), z(19)=2.95 p=0.002). In both of these regions, the correlation with the 38-dimensional space reached the lower bound of the noise ceiling (suggesting that although the average correlations are low, they approach the theoretical maximum given noise in the individual neural RDMs; see Experimental Procedures). Moreover, in both DMPFC and in MMPFC, the neural similarity space was more correlated with the space of 38 appraisal features than with either of the two dimensional spaces: the basic emotions space (DMPFC: 0.08 vs. 0.05, z(19)=3.02, p=0.002; MMPFC: 0.06, vs. 0.03, z(19)=2.31, p=0.021) and the circumplex space (DMPFC: 0.08 vs. 0.06, z(19)=2.84, p=0.005; MMPFC: 0.06 vs. 0.04, z(19)=2.80, p=0.005).

In both regions, the space of abstract appraisal features also outperformed a similarity space defined in terms of word-token frequencies (DMPFC: 0.08 vs. 0.02, z(19)=2.99, p=0.003, MMPFC: 0.06 vs. 0.02, z(19)=2.17, p=0.030), a representation frequently used in fully automated approaches to emotional text classification such as sentiment analysis of reviews or other social media [84, 85]. To control for lower-level properties of the verbal stimuli, we also compared the neural RDMs to the similarity of stimuli in their reading ease, their syntactic complexity, and rated intensity (confounded with motor response). In both regions, the correlation with the space of 38-appraisals was higher than for reading ease (DMPFC: 0.08 vs. 0.02, z(19)=2.39, p=0.017, MMPFC: 0.06 vs. 0.01, z(19)=2.02, p=0.044), syntactic complexity (DMPFC: 0.08 vs. 0.03, z(19)=2.50, p=0.012, MMPFC: 0.06, vs. 0.02, z(19)=1.98, p=0.048), and intensity (DMPFC: 0.08 vs. 0.02, z(19)=3.21, p=0.001, MMPFC: 0.06 vs. 0.03, z(19)=2.05, p=0.040).

In addition to our a priori ROIs, we conducted the same analyses in the remaining ToM regions (RTPJ, LTPJ, PC, RSTS, and VMPFC): these ROIs were also reliably correlated with the space of 38-appraials (RTPJ: M(SEM)=0.07(0.02), z(21)=3.59 p<0.001 (see Figure 5); see Figure S6 for results from other ToM regions), and no region was reliably more correlated with the basic emotion or circumplex spaces. The 38-dimensional space outperformed competing spaces in all ToM regions except for VMPFC (where the best performing space was the word frequency representation). In VMPFC, RSTS, and RTPJ (but not in PC and LTPJ) the neural-model correlations passed the lower bound of the noise ceiling (Figure S6). However, DMPFC and MMPFC were the only regions in which the high-dimensional space significantly outperformed all other models.

Region contributions:

We could reliably decode emotion in all of the theory of mind ROIs, and the same 38-dimensional feature space did the best job of capturing the neural similarity space in all regions other than VMPFC. Is the same information represented redundantly across these regions, or is there evidence that these regions contribute differently to the representation of emotions? To address this question, we first compare classification accuracies using single ROIs to the classification accuracy when combining regions across the ToM network. When classifying only valence, a model trained with voxels from all ToM ROIs (M(SEM)=0.581(0.016), t(21)=4.942, p<0.001) performs less well than a model trained only with voxels in DMPFC or MMPFC (58.1% relative to 61% in DMPFC). In contrast, when classifying the full set of 20 emotions, a model trained with voxels from all regions of the network outperforms any of the individual ROIs. Classification using the voxels from all regions of the ToM network (M(SEM)=0.108(0.006), t(21)=9.135, p<0.001; see Figure 5) was reliably higher than classification using only voxels in DMPFC (t(38)=2.684, p=0.015), MMPFC (t(38)=2.848, p=0.01), or RTPJ (t(42)=2.773, p=0.011), suggesting that the individual ROIs could contribute non-redundant information.

To further characterize representational differences across the ROIs, we explored whether the regions differ in the particular situation features they represent. Rather than compute separate RDMs for all 38 appraisal features, we identified a reduced set of 10 features that capture the most unique variance in behavioral ratings across items (see Figure S5; Experimental Procedures). We then computed the RDMs for this 10-dimensional space, and also for each of the 10 features individually, and correlated each with the neural RDMs in different regions. Thus in each subject, we obtained neural-feature correlations for each of the 10 features in each ROI. By testing for feature x ROI interactions across subjects, we can thus test for differences in the feature representations across ROIs. We focus in particular on comparing MPFC and RTPJ, as these regions have been proposed to be involved in distinct aspects of mental state reasoning (affective and epistemic respectively; see Koster-Hale et al., in review). The neural RDMs in DMPFC, MMPFC, and RTPJ were reliably correlated with the RDM of the 10-feature space (DMPFC: M(SEM)=0.08(0.02), z(19)=3.21 p=0.001; MMPFC: M(SEM)=0.05(0.02), z(19)=2.61 p=0.004; RTPJ: M(SEM)=0.06(0.01), z(21)=3.55 p<0.001), and this smaller set of features appears to capture much of the representational structure of the initial 38-d space (Figure 6; see Figure S7 for results from secondary ROIs). A repeated-measures ANOVA on the neural-model correlations for each feature (with ROI and feature as within-subjects factors) revealed a significant *ROI* x *feature* interaction for the comparison of DMPFC and RTPJ (F(9,171)=2.06, p=0.036) but not between MMPFC and RTPJ (F(9,171)=1.036, p=0.414). The differential feature representations between DMPFC and RTPJ suggest that although multiple ToM regions are involved in the attribution of emotion, some of these regions may contribute unique information to the final representational space that governs behavior. For example, exploratory analyses reveal that the correlation with the "self cause" feature ("Was this situation caused by <character> herself or by someone/something else?") is reliably higher in DMPFC than in RTPJ, where as the "distant past" feature ("Did this situation involve events from <character>'s distant past?") is more correlated with the RDM in RTPJ than in DMPFC.

DISCUSSION:

The ability to predict and infer the emotional states of others is central to our species' unique social and cooperative behaviors [86]. In the present research, we provide evidence that neural patterns in regions involved in mental state reasoning contain information relating to the emotional states of others. Moreover, we provide quantitative insight into the underlying representational structure that supports this inferential ability.

The structure of emotion knowledge

Decades of research in the science of emotion have aimed to characterize emotions in terms some low-dimensional space of basic affective primitives [1, 59, 67]. Behaviorally, we find that a space of 38 abstract event features, inspired by work in appraisal theory [80], reliably outperforms these simpler spaces in discriminating the 20 different emotions in our stimuli. While affective dimensions that make up the circumplex model [59] or basic emotions theory [1, 21] may capture the range of emotions we express and perceive with overt expressions, a higher-dimensional space is needed to encode the range attributions elicited by short verbal descriptions of events.

Interestingly, a model using the 38-dimensional space still falls short of human behavioral performance when labeling stimuli (57% versus 65% accurate), indicating that this collection of features does not adequately capture our intuitive emotion knowledge. There are at least two plausible reasons for this inadequacy. First, the 38 dimensions used in the present study were derived from prior literature on emotion appraisal without subsequent optimization. This list may therefore contain redundant or uninformative features, and some additional features are likely necessary.

A second, more fundamental limitation is that this approach aims to encode human emotion knowledge in terms of flat feature vectors (i.e. lists of appraisal checks applied to each stimulus). While this feature-based approach has been productive in other domains of perception and cognition [87–91] and proved useful in the present paradigm, it is unlikely that representations in a domain of high-level cognition such as theory of mind can be reduced to operations over lists of associated features [92, 93]. For example, emotions are caused by events that unfold over time: the emotion attributed depends critically on the temporal and causal order of the different elements of the event (e.g. eating a whole cake and then swearing to keep to your diet; versus swearing to keep your diet and then eating a whole cake). To capture the richly causal and compositional nature of the representations involved in emotion inference [12], future research will need to move beyond a feature-based approach, incorporating structured, generative knowledge representations from other areas of cognitive science [94, 95].

Nonetheless, present research makes important advances in our understanding of emotion inference. While constructivist theories of emotion have long acknowledged that attribution depends on emotion-specific conceptual knowledge [4, 6, 8, 68, 96, 97], the content and structure of that knowledge has remained unclear. Here, we provide an initial sketch of specific features that might structure human emotion concepts, and provide a framework for evaluating competing models of this knowledge.

Neural representations of others' emotions

Consistent with previous reports [36, 37], the present results suggest that neural representations in MPFC contain information about attributed emotions. Whereas prior studies focused on coarse distinctions (e.g. valence), we are able to classify a set of nuanced emotions at above chance levels, suggesting that emotion representations in this region are relatively fine-grained. Moreover, by expanding to a rich space of eliciting situations, we are able to decode attributed emotions in all regions of the ToM network. Using a whole brain searchlight, we find that although emotion information is present in many regions, this information is largely restricted to regions involved in theory of mind (particularly MPFC, RTPJ, and LTPJ). When combining information from voxels across all ToM regions, we were able to decode the emotion label of a stimulus with ~10% accuracy.

Although these classifications were reliably above chance (5%), they are far from reaching the accuracy observed in behavior (65%). This discrepancy between neural and behavioral classification could arise because the population code in these regions is insufficient to explain the behavior, or because single trial estimates of fMRI data provide a noisy, highly blurred measurement of the underlying neural code. However, across different emotions, there were reliable correlations in the average accuracy of the neural populations and of independent behavioral raters, providing support for the role of these regions in emotion attribution behaviors.

We then further probed the underlying representational structure that supports successful emotion discrimination. The previous literature [35–37] is consistent with the possibility that MPFC codes a limited space of affective dimensions such as valence and/or arousal. Moreover, even in our MPVA analyses, a region could support 20-way classification at above-chance levels by coding only a single dimension or feature that varies across emotions. Using RSA, we find that brain regions selective for theory of mind not only contain information about attributed emotions, they are also best captured by the high-dimensional space of event features.

In all but one of the ToM regions, the similarity of emotion conditions in their voxel response patterns was most correlated with the similarity of the emotions in the space of 38 appraisals. This result suggests that the neural code in these regions does not reduce to a simpler set of distinctions such as valence and arousal, and provides novel insight into the granularity of the emotion representations in MPFC and other ToM regions. Together, the behavioral and neural data suggest that human emotion attribution is organized around abstract features of the causal context in which different emotions occur, rather than the sorts of affective primitives that have dominated prior research.

A challenge for future work will be characterizing the scope and specificity of the neural representations in the observed regions. One possibility is that these neural populations contain representations specific to attributed emotion, and that these attributed states are coded within a space of emotion-relevant causal features. Alternatively, there could be neural populations that contain information about emotion-relevant features, but in the form of domain-general semantic representations. These event representations might serve as intermediate features in the service of diverse inferential processes beyond emotion attribution. Ultimately, successful emotion inference depends on a rich body of general world knowledge, and neural populations specific to social cognition must interface with more general-purpose semantic processing mechanisms. Characterizing information flow within and between these different networks will be an important avenue for future research.

Relation to prior research

In the present research, we provide a first attempt to characterize the feature space that governs emotion representation in the human brain. To do so, we draw heavily on methods and ideas that have been fruitful in recent research on visual object recognition and object semantics, where researchers have tested a range of high-level and low-level features that could capture neural similarity of different objects [46, 81, 83, 98–102]. In one study, Mitchell and colleagues [81] coded object words in terms of their co-occurrence with a set of 25 verbs hypothesized to pick out relevant semantic dimensions (e.g. “manipulate”, “taste”), and found that this representation was sufficient to support above chance neural classification of untrained stimuli. Further analyses of these data show that a corpus-based co-occurrence space is outperformed by a space derived from behavioral ratings on a set of a priori object properties (e.g. is it alive?) [101, 103]. The present research is most similar to this second approach, relying on behavioral ratings of a set of hypothesized event features. We show that it is possible to generate candidate representational spaces for domains of high-level cognition such as emotion inference, and to use these spaces to characterize patterns of activity in theory of mind brain regions.

With this approach, we hope to move beyond identifying regions that contain information about emotion attributions, and gain insight into the intermediate stages and corresponding features used to construct these attributions. In the study of object representation, researchers have made headway in understanding differences across regions and temporal stages [99, 102, 104]; representational similarity analysis in particular has provided a flexible framework for comparing the structure of the representations in different regions along the ventral pathway [105, 106]. Interestingly, the present results provide preliminary evidence that theory of mind regions differ in their contributions to emotion inference. When classifying the 20 emotions, we find a reliable advantage to using voxels from the whole network, compared to any region in isolation. Moreover, we observed region-by-feature interactions in the RSA analyses, suggesting that the regions differ in the specific appraisal features that dominate their response. Further work is needed to characterize the precise computational roles of these regions and how they interact with other networks to form a processing stream.

As has been the case in research on object representation, we assume that future studies of emotion attribution will yield feature spaces that outperform the 38-dimensional space explored here.

Future work might aim to not only better fit the neural data, but also to build computational models capable of extracting the relevant intermediate features directly. Many early approaches to modeling neural object representations involved hand-picked feature spaces (e.g. 25 chosen verbs) [81, 107] and often manual coding of stimuli within those spaces [99, 101, 108, 109]. However, recent research has yielded computational models that can be applied to raw stimuli (i.e. images) and achieve high quantitative fit to neural patterns [110]; even relevant feature spaces themselves can be discovered in a bottom-up manner [82, 111, 112]. In our study, candidate features were selected based on prior domain knowledge, and the stimuli required manual annotation into these feature spaces (MTurk ratings). Future research in this area should ideally identify new sets of optimized features (either event features or some other candidate basis), and new ways to infer these features from text alone, removing the need for a human subject or experimenter in the loop.

Despite these important open questions, the present data provide novel insight into the representations underlying human emotion inference and the neural populations that support them. Together, the results suggest that our knowledge of others' emotions is abstract and high-dimensional, that brain regions associated with emotion perception and inference contain information about relatively fine-grained emotional distinctions, and that the neural representations in these regions not reducible to more primitive affective dimensions such as valence and arousal.

EXPERIMENTAL PROCEDURES:

Further details on experimental procedures (e.g. ROI selection and univariate analyses) are provided in the Supplemental Experimental Procedures.

Stimuli: All experiments used a set of 200 verbal stimuli (2-3 sentences; M(SEM)= 50.68(0.28) words; see Table 1) describing a character experiencing one of 20 different emotions. In each item, the emotion was conveyed via a description of an emotion-eliciting event, without any labeling or description of the character's reaction.

Behavioral attributions: To verify that subjects make reliable attributions of the emotions conveyed in the 200 stimuli, subjects on Amazon Mechanical Turk (MTurk, N=139) were asked to choose which of the

20 emotions best described the character’s emotional state (see Supplemental Experimental Procedures). Predicted emotions for each stimulus were used to compute an overall accuracy level (relative to the intended emotion for each stimulus; see Figure 3A), as well as a confusion matrix (the proportion of time each intended emotion was labeled as each of the emotion categories: see Figure S4A).

Behavioral feature ratings: A separate set of MTurk subjects (N=250) provided ratings (1-10 scale) for each of the stimuli on each of the features of the three competing feature spaces. A given subject rated stimuli on either features from the 38-dimensional appraisal space (e.g. “Did someone cause this situation intentionally or did it occur by accident?” 1= caused accidentally, 5=neutral/not applicable, 10=caused intentionally; see Supplemental Table), or dimensions corresponding to the basic emotion space (e.g. “Was <character> happy in this situation?” 1= not at all happy, 5=somewhat/not applicable, 10=very happy) and the circumplex space (e.g. “Did <character> find this situation to be positive or negative?” 1=clearly negative, 5=neutral/not applicable, 10=clearly positive).

Feature-based classification of behavioral data: To test whether any of the 3 candidate spaces (basic emotion, circumplex, and 38 appraisals) capture the full range of attributed emotions, we created an item-by-feature matrix for each possible space, and tested whether a model trained on these features could classify the 20 distinct emotions. Specifically, we trained a linear SVM (SVC, one vs. one implementation) on emotion-labeled feature vectors for a subset of items, and tested whether the classifier could generate the appropriate label for a different set of items. Thus, we test whether each feature space provides a basis for emotion discrimination that generalizes across the different exemplars[3]. We conducted this procedure iteratively (n=1000), splitting items into 100 training and test exemplars (5 items for each emotion condition), and computed the average cross-item classification accuracy for each feature space,

[3] Note, because the set of 20 emotions includes some basic emotions (disgust, surprised), the comparison of the 38-appraisal spaces to the space of basic emotions is rather conservative. That is, the basic emotion space actually contains as features some of the labels we are trying to predict (i.e. to predict emotion categories like “disgusted”, “surprised”, and “terrified”, the basic emotion space uses ratings of the extent to which the event elicits “disgusted”, “surprised”, “afraid” emotions). If the 38-dimensional space outperforms the basic emotion space despite being at such a disadvantage, this would provide particularly compelling evidence for the role of causal context in structuring our representations of others’ emotions.

to compare to the behavioral benchmark (65%). We also computed the accuracy separately for each item, and tested for reliable differences between the three feature spaces using a t-test across items.

FMRI participants: 22 right-handed adults ages 18-40 (M_{age}= 25.39, STD_{age}=5.43; 13 female) participated in the study. All participants had normal/corrected-to-normal vision and no history of neurological disorders. Participants gave written, informed consent in accordance with the requirements of the MIT institutional review board. We collected behavioral measures of social-cognitive ability from each participant (see Supplementary Experimental Procedures)

FMRI tasks

Theory of mind localizer: Subjects were presented with short textual scenarios that required inferences about mental states (Belief condition) or physical representations such as a map or photo (Photo condition; [41, 113] (stimuli available at http://saxelab.mit.edu/superloc.php). Scenarios were presented for 10s, followed by a true or false question (4s) about either the representation (Belief or Photo) or the reality of the situation. Each run (4.53min) consisted of 10 trials separated by 12s inter-stimulus intervals, and 12s blocks of fixation were included at the beginning and end of each run. 1-2 runs were presented to each participant, with the order of stimulus type (Belief or Photo) and correct answer (True or False) counterbalanced within and across runs.

Emotion Attribution task: In the Emotion Attribution task, subjects viewed the 200 emotion stimuli, as well as a set of 10 stories describing physical pain [42]. The experiment consisted of 10 runs (7.37min/run), each containing 1 exemplar for each of the 21 conditions (20 emotions plus 1 pain stimulus). Each story was presented at fixation for 13s, followed by a 2s window during which subjects made a behavioral response. Subjects were instructed to press a button to indicate the intensity of the character's experience (1 to 4, neutral to extreme), which focused subjects' attention on the character's emotional state, but ensured that behavioral responses (intensity) were orthogonal to discriminations of interest. The stories were presented in a jittered, event-related design, with a central fixation cross presented between trials at a variable inter-stimulus interval of 3-5-7 seconds. The order of conditions was counterbalanced across runs and participants, and the order of individual stories for each condition was randomized.

FMRI 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 (1mm isotropic) T-1 weighted MPRAGE anatomical scan, followed by functional images acquired with a gradient-echo EPI sequence sensitive to blood-oxygen-dependent (BOLD) contrast (repetition time [TR] = 2s, echo time [TE] = 30ms, flip angle = 90°, voxel size 3x3x3mm, matrix 64x64, 32 axial slices).

FMRI analyses

Preprocessing: MRI data were preprocessed using SPM8 (www.fil.ion.ucl.ac.uk/spm/software/spm8/), freesurfer (for skull-stripping; http://surfer.nmr.mgh.harvard.edu/), and in-house code. SPM was used to motion correct each subject's data via rigid rotation and translation about the 6 orthogonal axes of motion, to register the functional data to the subject's high-resolution anatomical image, to normalize the data onto a common brain space (MNI), and to smooth images with a Gaussian filter (FWHM=5mm).

MVPA Classification Analyses: We first aimed to replicate previous valence decoding in MPFC (Skerry and Saxe, 2014) by choosing subset of conditions that most closely match the happy versus sad emotions used in that study ("”Excited”, “Joyful”, “Proud” vs. “Devastated”, “Disappointed”, “Annoyed”) and testing whether voxel patterns in MFPC could reliably classify the valence of these stimuli. We then tested whether voxel patterns in MFPC or other ToM regions could reliably classify the set of 20 emotions. Code for multi-voxel pattern classification was developed in Python using the publicly available PyMVPA toolbox (http://www.pymvpa.org/).

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, with linear detrending across the whole time-course. A timepoint was excluded if it was a global intensity outlier (> 3 SD above the mean intensity) or contained a large movement (> 2mm scan-to-scan). The data were temporally compressed to generate a voxel-wise average for each individual trial, and these single trial summaries were used for training and testing. The individual trial patterns were calculated by averaging the preprocessed bold images for the 12TR duration of the trial, offset by 5 TRs to account for the HRF and lag in relevant context (offset and duration selected based on subset of time-course in which the response magnitude differed between pain and emotion stimuli). Rest timepoints were removed and the trial summaries were concatenated. The pattern for each trial was then z-scored relative to the mean across all trial responses in that voxel.

The data were classified using a support vector machine; 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 responses in each voxel. For the 20-way discrimination, multi-class classification was conducted with a one-vs-one method [114], yielding a single condition prediction for each trial. We used a fixed regularization parameter (C=1) and restricted ourselves to linearly decodable signal under the assumption that a linear kernel implements a plausible readout mechanism for downstream neurons [115–117].

The data were partitioned into 10 run-based folds and the classifier was trained iteratively on all runs but one, and tested on the remaining run. Classification accuracy was averaged across folds to yield a single score for each subject in the ROI. A one-sample t-test was then performed over these individual accuracies, comparing to chance classification (.5 for positive versus negative, and .05 for the 20-way emotion classification; all t-tests on classification accuracies were one-tailed). We also performed this analysis using the full ToM network, where each subject's ROI was the union of his/her individually localized ROIs.

Whole brain searchlight classification: The searchlight procedure was identical to the ROI-based procedure except that the classifier was applied to voxels within local spheres rather than individually localized ROIs. For each voxel in a gray matter mask, we defined a sphere containing all voxels within a 3-voxel radius (123 voxels) of the center voxel. 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. A one-sample t-test over subjects' accuracy maps (comparing accuracy in each voxel to chance—0.05) yielded a group t-map, which was assessed at a p<.05 (K>25), FWE corrected (based on SPM's implementation of Gaussian Random Fields).

Representational Similarity Analyses: To create representational dissimilarity matrices (RDMs) for the competing feature spaces, we first averaged the feature vectors for each emotion condition (across stimuli), yielding the emotion-by-feature matrices shown in Figure 3. For each matrix, we then computed the Euclidean distance of feature vectors for each pair of emotions. We conducted this analysis iteratively (n=1000) across split halves of the data (5 items per condition in each half), such that the self-distances along the diagonal are meaningful.

In addition to the three candidate feature spaces described above (circumplex model, basic emotions, and abstract appraisals), we generated an additional space defined in terms of the similarity in word occurrences across stimuli. Features vectors were created based on frequencies of individual words from the stimuli, excluding English stop words and stripping common morphological endings. We used a term frequency-inverse document frequency (tf-idf) vectorizer such that the exemplar value for each word increases with the frequency of the word in the exemplar, but decreases with the frequency of the word in the full stimulus set. This type of flat, unordered word-level representation is popular in existing machine learning approaches to sentiment analysis/emotion classification [84, 85]. Finally, we computed several additional control spaces to confirm that neural RDMs could not be explained in terms of lower-level properties of the stimuli: reading ease, syntactic complexity, and behavioral ratings of intensity (see Supplemental Procedures).

Neural RDMs were computed separately for each region in each subject. These were computed with a procedure analogous to that described for feature space RDMs, except that the features were voxel-wise neural responses rather than the behavioral feature ratings. We averaged voxel response vectors for each condition separately split halves of the data, yielding two condition-by-voxel matrices (for even and odd runs). We then computed similarity of the conditions in terms of Euclidean distance of the voxel patterns across runs, yielding a RDM for each region (again this is done across even and odd subsets so that the diagonal is interpretable). Each neural RDM was normalized by subtracting its minimum value and dividing by the range, yielding a matrix with distances ranging from 0 to 1. This procedure was conducted separately for each individual subject, and individual subject neural RDMs for each region were averaged to generate a group RDM for the region.

To compare neural and model similarity spaces, we then computed the rank correlation (kendall's tau-a) between the neural and model RDMS for each region. The group neural RDM will be least noisy, and therefore provide the best estimate of the relationship between the true neural RDM and each of the model spaces. However, to assess the reliability of the neural-model relationships, we compute the neural-model correlations separately for each subject, and perform a Wilcoxon test comparing the individual subject correlations to chance (average kendall's tau = 0). We also compare the fit of different models by conducting a one-tailed Wilcoxon signed-rank test on the correlations for different pairs of models.

We compare these neural-model correlations to a behavioral benchmark (dotted line in Figures 5 and 6) was defined as the correlation of the neural RDM with an RDM computed from the confusion matrix of independent behavioral raters. From the behavioral classification study, we have scores representing the frequency with which a given emotion is misclassified as another emotion. We then computed Euclidean distances of the conditions with this matrix to yield a RDM capturing the similarity space of the emotion conditions (though because behavioral performance was relatively high, the confusion matrix is sparse and may underestimate the similarity structure of the emotions). We then test whether any of our intermediate feature spaces (38-dimensions, basic emotions, circumplex model, etc.) meet or surpass this benchmark correlation. Given that our analyses depends on neural-model correlations computed in individual subjects, we wished to assess the extent to which our correlations are limited by noise in the individual subject neural RDMs. We estimated a noise ceiling for this analyses approach, on the assumption that individual neural RDMs can be no more correlated with a model than they are with the true neural RDM [118]. We computed the correlation of each individual RDM to the mean RDM of the full group, which potentially overestimates the reliability of the individual RDMs since the individual subject's data is included in the group. We then used a leave-one-out procedure to correlate the individual RDMs to the mean of the group RDM excluding that subject; this potentially underestimates the reliability of the individual neural RDMs with the true neural representation, since the group mean is an average of a small sample of subjects. These two values serve as upper and lower bounds on the neural-model correlations we can expect to observe with this analysis approach.

Region analysis—comparisons of individual features: Using a set of 10 features that explain the most unique variance across stimuli (see Supplemental Procedures and Data), we created a reduced 10-dimensional space and subject it to the same behavioral and neural analyses described above. Specifically, we compute RDMs for the 10-feature space, and for each feature in isolation, and correlate these with the neural RDMs in each region. To assess whether regions differ in the individual features they represent, we conducted repeated measures ANOVA on the correlations (kendall's tau) between neural and the feature RDMs, with *ROI* and *feature* as within-subject factors.

Contributions: AES and RS designed and performed the research, analyzed the data, and wrote the paper.

Acknowledgements: This research was supported by National Science Foundation Graduate Research Fellowship (A.E.S.) and NIH Grant 1R01 MH096914-01A1 (R.S.). We thank Laura Schulz, Josh Tenenbaum, Nancy Kanwisher, Michael Cohen, Jorie Koster-Hale, Todd Thompson, and Erik Nook for helpful comments and discussion.

REFERENCES

1. Ekman, P. (1992). Are there basic emotions? Psychol. Rev. *99*, 550–553.

2. Izard, C. E. (1971). The face of emotion (Appleton-Century-Crofts).

3. Bachorowski, J.-A., and Owren, M. J. (2003). Sounds of Emotion. Ann. N. Y. Acad. Sci. *1000*, 244–265.

4. Clore, G. L., and Ortony, A. (2013). Psychological Construction in the OCC Model of Emotion. Emot. Rev. *5*, 335–343.

5. Sauter, D. A., Eisner, F., Ekman, P., and Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. Proc. Natl. Acad. Sci. *107*, 2408–2412.

6. Brosch, T. (2013). Comment: On the Role of Appraisal Processes in the Construction of Emotion. Emot. Rev. *5*, 369–373.

7. Aviezer, H., Trope, Y., and Todorov, A. (2012). Body Cues, Not Facial Expressions, Discriminate Between Intense Positive and Negative Emotions. Science *338*, 1225–1229.

8. Lindquist, K. A. (2013). Emotions Emerge from More Basic Psychological Ingredients: A Modern Psychological Constructionist Model. Emot. Rev. *5*, 356–368.

9. Dael, N., Mortillaro, M., and Scherer, K. R. (2012). Emotion expression in body action and posture. Emot. Wash. DC *12*, 1085–1101.

10. Lindquist, K. A., MacCormack, J. K., and Shablack, H. (2015). The role of language in emotion: Predictions from psychological constructionism. Lang. Sci. *6*, 444.

11. De Gelder, B. (2006). Towards the neurobiology of emotional body language. Nat. Rev. Neurosci. *7*, 242–249.

12. Ortony, A. (1990). The Cognitive Structure of Emotions (Cambridge University Press).

13. Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. Psychol. Sci. *18*, 1050–1057.

14. Frijda, N. H. (1986). The Emotions (Cambridge University Press).

15. Ortony, A., and Turner, T. J. (1990). What’s basic about basic emotions? Psychol. Rev. *97*, 315–331.

16. Barrett, L. F., Lindquist, K. A., Bliss-Moreau, E., Duncan, S., Gendron, M., Mize, J., and Brennan, L. (2007). Of Mice and Men: Natural Kinds of Emotions in the Mammalian Brain? A Response to Panksepp and Izard. Perspect. Psychol. Sci. J. Assoc. Psychol. Sci. *2*, 297–311.

17. Barrett, L. F., Mesquita, B., and Gendron, M. (2011). Context in Emotion Perception. Curr. Dir. Psychol. Sci. *20*, 286–290.

18. Barrett, L. F., and Kensinger, E. A. (2010). Context is routinely encoded during emotion perception. Psychol. Sci. *21*, 595–599.

19. Hassin, R. R., Aviezer, H., and Bentin, S. (2013). Inherently Ambiguous: Facial Expressions of Emotions, in Context. Emot. Rev. *5*, 60–65.

20. Abelson, R. P., and Sermat, V. (1962). Multidimensional scaling of facial expressions. J. Exp. Psychol. *63*, 546–554.

21. Ekman, P., and Rosenberg, E. L. (1997). What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS) (Oxford University Press).

22. Russell, J. A., and Bullock, M. (1986). On the dimensions preschoolers use to interpret facial expressions of emotion. Dev. Psychol. *22*, 97–102.

23. Adolphs, R. (2002). Neural systems for recognizing emotion. Curr. Opin. Neurobiol. *12*, 169–177.

24. Calder, A. J., and Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. Nat. Rev. Neurosci. *6*, 641–651.

25. Furl, N., Hadj-Bouziane, F., Liu, N., Averbeck, B. B., and Ungerleider, L. G. (2012). Dynamic and Static Facial Expressions Decoded from Motion-Sensitive Areas in the Macaque Monkey. J. Neurosci. Off. J. Soc. Neurosci. *32*, 15952–15962.

26. Harry, B., Williams, M. A., Davis, C., and Kim, J. (2013). Emotional expressions evoke a differential response in the fusiform face area. Front. Hum. Neurosci. *7*. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3809557/ [Accessed January 1, 2014].

27. Said, C. P., Moore, C. D., Engell, A. D., Todorov, A., and Haxby, J. V. (2010). Distributed representations of dynamic facial expressions in the superior temporal sulcus. J. Vis. *10*, 11.

28. Said, C. P., Moore, C. D., Norman, K. A., Haxby, J. V., and Todorov, A. (2010). Graded Representations of Emotional Expressions in the Left Superior Temporal Sulcus. Front. Syst. Neurosci. *4*. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2842102/ [Accessed November 24, 2013].

29. Ethofer, T., Van De Ville, D., Scherer, K., and Vuilleumier, P. (2009). Decoding of Emotional Information in Voice-Sensitive Cortices. Curr. Biol. *19*, 1028–1033.

30. Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations. Psychol. Rev. *99*, 561–565.

31. Izard, C. E. (1994). Innate and universal facial expressions: evidence from developmental and cross-cultural research. Psychol. Bull. *115*, 288–299.

32. Jack, R. E., Garrod, O. G. B., and Schyns, P. G. (2014). Dynamic Facial Expressions of Emotion Transmit an Evolving Hierarchy of Signals over Time. Curr. Biol. *24*, 187–192.

33. Russell, J. A. (1994). Is there universal recognition of emotion from facial expressions? A review of the cross-cultural studies. Psychol. Bull. *115*, 102–141.

34. Aviezer, H., Hassin, R. R., Ryan, J., Grady, C., Susskind, J., Anderson, A., Moscovitch, M., and Bentin, S. (2008). Angry, disgusted, or afraid? Studies on the malleability of emotion perception. Psychol. Sci. *19*, 724–732.

35. Chikazoe, J., Lee, D. H., Kriegeskorte, N., and Anderson, A. K. (2014). Population coding of affect across stimuli, modalities and individuals. Nat. Neurosci. *17*, 1114–1122.

36. Peelen, M. V., Atkinson, A. P., and Vuilleumier, P. (2010). Supramodal Representations of Perceived Emotions in the Human Brain. J. Neurosci. *30*, 10127–10134.

37. Skerry, A. E., and Saxe, R. (2014). A common neural code for perceived and inferred emotion. J. Neurosci. Off. J. Soc. Neurosci. *34*, 15997–16008.

38. Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. Science *293*, 2425–2430.

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

40. Mitchell, J. P. (2009). Inferences about mental states. Philos. Trans. R. Soc. B Biol. Sci. *364*, 1309–1316.

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

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

43. Spunt, R. P., and Lieberman, M. D. (2012). An integrative model of the neural systems supporting the comprehension of observed emotional behavior. NeuroImage *59*, 3050–3059.

44. Zaki, J., Weber, J., Bolger, N., and Ochsner, K. (2009). The neural bases of empathic accuracy. Proc. Natl. Acad. Sci. *106*, 11382–11387.

45. Kriegeskorte, N., and Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. *17*, 401–412.

46. Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational Similarity Analysis - Connecting the Branches of Systems Neuroscience. Front. Syst. Neurosci. *2*. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2605405/ [Accessed November 21, 2014].

47. Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Comput Biol *10*, e1003915.

48. Darwin, C. (1872). The expression of the emotions in man and animals (London, John Murray) Available at: http://archive.org/details/expressionofemot1872darw [Accessed October 14, 2014].

49. Ekman, P. (1993). Facial expression and emotion. Am. Psychol. *48*, 384–392.

50. Ekman, P., and Cordaro, D. (2011). What is Meant by Calling Emotions Basic. Emot. Rev. *3*, 364–370.

51. Levenson, R. W. (2011). Basic Emotion Questions. Emot. Rev. *3*, 379–386.

52. Panksepp, J. (1992). A critical role for "affective neuroscience" in resolving what is basic about basic emotions. Psychol. Rev. *99*, 554–560.

53. Shariff, A. F., and Tracy, J. L. (2011). What Are Emotion Expressions For? Curr. Dir. Psychol. Sci. *20*, 395–399.

54. Oatley, K., and Johnson-Laird, P. N. (1996). The communicative theory of emotions: Empirical tests, mental models, and implications for social interaction. In Striving and feeling: Interactions among goals, affect, and self-regulation, L. L. Martin and A. Tesser, eds. (Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc), pp. 363–393.

55. Plutchik, R. (2001). The Nature of Emotions. Am. Sci. *89*, 344.

56. Turner, J. H. (2000). On the Origins of Human Emotions: A Sociological Inquiry Into the Evolution of Human Affect (Stanford University Press).

57. Du, S., Tao, Y., and Martinez, A. M. (2014). Compound facial expressions of emotion. Proc. Natl. Acad. Sci., 201322355.

58. Barrett, L. F. (2006). Valence is a basic building block of emotional life. J. Res. Personal. *40*, 35–55.

59. Russell, J. A. (1980). A circumplex model of affect. J. Pers. Soc. Psychol. *39*, 1161–1178.

60. Russell, J. A., and Bullock, M. (1986). Fuzzy Concepts and the Perception of Emotion in Facial Expressions. Soc. Cogn. *4*, 309–341.

61. Watson, D., Wiese, D., Vaidya, J., and Tellegen, A. (1999). The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. J. Pers. Soc. Psychol., 820–838.

62. Barrett, L. F., and Bliss-Moreau, E. (2009). Chapter 4 Affect as a Psychological Primitive. In Advances in Experimental Social Psychology, Mark P. Zanna, ed. (Academic Press), pp. 167–218. Available at: http://www.sciencedirect.com/science/article/pii/S0065260108004048 [Accessed October 24, 2014].

63. Barrett, L. F., and Wager, T. D. (2006). The Structure of Emotion Evidence From Neuroimaging Studies. Curr. Dir. Psychol. Sci. *15*, 79–83.

64. Lindquist, K. A., Satpute, A. B., Wager, T. D., Weber, J., and Barrett, L. F. (2015). The Brain Basis of Positive and Negative Affect: Evidence from a Meta-Analysis of the Human Neuroimaging Literature. Cereb. Cortex, bhv001.

65. Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., and Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. Behav. Brain Sci. *35*, 121–143.

66. Wilson-Mendenhall, C. D., Barrett, L. F., and Barsalou, L. W. (2013). Neural Evidence That Human Emotions Share Core Affective Properties. Psychol. Sci. *24*, 947–956.

67. Posner, J., Russell, J. A., Gerber, A., Gorman, D., Colibazzi, T., Yu, S., Wang, Z., Kangarlu, A., Zhu, H., and Peterson, B. S. (2009). The neurophysiological bases of emotion: An fMRI study of the affective circumplex using emotion-denoting words. Hum. Brain Mapp. *30*, 883–895.

68. Barrett, L. F. (2006). Are Emotions Natural Kinds? Perspect. Psychol. Sci. *1*, 28–58.

69. Posner, J., Russell, J. A., and Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev. Psychopathol. *null*, 715–734.

70. Engen, T., Levy, N., and Schlosberg, H. (1958). The dimensional analysis of a new series of facial expressions. J. Exp. Psychol. *55*, 454–458.

71. Hamann, S. (2012). Mapping discrete and dimensional emotions onto the brain: controversies and consensus. Trends Cogn. Sci. *16*, 458–466.

72. Vytal, K., and Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. J. Cogn. Neurosci. *22*, 2864–2885.

73. Barrett, L. F., Lindquist, K. A., and Gendron, M. (2007). Language as context for the perception of emotion. Trends Cogn. Sci. *11*, 327–332.

74. Ellsworth, P. C. (2013). Appraisal Theory: Old and New Questions. Emot. Rev. *5*, 125–131.

75. Lazarus, R. S. (1991). Progress on a cognitive-motivational-relational theory of emotion. Am. Psychol. *46*, 819–834.

76. Scherer, K. R. (1999). Appraisal theory. In Handbook of cognition and emotion, T. Dalgleish and M. J. Power, eds. (New York, NY, US: John Wiley & Sons Ltd), pp. 637–663.

77. Meuleman, B., and Scherer, K. (2013). Nonlinear Appraisal Modeling: An Application of Machine Learning to the Study of Emotion Production. IEEE Trans. Affect. Comput. *Early Access Online*.

78. Mortillaro, M., Meuleman, B., and Scherer, K. R. (2012). Advocating a Componential Appraisal Model to Guide Emotion Recognition: Int. J. Synth. Emot. *3*, 18–32.

79. Roseman, I. J., and Smith, C. A. (2001). Appraisal theory: Overview, assumptions, varieties, controversies. In Appraisal processes in emotion: Theory, methods, research Series in affective science., K. R. Scherer, A. Schorr, and T. Johnstone, eds. (New York, NY, US: Oxford University Press), pp. 3–19.

80. Scherer, K. R., and Meuleman, B. (2013). Human Emotion Experiences Can Be Predicted on Theoretical Grounds: Evidence from Verbal Labeling. PLoS ONE *8*, e58166.

81. Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting Human Brain Activity Associated with the Meanings of Nouns. Science *320*, 1191–1195.

82. Murphy, B., Talukdar, P., and Mitchell, T. Selecting Corpus-Semantic Models for Neurolinguistic Decoding.

83. Pereira, F., Detre, G., and Botvinick, M. (2011). Generating text from functional brain images. Front. Hum. Neurosci. *5*, 72.

84. Pang, B., and Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics ACL ’04. (Stroudsburg, PA, USA: Association for Computational Linguistics). Available at: http://dx.doi.org/10.3115/1218955.1218990 [Accessed March 30, 2015].

85. Tan, S., Cheng, X., Wang, Y., and Xu, H. (2009). Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis. In Advances in Information Retrieval Lecture Notes in Computer Science., M. Boughanem, C. Berrut, J. Mothe, and C. Soule-Dupuy, eds. (Springer Berlin Heidelberg), pp. 337–349. Available at: http://link.springer.com/chapter/10.1007/978-3-642-00958-7_31 [Accessed March 30, 2015].

86. Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. Behav. Brain Sci. *28*, 675–691.

87. Freiwald, W. A., Tsao, D. Y., and Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. Nat. Neurosci. *12*, 1187–1196.

88. Just, M. A., Cherkassky, V. L., Aryal, S., and Mitchell, T. M. (2010). A Neurosemantic Theory of Concrete Noun Representation Based on the Underlying Brain Codes. PLoS ONE *5*, e8622.

89. Koster-Hale, J., Bedny, M., and Saxe, R. (2014). Thinking about seeing: perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults. Cognition *133*, 65–78.

90. Koster-Hale, J., Saxe, R., Dungan, J., and Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. Proc. Natl. Acad. Sci., 201207992.

91. Sudre, G., Pomerleau, D., Palatucci, M., Wehbe, L., Fyshe, A., Salmelin, R., and Mitchell, T. (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. NeuroImage *62*, 451–463.

92. Laurence, S., and Margolis, E. (1999). Concepts and Cognitive Science. In Concepts: Core Readings, E. Margolis and S. Laurence, eds. (MIT), pp. 3–81.

93. Murphy, G. L., and Medin, D. L. (1985). The role of theories in conceptual coherence. Psychol. Rev. *92*, 289–316.

94. Baker, C. L., Saxe, R., and Tenenbaum, J. B. (2009). Action understanding as inverse planning. Cognition *113*, 329–349.

95. Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. Science *331*, 1279–1285.

96. Lindquist, K. A., and Barrett, L. F. (2008). Constructing emotion: the experience of fear as a conceptual act. Psychol. Sci. *19*, 898–903.

97. Lindquist, K. A., Gendron, M., Barrett, L. F., and Dickerson, B. C. (2014). Emotion perception, but not affect perception, is impaired with semantic memory loss. Emotion *14*, 375–387.

98. Baldassi, C., Alemi-Neissi, A., Pagan, M., DiCarlo, J. J., Zecchina, R., and Zoccolan, D. (2013). Shape Similarity, Better than Semantic Membership, Accounts for the Structure of Visual Object Representations in a Population of Monkey Inferotemporal Neurons. PLoS Comput Biol *9*, e1003167.

99. Carlson, T. A., Simmons, R. A., Kriegeskorte, N., and Slevc, L. R. (2013). The Emergence of Semantic Meaning in the Ventral Temporal Pathway. J. Cogn. Neurosci., 1–12.

100. Liu, N., Kriegeskorte, N., Mur, M., Hadj-Bouziane, F., Luh, W.-M., Tootell, R., and Ungerleider, L. (2013). Intrinsic Structure of Visual Exemplars and Category Representations in Macaque Brain. J. Vis. *13*, 674–674.

101. Palatucci, M., Pomerleau, D., Hinton, G. E., and Mitchell, T. M. (2009). Zero-shot Learning with Semantic Output Codes. In, pp. 1410–1418. Available at: http://machinelearning.wustl.edu/mlpapers/papers/NIPS2009_0395 [Accessed March 30, 2015].

102. Rust, N. C., and DiCarlo, J. J. (2010). Selectivity and Tolerance (“Invariance”) Both Increase as Visual Information Propagates from Cortical Area V4 to IT. J. Neurosci. *30*, 12978–12995.

103. Sudre, G., Pomerleau, D., Palatucci, M., Wehbe, L., Fyshe, A., Salmelin, R., and Mitchell, T. (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. NeuroImage *62*, 451–463.

104. Isik, L., Meyers, E. M., Leibo, J. Z., and Poggio, T. A. (2013). The dynamics of invariant object recognition in the human visual system. J. Neurophysiol., jn.00394.2013.

105. Cichy, R. M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. Nat. Neurosci. *17*, 455–462.

106. Leeds, D. D., Seibert, D. A., Pyles, J. A., and Tarr, M. J. (2013). Comparing visual representations across human fMRI and computational vision. J. Vis. *13*, 25.

107. Ahmad Babaeian Jelodar, M. A. (2010). WordNet based features for predicting brain activity associated with meanings of nouns. 18–26.

108. Chang, K. K., Mitchell, T., and Just, M. A. (2011). Quantitative modeling of the neural representation of objects: how semantic feature norms can account for fMRI activation. NeuroImage *56*, 716–727.

109. Huth, A. G., Nishimoto, S., Vu, A. T., and Gallant, J. L. (2012). A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. Neuron *76*, 1210–1224.

110. Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. *111*, 8619–8624.

111. Devereux, B., Kelly, C., and Korhonen, A. (2010). Using fMRI Activation to Conceptual Stimuli to Evaluate Methods for Extracting Conceptual Representations from Corpora. In Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics CN '10. (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 70–78. Available at: http://dl.acm.org/citation.cfm?id=1866686.1866695 [Accessed January 9, 2015].

112. Stansbury, D. E., Naselaris, T., and Gallant, J. L. (2013). Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex. Neuron *79*, 1025–1034.

113. Dodell-Feder, D., Koster-Hale, J., Bedny, M., and Saxe, R. (2011). fMRI item analysis in a theory of mind task. NeuroImage *55*, 705–712.

114. Knerr, S., Personnaz, L., and Dreyfus, G. (1990). Single-layer learning revisited: a stepwise procedure for building and training a neural network. In Neurocomputing NATO ASI Series., F. F. Soulié and J. Hérault, eds. (Springer Berlin Heidelberg), pp. 41–50. Available at: http://link.springer.com/chapter/10.1007/978-3-642-76153-9_5 [Accessed March 30, 2015].

115. Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. Science *310*, 863–866.

116. Seung, H. S., and Sompolinsky, H. (1993). Simple models for reading neuronal population codes. Proc. Natl. Acad. Sci. *90*, 10749–10753.

117. Shamir, M., and Sompolinsky, H. (2006). Implications of Neuronal Diversity on Population Coding. Neural Comput. *18*, 1951–1986.

118. Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. PLoS Comput. Biol. *10*. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3990488/ [Accessed November 21, 2014].

Figure and Table Legends

Table 1. Example stimuli.

Figure 1. MVPA classification results: A. Above chance 20-way classification of emotions in all ToM regions. B. Whole-brain random-effects analysis of ToM localizer (FB>FP, green); searchlight map for 20-way emotion classification (red); overlap (yellow).

Figure 2. Classification accuracy broken down by emotion: A. Average classification accuracy for each emotion condition (+/- SEM across exemplars) in behavioral judgments. B. Correlation between behavioral classification accuracies (from A) and neural classification accuracies for each emotion class (based on errors of an SVM trained and tested on DMPFC voxel patterns).

Figure 3. Competing behavioral feature spaces: Matrix of emotions x average dimension scores for A) the 38-dimensional appraisal space, B) the 6 basic emotion space, and C) the circumplex space. D. Classification of 20 emotions (across stimulus exemplars) using information from each of the 3 competing spaces (+/- SEM across exemplars). Orange dotted line reflects chance (.05); blue dotted line reflects behavioral performance (.65).

Figure 4. RSA Methods: Representational dissimilarity matrices (RDMs) encode the pairwise Euclidean distances between different emotions within each feature space. For each region, a neural RDM captures the pairwise Euclidean distances between different emotions in the patterns of activity elicited across voxels (DMPFC shown here). Feature spaces are fit to the neural data by computing correlations between feature space RDMs and neural RDMs for each region in each subject.

Figure 5. RSA Results: Mean correlation (kendall's tau) between model RDMs and individual subject neural RDMs (+/- SEM across subjects). Dotted line shows the correlation of a similarity space defined by the raw behavioral confusion matrix.

Figure 6. RSA Feature Results: For each region in each subject, the neural RDM was correlated with the full 38-dimensional space, the reduced space of 10 features, and with RDMs encoding each of the features individually. Plot shows the mean correlation between neural and model spaces (+/- SEM across subjects).

Figures and Tables

Table 1.

Stimulus Type	Example Stimulus
Emotion	After an 18-hour flight, Caitlin arrived at her vacation destination to learn that her baggage (including necessary camping gear for her trip) hadn't made the flight. After waiting at the airport for 2 nights, Caitlin was informed that the airline had lost her luggage altogether and wouldn't provide any compensation.
	For months, Naomi had been struggling to keep up with her various projects at work. One week, the company announced that they would be making massive payroll cuts. The next day, Naomi 's boss asked her to come into his office and close the door behind her.
	Linda was having financial difficulties after graduating from college. She worked overtime and lived very meagerly, but still had trouble making her loan payments. One day, she received a letter from her grandfather saying that he wanted to help. A check for $8,000 was enclosed.
	Dana always wanted a puppy, but her parents said it was too much of a hassle. One summer afternoon, Dana's parents returned from a supposed trip to the grocery store, and Dana heard barking from inside her garage. She opened the door to see her parents holding a golden retriever puppy.
Physical Pain	One afternoon, Caitlin was running through her house while playing tag with her friend. After going through a doorway, Caitlin slammed the door behind her, but her fingers were caught in the door. When they opened the door, two of her fingers were broken.

Figure 1

Click here to download Figure: Figure 1.jpg

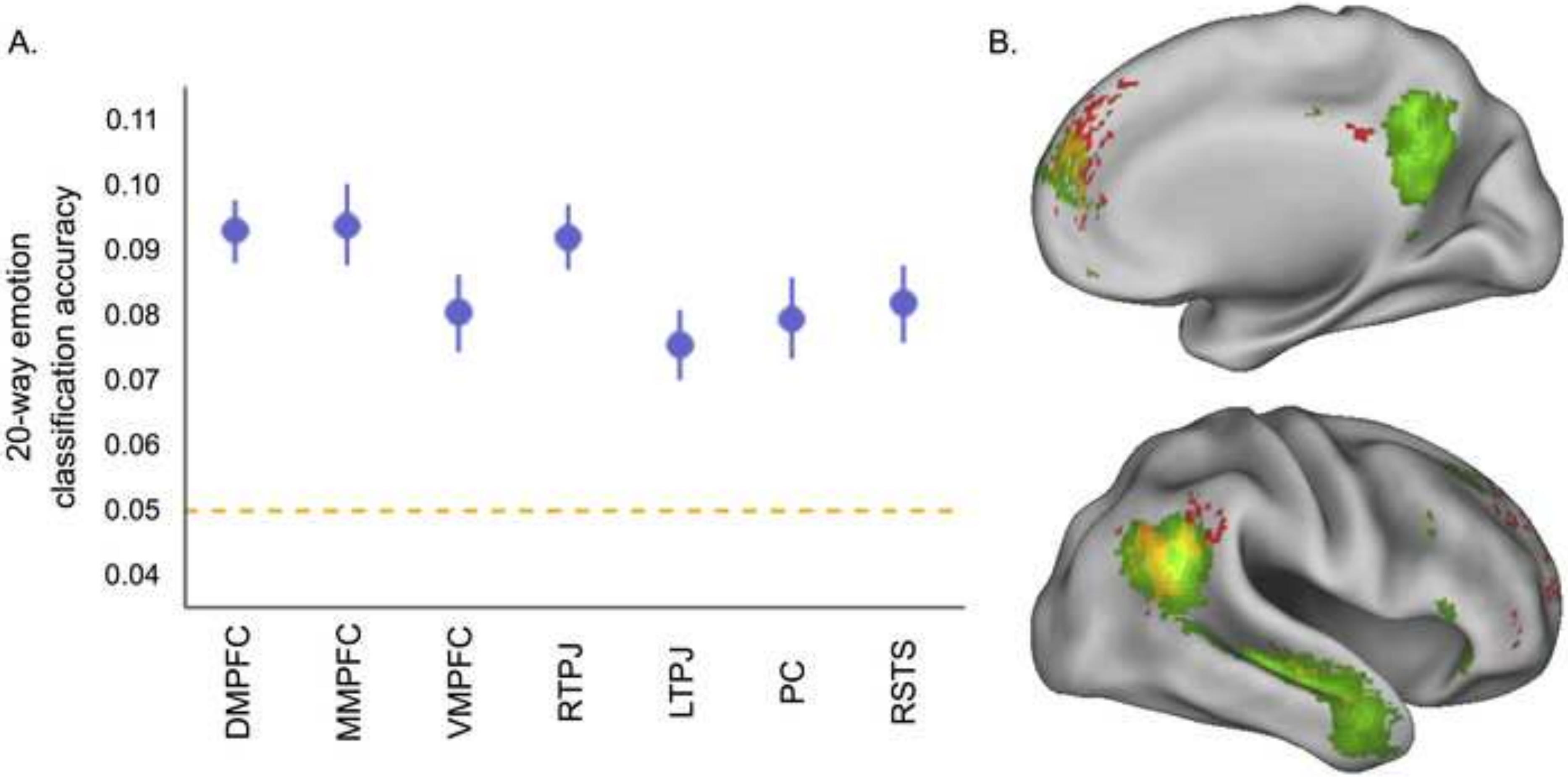


Figure 2

Click here to download Figure: Figure 2.jpg

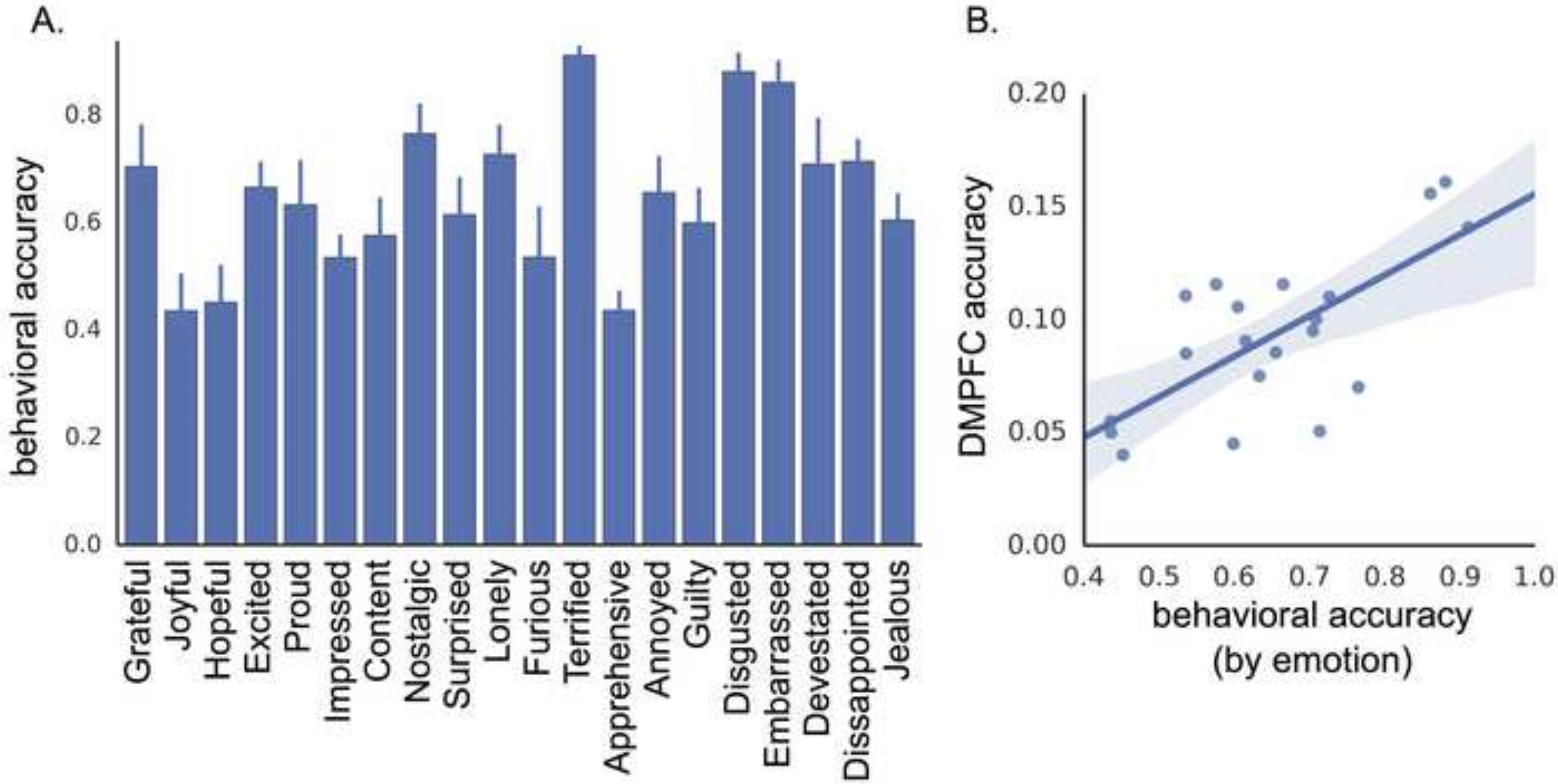


Figure 3

Click here to download Figure: Figure 3.jpg

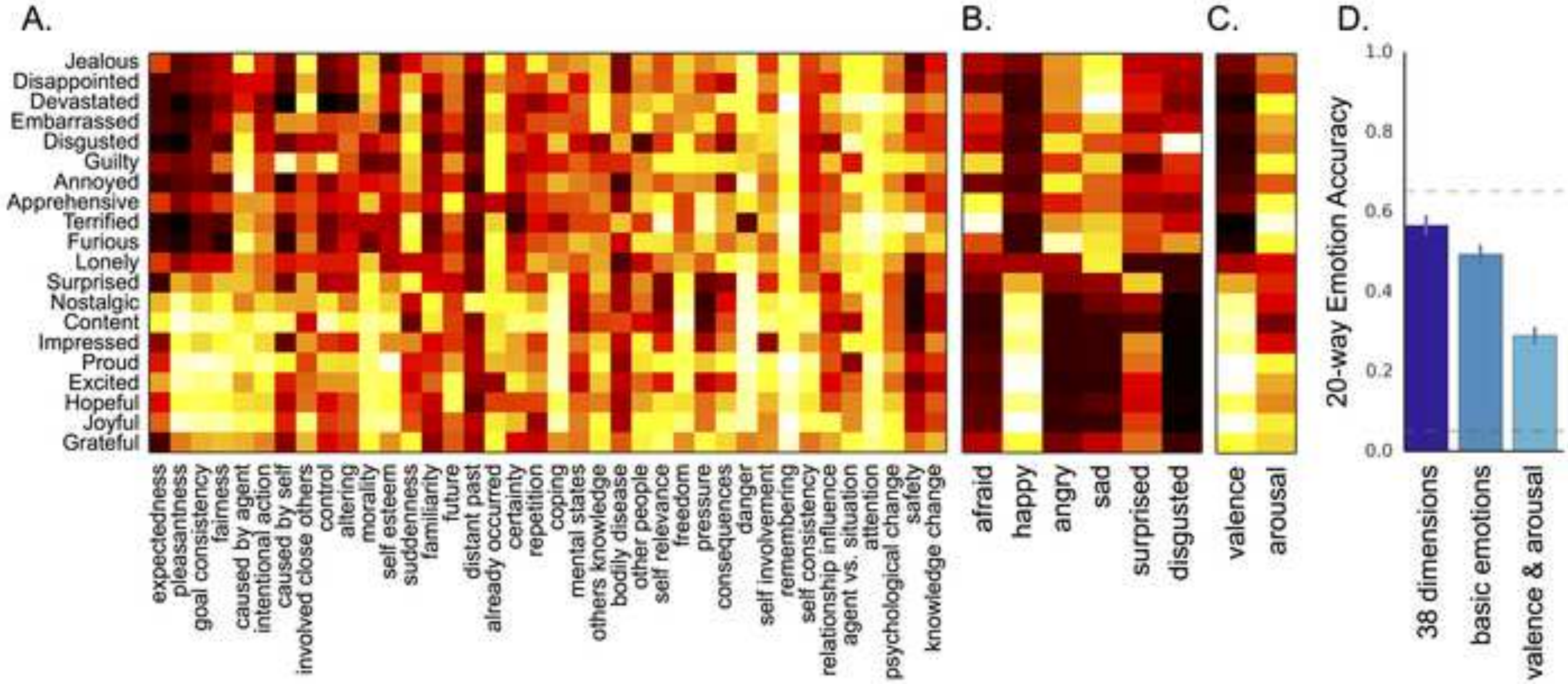


Figure 4

Click here to download Figure: Figure 4.jpg

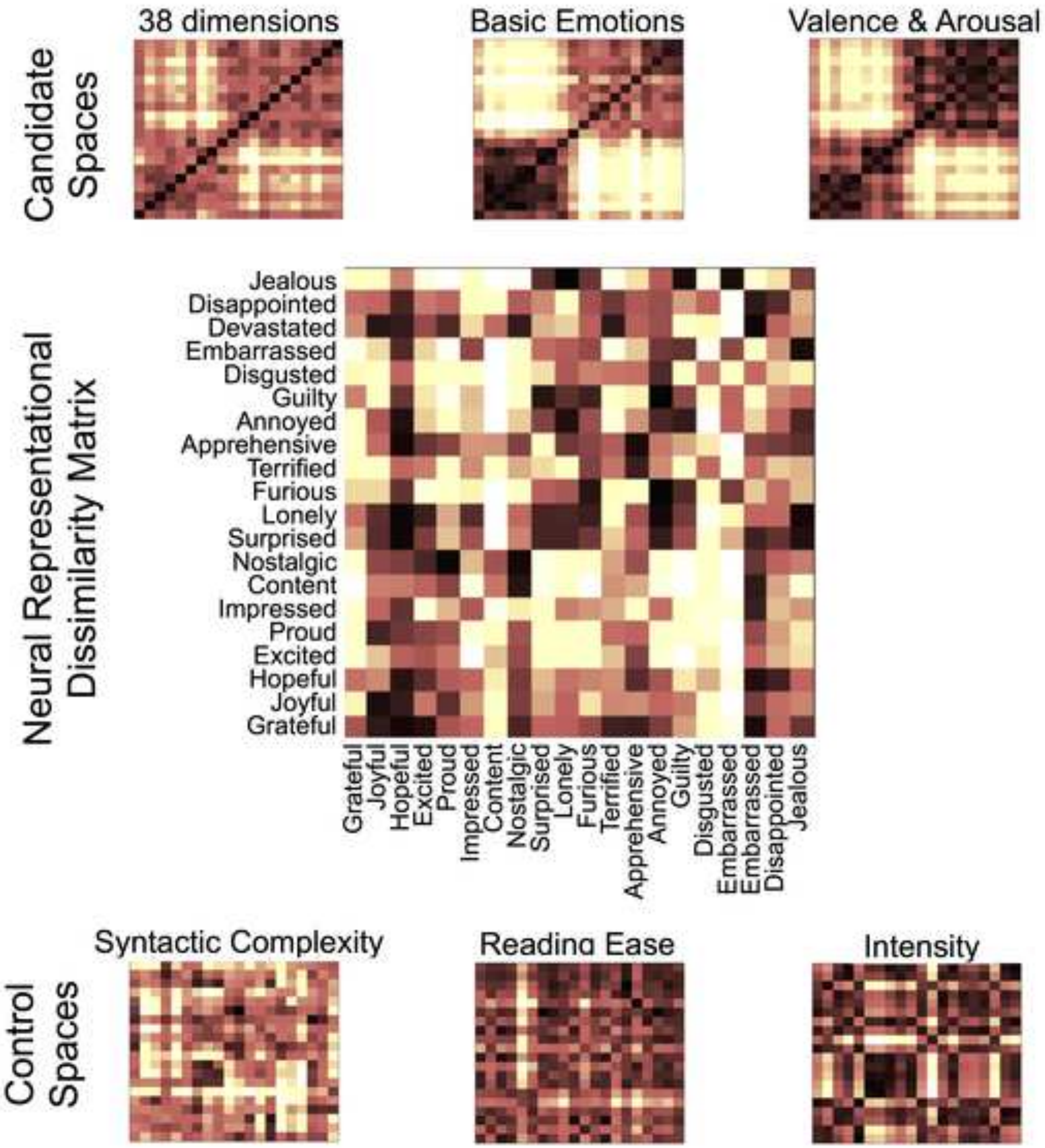


Figure 5

Click here to download Figure: Figure 5.jpg

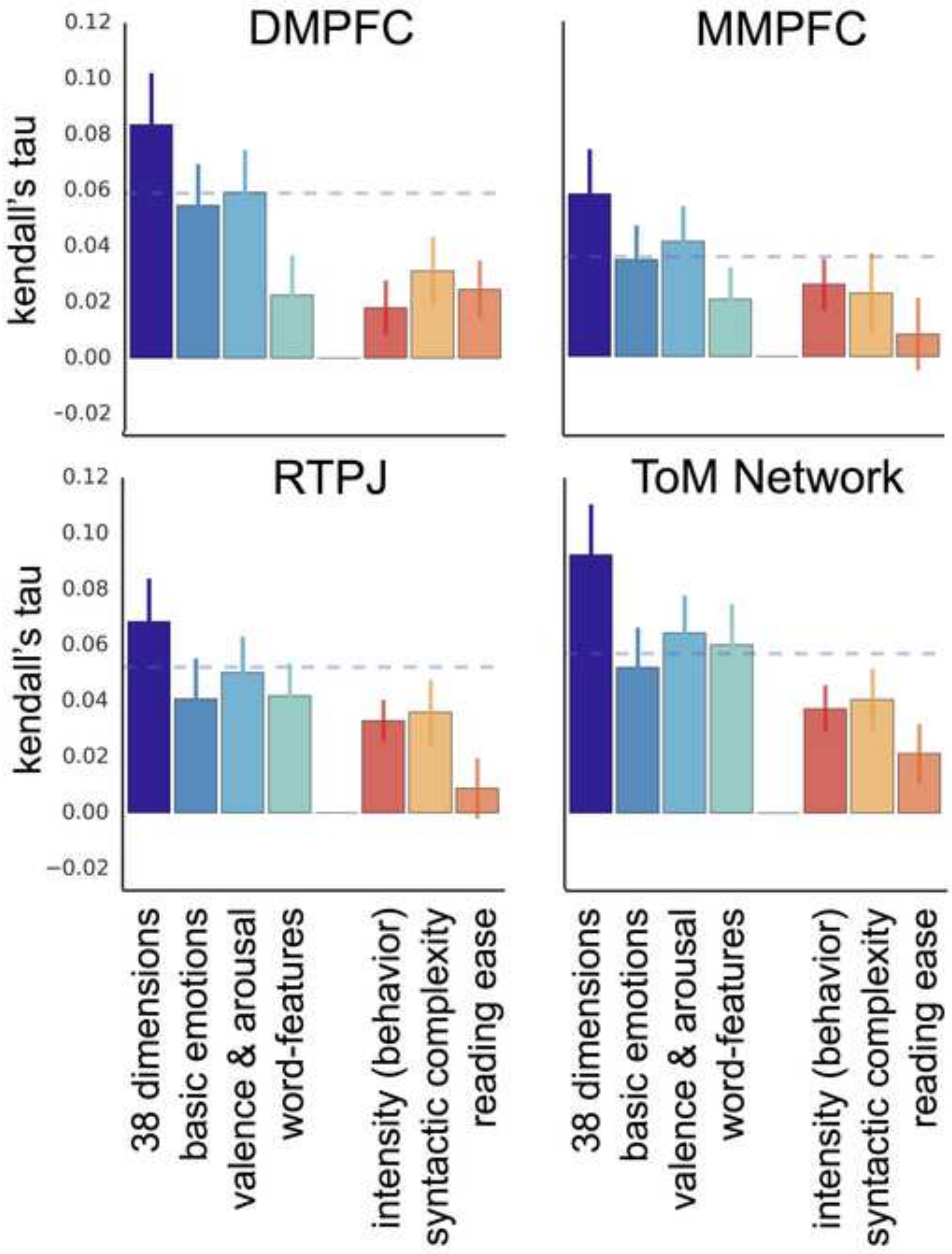
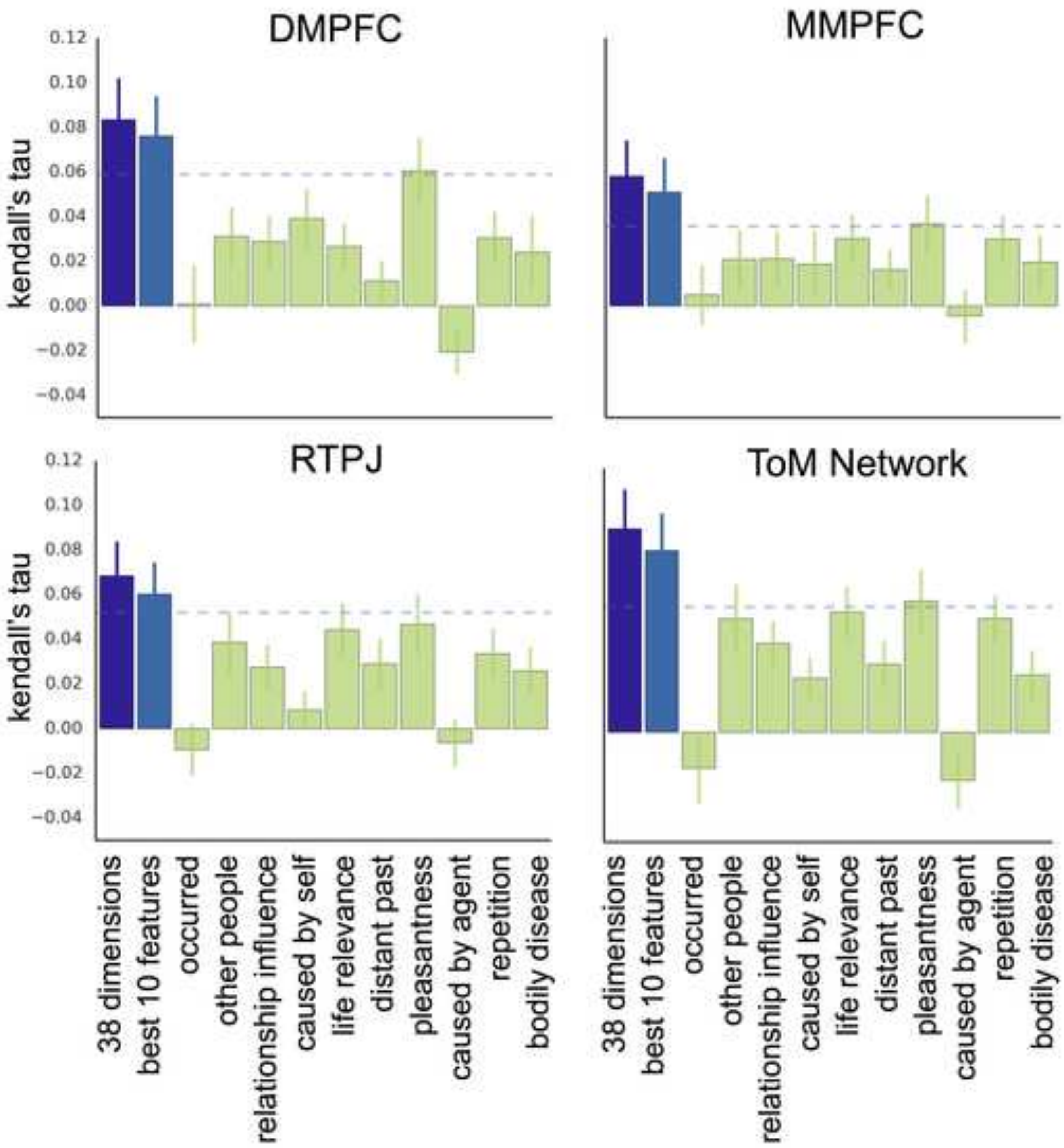


Figure 6

Click here to download Figure: Figure 6.jpg



SUPPLEMENTAL INFORMATION

Supplemental Data

Table S1. Results from whole brain searchlight for 20-way emotion classification ($p<.05$, FWE corrected, $k>25$)

# voxels	Peak T	X	Y	Z	Region
1204	11.91	12	60	18	DMPFC/MMPFC
527	12.27	50	-58	18	RTPJ
233	12.99	-46	-66	20	LTPJ
147	11.40	-46	-56	46	LIPL
107	9.83	54	2	-22	RSTS
53	10.55	-58	-18	-16	LSTS/LMTG
41	9.06	0	-34	32	PC
36	10.51	-40	24	32	LMFG
36	9.33	-48	36	12	LIFG
32	9.27	54	-24	-6	RSTS
32	8.83	36	48	0	RIFG/MFG
28	10.39	-2	-18	40	PC
26	9.40	-50	-36	42	LIPL
26	8.56	16	52	28	DMPFC

Supplemental Figure Legends

Figure S1. Univariate Analyses of ToM ROIs: We localized theory of mind regions in the majority of subjects using the localizer contrast FB>FP (DMPFC: 20 subjects, MMPFC: 21 subjects, VMPFC: 22 subjects, RTPJ: 22 subjects, LTPJ: 22 subjects, PC: 22 subjects, RSTS: 22 subjects). Previous results suggest that these regions are selectively involved in processing the mental states of other people relative to physical or bodily states (Bruneau et al, 2012; 2013). We confirmed the selectivity of localized ROIs by comparing the response to the emotion stimuli and physical pain stimuli. Plots of percent signal change in each of the ToM ROIs show the BOLD response to the 20 emotion stimuli (negative situations in yellow/green, positive situations in blue/pink) relative to the response to non-mental stimuli describing physical pain (dark red). Consistent with prior work, we found robustly higher response to the emotional items relative to physical pain items (average Emotion beta > Pain beta) in all regions: DMPFC: $t(19)=6.224$, $p<.001$; MMPFC: $t(20)=6.115$, $p<.001$; VMPFC: $t(21)=6.065$, $p<.001$; RTPJ: $t(21)=4.571$, $p<0.001$; LTPJ: emo>pain: $t(21)=8.085$, $p<0.001$;PC: $t(21)=7.620$, $p<0.001$; RSTS: $t(21)=5.182$, $p<0.001$)

Figure S2. A. Confusion matrix from behavioral stimulus categorizations performed by independent subjects on MTurk. B. Confusion matrices from neural classification analysis (linear SVM trained and tested on 20 emotion categories across runs).

Figure S3. The behavioral classification results show that different emotions have distinct profiles across the 38 appraisal dimensions. Plot shows examples of the feature averages (+/- SEM across items) for several different emotions.

Figure S4. Correlation matrix for 38-dimensional feature space. Pairwise correlations between individual appraisal features show high collinearities amongst certain features.

Figure S5. Reduced appraisal space: A. Stimulus-by-appraisal matrix reconstructed using the 10 selected features, capturing 75% of the variance of the original matrix of stimuli-by-38-dimensions. B. Classification of 20 emotions (generalizing across stimulus exemplars) using information from the full 38-dimensional space (dark blue), the space of 10 selected features (light blue), and each of those 10 features individually (green). A model trained to distinguish the 20 emotion labels using only the 10 features could classify the emotions of novel stimuli at 45% accuracy (compared to 57% observed with the full appraisal space), and the emotional discriminations were not captured by any of the individual features in isolation (all accuracies <20%).

Figure S6. Representational Similarity Analysis Results (secondary ROIs): Mean correlation (kendall's tau) between model RDMs and individual subject neural RDMs (+/- SEM across subjects). Dotted line shows the correlation of a similarity space defined by the raw behavioral confusion matrix. ROIs were all significantly correlated with the space of 38-appraials: LTPJ: M(SEM)=0.06(0.02), z(21)=2.97 p<0.001, PC: M(SEM)=0.05(0.01), z(21)=3.07 p<0.001, RSTS: M(SEM)=0.06(0.02), z(20)=2.97 p<0.001, VMPFC: M(SEM)=0.03(0.01), z(17)=1.81 p=0.035.

Figure S7. Representational Similarity Analysis Feature Results (secondary ROIs): Mean correlation between neural RDMs (from LTPJ, RSTS, VMPFC, and PC) and the RDM encoding the 38-dimensional space, an RDM encoding the reduced space of 10 features, and separate RDMs encoding each of the features individually (+/- SEM across subjects).

Figure S8. Representational Similarity Time-course Analysis: To explored the temporal profile of representation in each region, RSA analyses were conducted separately for overlapping 4s windows with onsets ranging from 0 to 11 TRs post stimulus presentation. This analysis reveals relatively comparable time-courses across ROIs; for example, the regions with the strongest correlations (DMPFC, MMPFC, and RTPJ) all exhibited a peak in similarity in the window 6-7 TRs post stimulus onset.

Supplemental Figures

Figure S1.

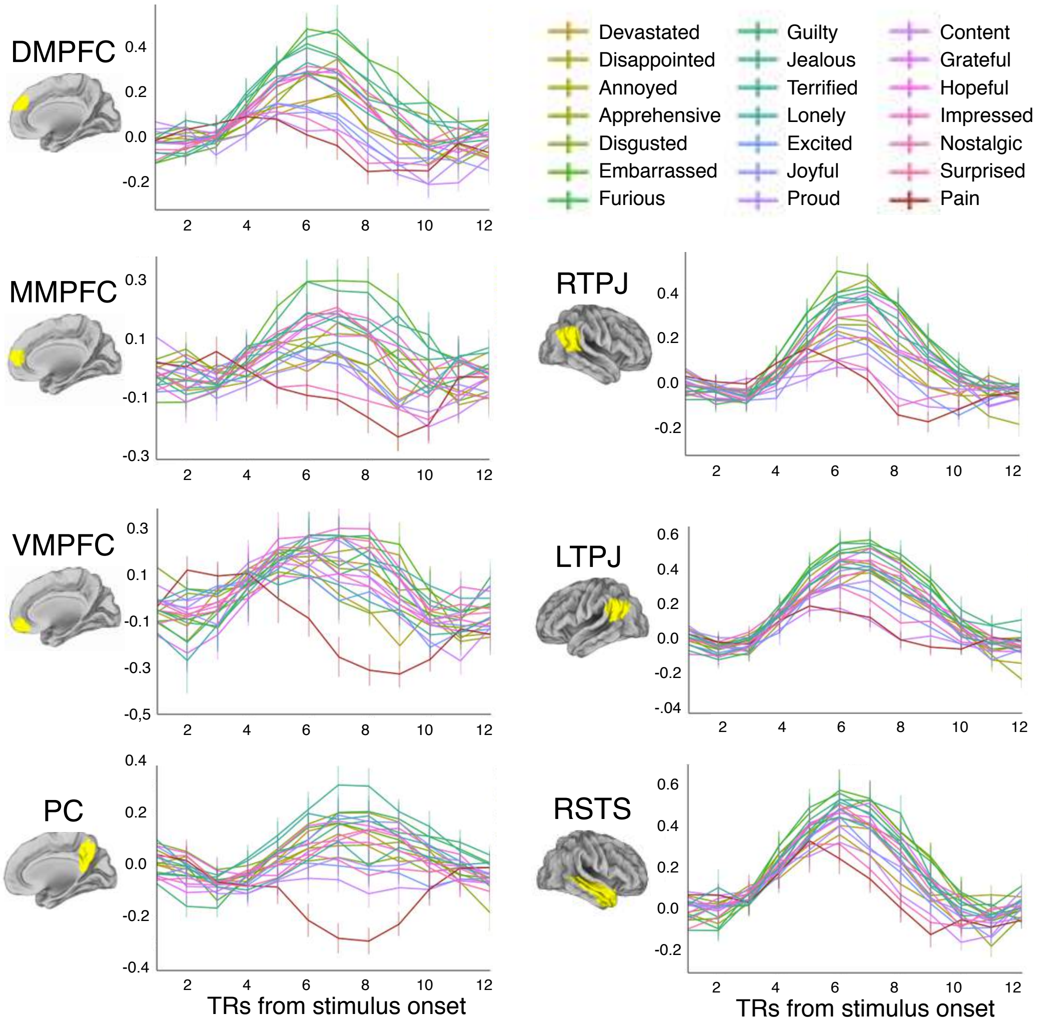


Figure S2.

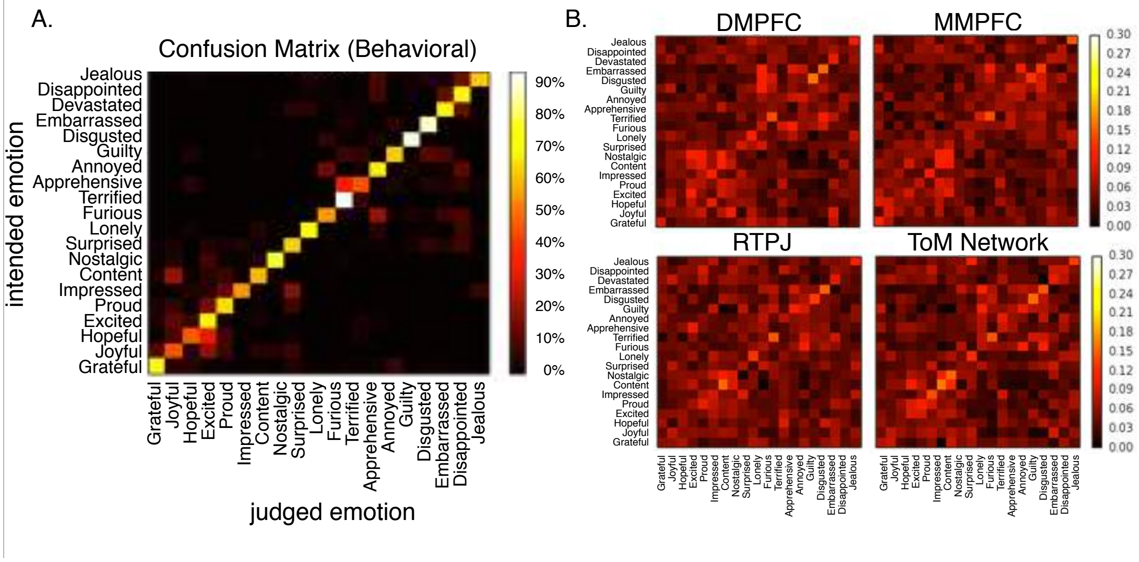


Figure S3.

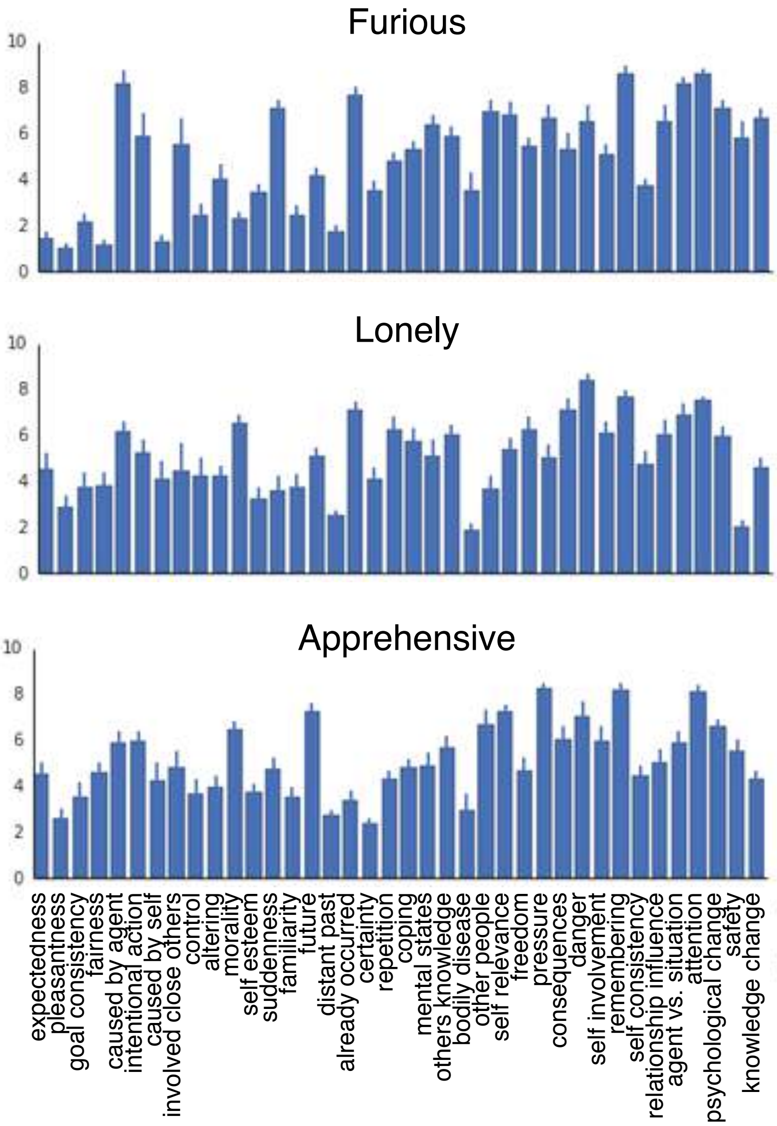


Figure S4.

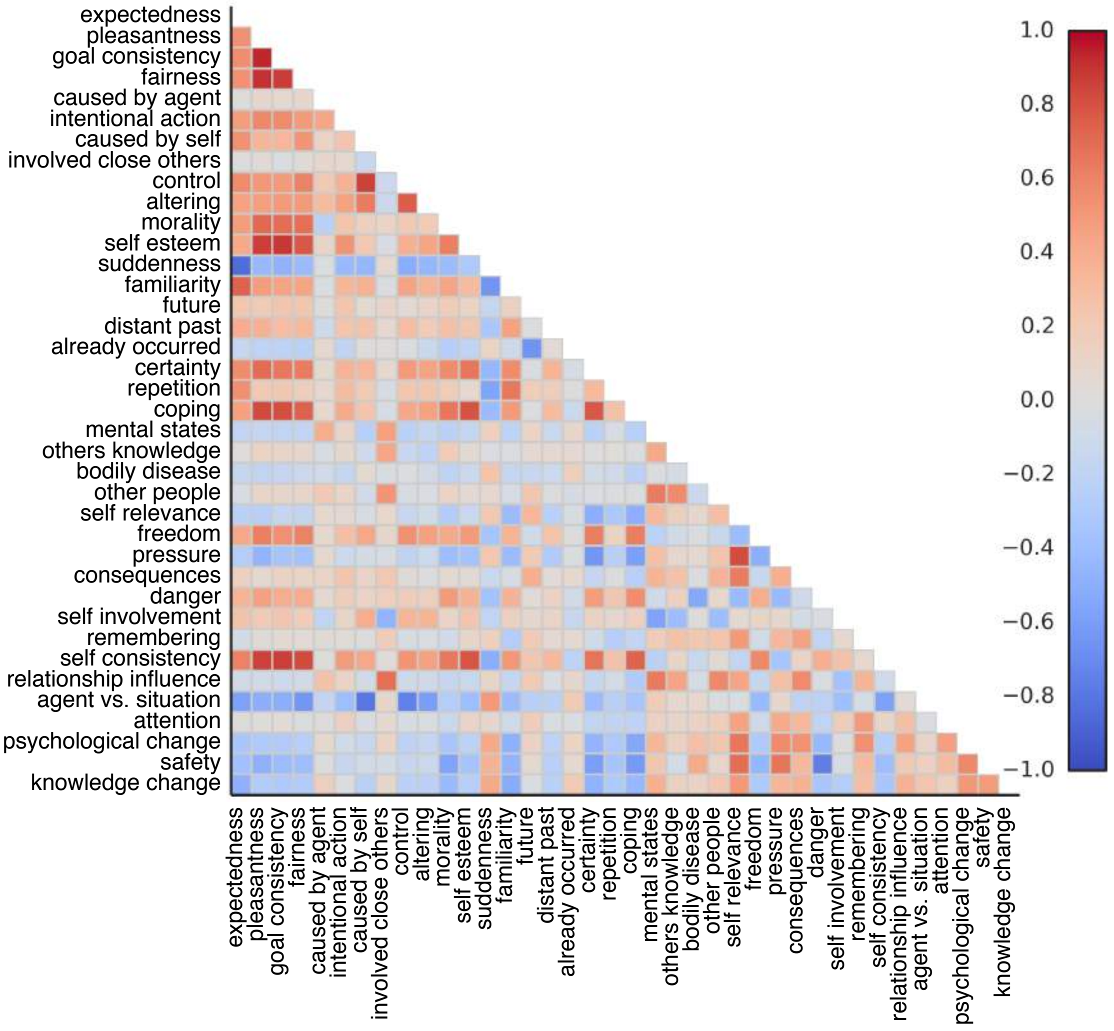


Figure S5.

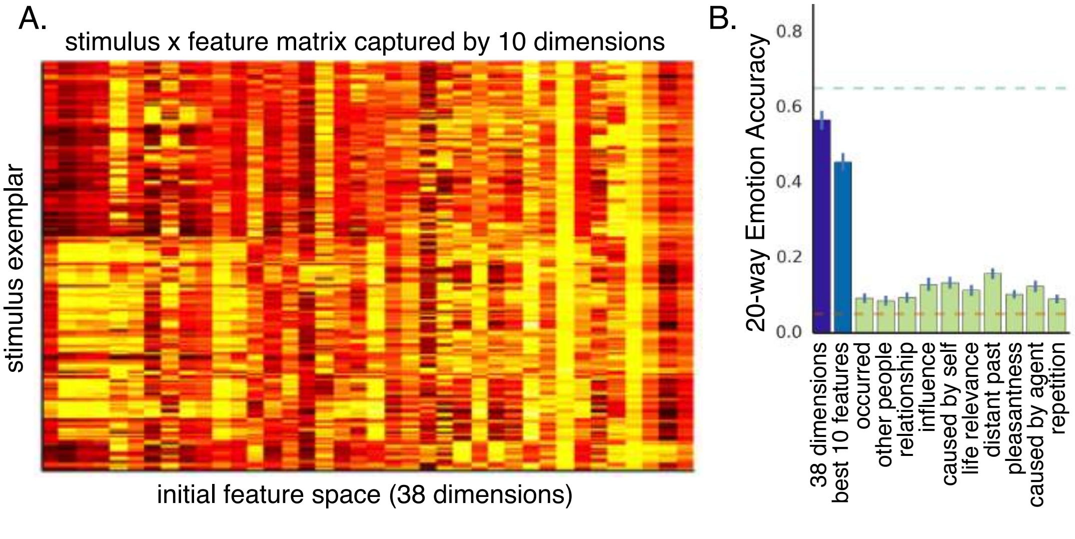


Figure S6.

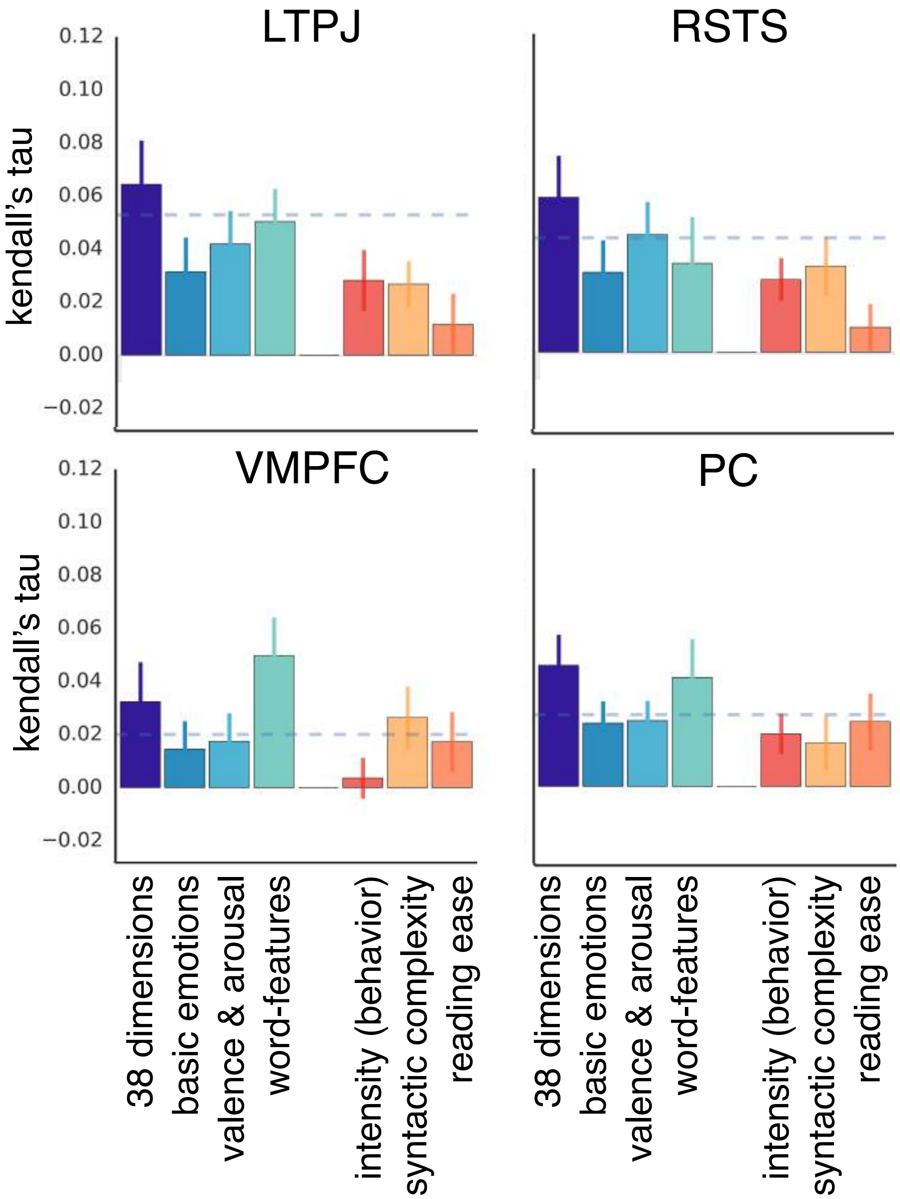

LTPJ
RSTS
VMPFC
PC
kendall's tau
kendall's tau
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−0.02
38 dimensions
basic emotions
valence & arousal
word-features
intensity (behavior)
syntactic complexity
reading ease

Figure S7.

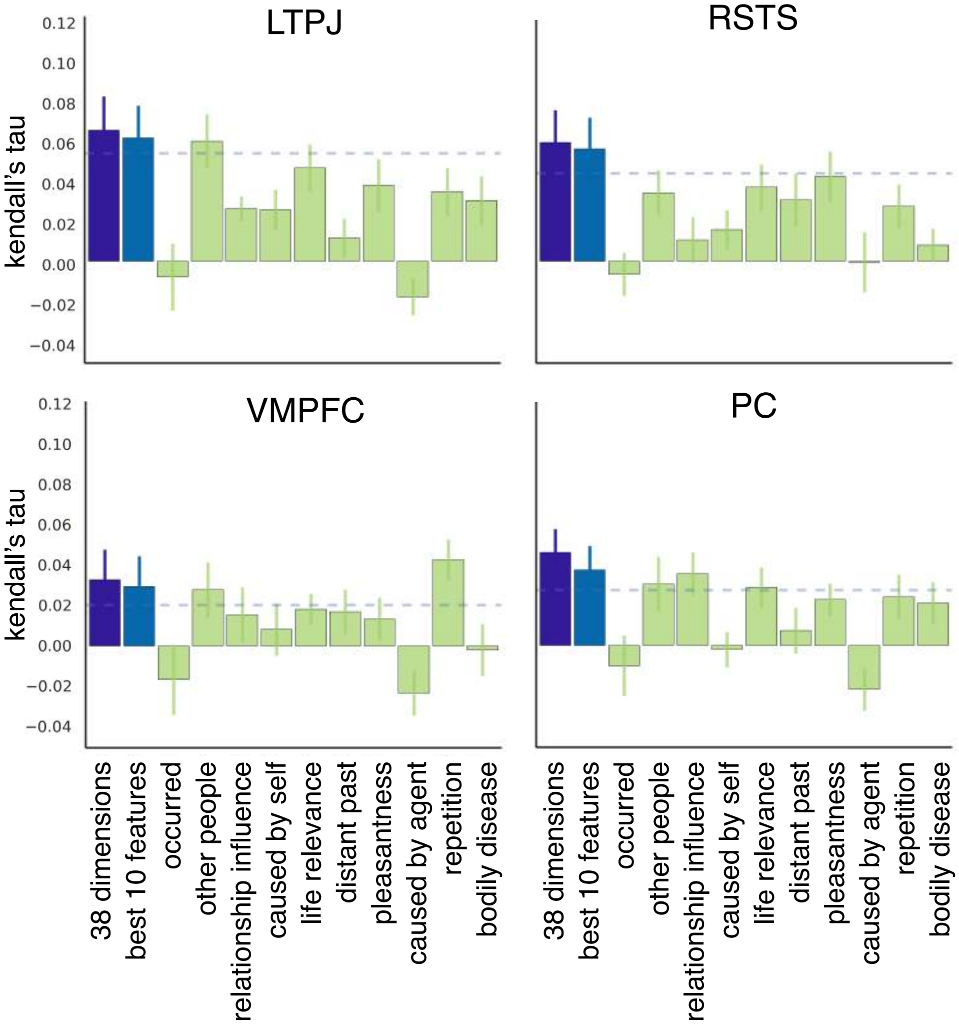
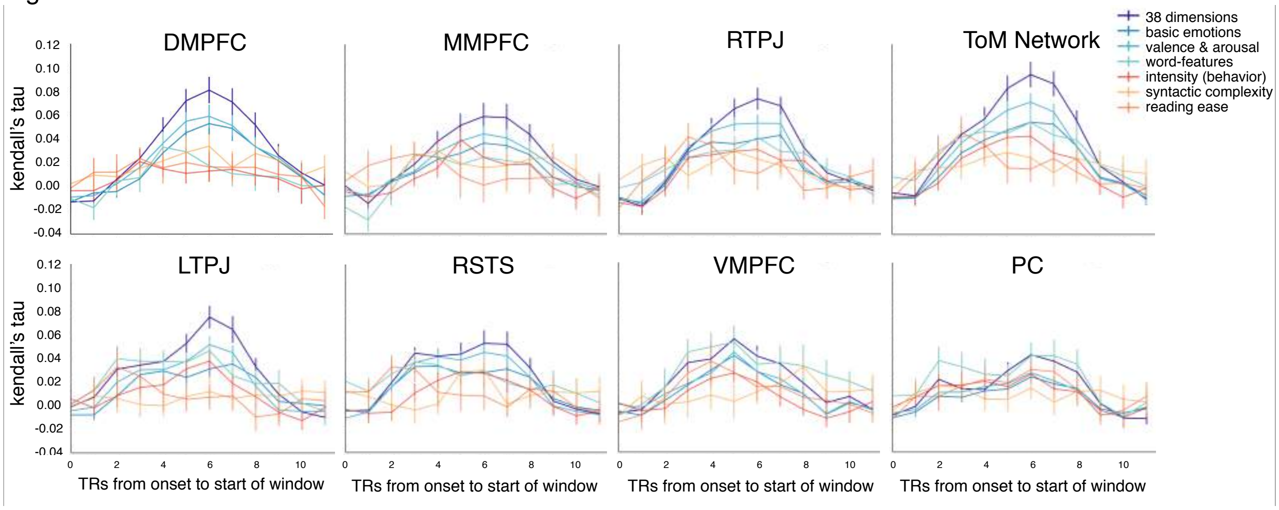


Figure S8.



Supplemental Experimental Procedures

38 features used to construct the abstract appraisal space (RSA analysis)

Feature Name	**Feature Question**
expectedness	Did <character> expect this situation to occur?
pleasantness	Did the situation involve a hedonically positive or pleasant experience for <character>?
goal consistency	Was the situation consistent or inconsistent with <character>'s goals, needs, or desires?
fairness	Was this situation fair or unfair for <character>?
caused by agent	Was this situation caused by a person or some other external force (e.g. randomness)?
intentional action	Did someone cause this situation intentionally or did it occur by accident?
caused by self	Was this situation caused by <character> herself or by someone/something else?
involved close others	Did this situation involve people that <character> felt close to?
control	Were the events in this situation primarily within <character>'s control?
altering	Did <character> think she had the power to alter the situation in the future?
morality	Did this situation involve people behaving in a way that would be considered proper or moral?
self esteem	Did this situation affect <character>'s self-esteem or opinion of herself?
suddenness	Did this situation occur suddenly/out of the blue?
familiarity	Was this situation a familiar event/situation for <character>?
future	Did <character>'s emotion involve an event that would or might occur in the future?
distant past	Did this situation involve events from <character>'s distant past?
already occurred	Was <character>'s emotion based on an something that had already occurred?
certainty	Did <character> feel certain about the situation/outcome?
repetition	Did <character> think the situation was likely to occur again?
coping	Did <character> think she could cope with/handle the situation?
mental states	Was <character>'s emotion related to the mental states (e.g. beliefs, attitudes) of other people?
others knowledge	Did people other than <character> know about the situation that occurred?
bodily disease	Did the situation involve events relating to the physical body?
other people	Was <character> interacting with people in this situation?
self relevance	Was there a lot at stake for <character> in this situation? Did the events have high-relevance for <character>'s life?
freedom	Was <character> free to act or behave however she wanted in this situation?
pressure	Was <character> under a lot of pressure in this situation?
consequences	Was <character>'s situation an isolated incident, or did it have long-term consequences?
danger	Was <character> in physical danger in this situation?
self involvement	Did the situation describe an outcome directly involving <character> herself or primarily involving other people?
remembering	Was the situation something that <character> is likely to remember in the future?
self consistency	Did the situation involve events consistent with <character>'s personality or self-concept?
relationship	Did this situation affect <character>'s relationships with other people?

influence	
agent vs. situation	Was this event primarily a reflection of <character> (e.g. her personality, her abilities) or a reflection of the surrounding situation?
attention	How much of <character>'s attention did this situation occupy?
psychological change	Did this situation involve a change in <character>'s psychological state?
safety	Did this situation involve risks for <character> or others?
knowledge change	Did this situation involve a change in <character>'s knowledge or belief about something?

Behavioral data acquisition for Classification Analysis: Behavioral data were collected on Amazon's Mechanical Turk. To obtain behavioral classification for each of the 200 stimuli, a set of subjects were presented with a single story on each trial and asked to choose which emotion (one of the 20 emotion categories, plus Neutral) best described the emotional state of the character. In addition to the stimuli from the 20 emotion categories, subjects were presented two stories in which the character was explicitly described as feeling neutral, which was used as an attention/quality check. We obtained judgments from 172 subjects (96 female; $M(SEM)_{age}$=33.326(0.873), and reduced the sample to 139 subjects who passed the quality check questions (by rating the neutral stories as neutral).

Behavioral data acquisition for RSA analysis: To construct RDMs, a separate set of subjects was used to obtain ratings of the stimuli within each of our feature spaces. MTurk subjects (n=250) were presented with a single stimulus item and used a 1 to 10 scale to rate the extent to which the event contained each feature in a given space. As an attention check, subjects were asked to rate the extent to which the story involved the character named in the story (subjects were excluded if their average response to this item was <7, and individual items were excluded if the response on this question was <5). 22 subjects were excluded for failing the attention check, leaving 238 subjects (108 female; $M(SEM)_{age}$= 34.47(0.77)) and an average of 15.4 responses for each of the 200 items). Subjects were allowed to rate more than one stimulus, and a given subject rated stimuli either on features from the 38-dimensional appraisal space (e.g. "Did someone cause this situation intentionally or did it occur by accident?"), or on dimensions corresponding to the basic emotion space (e.g. "What <character> happy in this situation") and the circumplex space ("Did <character> find this situation to be positive or negative?").

Region of interest selection: To define individual ToM ROIs, we used hypothesis spaces for bilaterial TPJ, right STS, PC, ventral, dorsal and middle subregions of MPFC, which were derived from previous random effects analyses with this task (see Figure S1 for hypothesis spaces). The task was modeled as a 14s boxcar (the full length of the story and question period) convolved with a standard hemodynamic response function, and a general linear model (implemented in SPM8) was used to estimate beta values for Belief trials and Photo trials. We conducted high-pass filtering at 128hz, normalized the global mean signal, and included nuisance covariates to remove effects of run. For each subject, we used a t-test implemented in SPM8 to generate a map of t values for the contrast of Belief>Photo and 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 9mm sphere surrounding this peak. If no cluster was found at p<0.001, we repeated this procedure at p<0.01 and p<.05 (see Figure S1). 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.

Univariate analyses of mental state selectivity in ToM ROIs: To confirm the selectivity of the individually localized ROIs, we compared the average BOLD response to the emotion stimuli to the average response to non-mental stories describing events in which a character experienced physical pain. This task was modeled as a boxcar (the full length of the story and response period) convolved with a standard HRF. We conducted high-pass filtering at 128hz, normalized the global mean signal, and included nuisance covariates to remove effects of run. We computed for each ROI the average beta value across voxels for each condition, and then averaged the beta values for the different emotion conditions to compare to the response to pain stimuli. In each ROI, we conducted a paired sample t-test comparing the beta values for emotion and pain conditions (see Figure S1)

Percent signal change in ToM ROIs: To visualize the univariate response to all conditions in these regions (see Figure S1), we computed the percent signal change (PSC) relative to baseline for each of the 21 conditions (20 emotions, plus physical pain). Baseline response for each ROI was computed as the average BOLD response at all rest time points, excluding the first 4s after stimulus offset for each trial. The PSC relative to baseline was calculated for each time point in each condition, averaging across all voxels in the ROI and across all trials in the condition, where PSC (at time t) = (average BOLD for condition at time t – average BOLD for fixation)/average BOLD magnitude for fixation. We plot this event-related average to visualize the average BOLD response at each time point after the trial onset for each condition.

Construction of control feature spaces for RSA analysis: To control for possible confounding sources of variation in our stimulus set, we computed three control spaces: reading ease, syntactic complexity, and intensity as rated by subjects in the scanner (confounded with motor response). The similarity spaces for reading ease and syntactic complexity were both derived from features extracted using CohMetrix (www.cohmetrix.com). Reading ease was made up of each exemplar's scores on Flesch Reading Ease (measuring average sentence length and number of syllables per word, where higher scores indicates easier text and increased readability) and Standardized Cohesion (measuring extent to which words overlap across the text; text with low referential cohesion is usually more difficult to process as there are fewer repetitions connecting ideas across the text). Syntactic complexity was defined in terms of Negation (measuring the number of negative expressions in the text, such as no, not, un-, without), Noun Phrase Modification (measuring the average number of modifiers, such as adjectives, adverbs, and determiners, per noun phrase), and Left-Embeddedness (measuring the average number of words before the main verb in each sentence). Finally, we computed the similarity of each stimulus in terms of its average intensity, derived from the in scanner behavioral judgments (1=neutral/low intensity, 4=extreme/high intensity). This RDM allows us to control for neural patterns reflecting possible differences in motor responses across emotion conditions.

Feature selection for region comparison analysis: To characterize the representation of specific appraisal features, we identified a reduced set of features that capture unique variance across stimuli (i.e. to eliminate redundant features or features that do not reliably vary across the stimuli). Computing pairwise correlations between each of the features (see Figure S4), we observe substantial correlations amongst features, suggesting that a smaller feature space may be sufficient. While many approaches to dimensionality reduction involve transforming data into linear combinations of the initial features, we wished to maintain the interpretability of our semantically meaningful features, and therefore used a forward step-wise regression procedure. On each of 38 iterations, we computed a separate regression fitting each of the available appraisal dimensions to the data (original 200 stimuli x 38 dimension matrix), yielding a separate R^2 value for each dimension. We then identified the appraisal dimension with the highest R^2 for that iteration and added it to an ordered list of appraisals. We then performed a regression on the initial data matrix using that dimension and the previously selected appraisals, with the R^2 characterizing the variance explained in the initial data matrix that can by this reduced set of features. The residuals of this regression (observed-predicted) served as the data matrix for the subsequent iteration. Thus, this procedure selects appraisal features that explain unique variance in the data. The first 10 appraisal features extracted with this method together capture 75.95% of the variance in the full input matrix, and are used as a reduced space. With the behavioral data, we perform 20-way classification using a feature vector in this 10-dimensional space, and using each of the 10 features in isolation (see Figure 6).

Analysis of RSA time-courses for region comparison: We also explored the temporal profile of representation in each region (see [1]) by computing RSA time-courses for each region. To do so, we separately analyzed a series of overlapping 2 TR (4 sec) windows, with onsets ranging from 0 to 11 TRs post stimulus presentation. For each window, we conducted the RSA analyses described above (compute neural RDM within that temporal window, and compute kendall's tau between the neural RDM and the model RDM for that time period). We can then plot these neural-model correlations over time to identify differences in the time-course of similarities across different region.

Behavioral individual difference measures: To assess the relevance of emotion-specific neural patterns to emotional and social competence, we collected behavioral measures of empathy and emotion recognition abilities and sought to relate these to individual differences in neural classification accuracy. The AQ [2] and the Interpersonal Reactivity Index (IRI, [3]) were completed via online Qualtrics surveys (www.qualtrics.com), yielding for each participant a single AQ score, and separate IRI scores for Empathic Concern (EC), Fantasy (FS), Personal Distress (PD), and Perspective Taking (PT). Participants also completed an Empathic Accuracy task [4] in which they made continuous ratings of a target's emotional state over the course of 16 trials. Empathic accuracy was operationalized as the correlation between subjects' emotion ratings and ratings generating by the individuals who recorded the stimuli. We tested for relationships between the behavioral measures of interest (Empathic Accuracy, AQ score, and IRI-Empathic Concern score) and individual subject classification accuracies in each of the ROIs by computing Pearson correlations (testing for positive relationship between neural classification accuracy and behavioral measures of EA and IRI-EC and negative relationships between classification accuracy and AQ score). No reliable relationships were observed.

Supplemental References

1. Cichy, R. M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. Nat. Neurosci. *17*, 455–462.

2. Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., and Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Malesand Females, Scientists and Mathematicians. J. Autism Dev. Disord. *31*, 5–17.

3. Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. J. Pers. Soc. Psychol. *44*, 113–126.

4. Zaki, J., Bolger, N., and Ochsner, K. (2008). It Takes Two The Interpersonal Nature of Empathic Accuracy. Psychol. Sci. *19*, 399–404.