Article

Current Biology

Neural Representations of Emotion Are Organized around Abstract Event Features

Highlights

- Patterns in ToM brain regions represent subtle emotion attributions
- These emotion attributions are well captured by a space of abstract event features
- This space outperforms competitors in capturing representations in ToM regions
- These neural representations are not reducible to primitive dimensions like valence

Authors

Amy E. Skerry, Rebecca Saxe

Correspondence

amy.skerry@gmail.com

In Brief

Skerry and Saxe find patterns of neural activity representing fine-grained emotional attributions, well captured by a space of abstract event features. These findings show that it is possible to characterize the detailed representational structure of an essential human reasoning capacity—the ability to infer the emotional states of others.

Skerry and Saxe, 2015, Current Biology *25*, 1–10
August 3, 2015
http://dx.doi.org/10.1016/j.cub.2015.06.009

CellPress

Please cite this article in press as: Skerry and Saxe, Neural Representations of Emotion Are Organized around Abstract Event Features, Current Biology (2015), http://dx.doi.org/10.1016/j.cub.2015.06.009

Current Biology
Article

Neural Representations of Emotion Are Organized around Abstract Event Features

Amy E. Skerry[1,*] and Rebecca Saxe[2]
[1]Department of Psychology, Harvard University, Cambridge, MA 02138, USA
[2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
*Correspondence: amy.skerry@gmail.com
http://dx.doi.org/10.1016/j.cub.2015.06.009

SUMMARY

Research on emotion attribution has tended to focus on the perception of overt expressions of at most five or six 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

Others' emotional states can be identified by diverse cues including facial expressions [1], vocalizations [2], or body posture [3]. However, we can also attribute subtle emotions based solely on the situation a person encounters [4, 5], and our vocabulary for attributing these states extends beyond the emotions associated with canonical emotional displays [6]. While the space of emotions perceived in faces has been studied extensively [7–9], little is known about how conceptual knowledge of others' emotions is organized, or how that knowledge is encoded in the human brain. What neural mechanisms underlie fine-grained attributions (e.g., distinguishing when someone will feel angry versus disappointed)? Here, we suggest that emotion attribution recruits a rich theory of the causal context of different emotions and show that dimensions of this intuitive knowledge underlie emotion representations in brain regions associated with theory of mind (ToM).

Previous research suggests that others' emotions are represented at varying levels of abstraction throughout face-selective and ToM brain regions. For example, different facial expressions elicit discriminable patterns of activity in the superior temporal sulcus (STS) and fusiform gyrus [10, 11]. In contrast, the medial prefrontal cortex (MPFC) has been shown to contain representations of emotion that are invariant to perceptual modality [12, 13] and generalize to emotions inferred in the absence of any overt display [14]. However, all of these studies focused on coarse distinctions, decoding either valence [14] or five basic emotions [13]. Does the MPFC also support more fine-grained emotional discriminations? To address this question, we constructed verbal stimuli (see Table 1) describing situations that would elicit 1 of 20 different emotions in a character (validated using 20-alternative-forced-choice [AFC] behavioral experiment with independent subjects; see Supplemental Experimental Procedures) and used multi-voxel pattern analysis [15] to test which regions contain information about these subtle emotional distinctions.

As a first step, we trained a classifier to distinguish the 20 emotions using distributed patterns of activity across voxels in a region and tested 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 focused on a priori regions of interest (ROIs), the strongest candidates being subregions of MPFC—dorsal medial prefrontal cortex (DMPFC) and middle medial prefrontal cortex (MMPFC) [13, 14]. We also tested other regions of the ToM network [16]: precuneus (PC), bilateral temporal parietal junction (TPJ), and right STS (RSTS).

We then used representational similarity analysis (RSA; [17]) to test competing hypotheses about the representational spaces in these regions. RSA complements classification analyses by providing a framework for characterizing representational structure and for testing competing models of that structure [17, 18]. In RSA, neural population codes are represented in terms of the similarity of neural patterns elicited by different stimuli or conditions. A neural representational dissimilarity matrix (RDM) of the conditions can then be compared to the similarity spaces captured by a number of different models [18, 19]. Importantly, RSA allows for comparison of hypotheses that take different forms and have different numbers of parameters. The correlation between model and neural RDMs has no free parameters, meaning that a model will not provide a better fit to the data simply

Table 1. Example Stimuli

Stimulus Type	Example Stimulus
Emotion	After an 18-hr flight, Caitlin arrived at her vacation destination to learn that her baggage (including necessary camping gear for her trip) had not made the flight. After waiting at the airport for two nights, Caitlin was informed that the airline had lost her luggage altogether and would not 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.

All experiments used the same set of 200 verbal stimuli in which a character experienced 1 of 20 different emotions (validated with 20-AFC experiment on MTurk), conveyed via a description of an emotion-eliciting event (see Supplemental Experimental Procedures).

because it is higher dimensional. Thus, RSA can go beyond classification to test specific alternative models of the dimensions that structure the representation of others' emotions.

Candidate Feature Spaces for Emotion Inference

Research in affective neuroscience has typically examined representations involved in both first-person emotional experience and emotional face perception. Here, we address a different question, concerning observers' inferences about others' emotions. Nevertheless, it is plausible that intuitive theories of emotion are fairly veridical (in order to be maximally useful in social interactions) and even informed by one's own emotional experiences. Therefore, models of the structure of first-person emotional experience may also capture the basis for third-person emotion attribution. We drew from prior literature on emotional experience three alternative models of the representational space of emotions.

A dominant approach has been to represent emotions as combinations of more basic affective states. According to basic emotion theory, complex and subtle emotions can be understood as combinations of 5–6 basic emotional states, each associated with a prototypical facial expression and innate neural substrate [1, 20, 21]. A second theory is the "circumplex" model, which posits that emotions are composed of only two primitive dimensions—valence and arousal [9, 22, 23]—corresponding to two innate systems implemented in distinct neural circuits and recruited to varying degrees across different emotions [24–26]. In this view, neural representations of emotion may be reduced to a linear combination of these two neurophysiological dimensions [27].

Although many have focused on the differences between these two proposals [28, 29], both aim to represent emotions in terms of combinations of a small number of basic affective states, rooted in innate neural substrates. An alternative approach in affective science, termed "appraisal theory," aims to instead characterize emotions in terms of people's interpretations or "appraisals" of the events around them [30, 31]. Researchers have proposed specific sets of event appraisals that correspond to different emotions (see Supplemental Experimental Procedures for further details) and shown that these features capture differences in the emotions subjects experience across different situations [32, 33].

All three of these theories have shown some utility in characterizing first-person emotional experiences. Here, we investigated whether any of these approaches successfully capture subjects' intuitive attributions of others' emotions and whether they could explain the representational spaces in MPFC and other ToM regions. If people reason about others' emotions using an intuitive causal theory (embedded in a larger intuitive ToM), this theory should capture regularities in the situations that cause different emotions. Thus, we hypothesized that the representations involved in inferring the emotions of others, especially based on short verbal narratives, would be better captured by abstract event features than by combinations of basic emotional dimensions.

We therefore used RSA to determine whether representations in regions that discriminate our 20 categories are best captured by one of three candidate spaces (see Figure 2): a "circumplex" space defined by independent subjects' judgments (Amazon Mechanical Turk [MTurk]; see Supplemental Experimental Procedures) of valence and arousal for each stimulus, a "basic emotion" space defined by judgments of the extent to which the stimulus elicited each of six basic emotions (happy, sad, angry, afraid, disgusted, or surprised), and a space of abstract event features derived from appraisal theory. For this third model, we generated a set of 38 event features thought to reliably vary across different emotion concepts (e.g., "Did someone cause this situation intentionally, or did it occur by accident?"; See Supplemental Experimental Procedures for appraisal features and selection process). Importantly, the latter space differs from the other two not only in its dimensionality (38 dimensions versus 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. To test which feature space best explains the neural representation of these stimuli, we computed the similarity of emotion conditions within each proposed feature space and compared the RDMs of candidate models to neural RDMs derived from patterns of activity across voxels in each ROI.

Of course, the hypothesis that neural representations of emotion concepts are best captured by a high-dimensional space of abstract event features is not incompatible with the

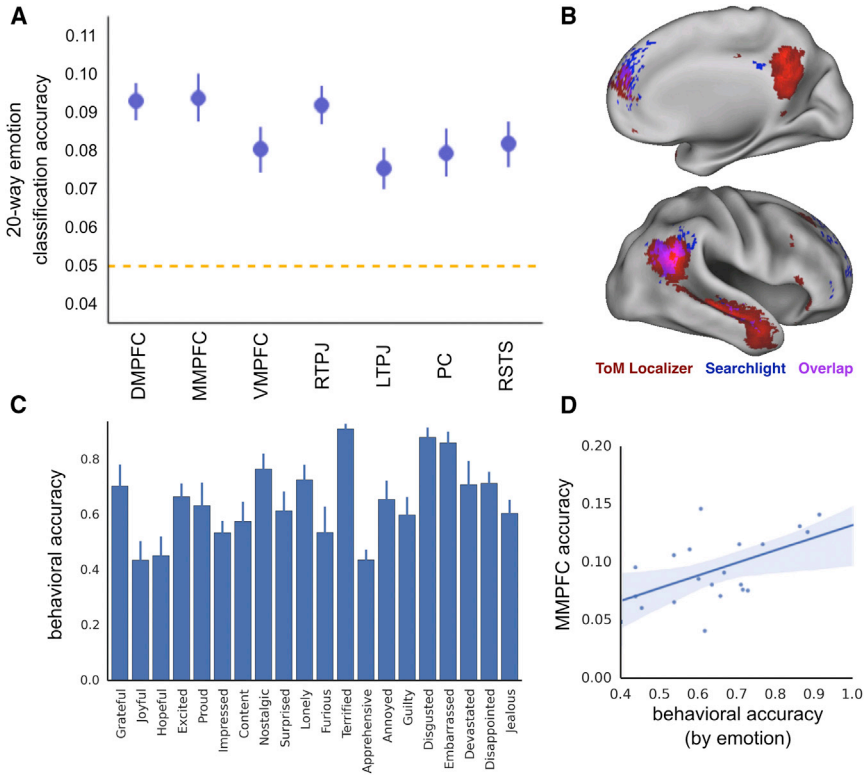

Figure 1. Classification Results

(A) Above-chance 20-way classification of emotions in all ToM regions.

(B) Whole-brain random-effects analysis of ToM localizer (false belief > false photo, red); searchlight map for 20-way emotion classification (blue); overlap (purple).

(C) Classification accuracy broken down by emotion: average classification accuracy for each emotion condition (±SEM across exemplars) in behavioral judgments.

(D) Correlation between behavioral classification accuracies (from C) and neural classification accuracies for each emotion class (based on errors of an SVM trained and tested on MMPFC voxel patterns).

claim that simpler dimensions like valence and arousal contribute to the organization of our emotion knowledge. For example, we included 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. With this approach, we show that it is possible to characterize the fine-grained representational structure of a high-level human reasoning capacity like emotion attribution.

RESULTS

Classification

In the scanner, subjects (n = 22) read 200 stimuli describing situations that would cause a particular emotion (see Experimental Procedures; example stimuli provided in Table 1). To confirm that these stimuli elicit reliable fine-grained emotion attributions, a group of subjects on MTurk 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 intended emotion), classifying the stimuli with 65% accuracy (chance = 5%; Figure 1C; see Supplemental Information for evidence that subjects attribute consistent emotions). This classification accuracy provided a benchmark with which to compare different models and brain regions.

To identify regions in which neural patterns contain information about emotions, we first replicated the finding that MPFC contains abstract emotion representations by testing whether neural patterns in MPFC could distinguish the valence in single trial estimates of these verbal stimuli. We functionally localized MPFC and other ToM regions in individual subjects (see Figure S1). We selected a subset of conditions that most closely align with the positive and negative conditions used previously [14] and tested whether neural patterns in MPFC would support decoding of valence. 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 ($0.603(0.019)$, $t(19) = 5.530$, $p < 0.001$).

We then investigated whether these or other regions contain information about the full set of 20 emotions. A whole-brain searchlight revealed that 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 Table S1). The searchlight analysis exhibited striking overlap with the set of regions recruited for ToM (Figure 1B shows overlap between the searchlight [family-wise error, FWE $p < .05$, $k > 25$] and the random effects analysis of false belief > false photograph from the localizer task, shown at $p < .001$ uncorrected) and justified our continued focus on these a priori ROIs. Consistent with the searchlight results, we were able to classify emotions with above-chance accuracy (1 out of 20 emotions, 5%) based on neural patterns in all individually localized ToM regions (Figure 1A; Table S2). Because these analyses involved training and testing across stimulus items, above-chance classification indicates a representation of emotion that generalizes across otherwise highly variable verbal scenarios.

Moreover, in the judgments provided by subjects on MTurk, there were reliable differences across the emotion categories in the extent to which subjects provided the expected emotion label (one-way ANOVA: $F(19,180) = 4.99$, $p < 0.001$; see Figure 1C), which provided another signature with which to compare neural representations. We computed separate accuracies for each emotion category in each ROI and correlated these with the behavioral emotion labeling 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 MTurk behavioral raters (see Table S2; see Figure 1D). Thus, the reliable across-emotion accuracy differences observed behaviorally were paralleled in the emotion-specific

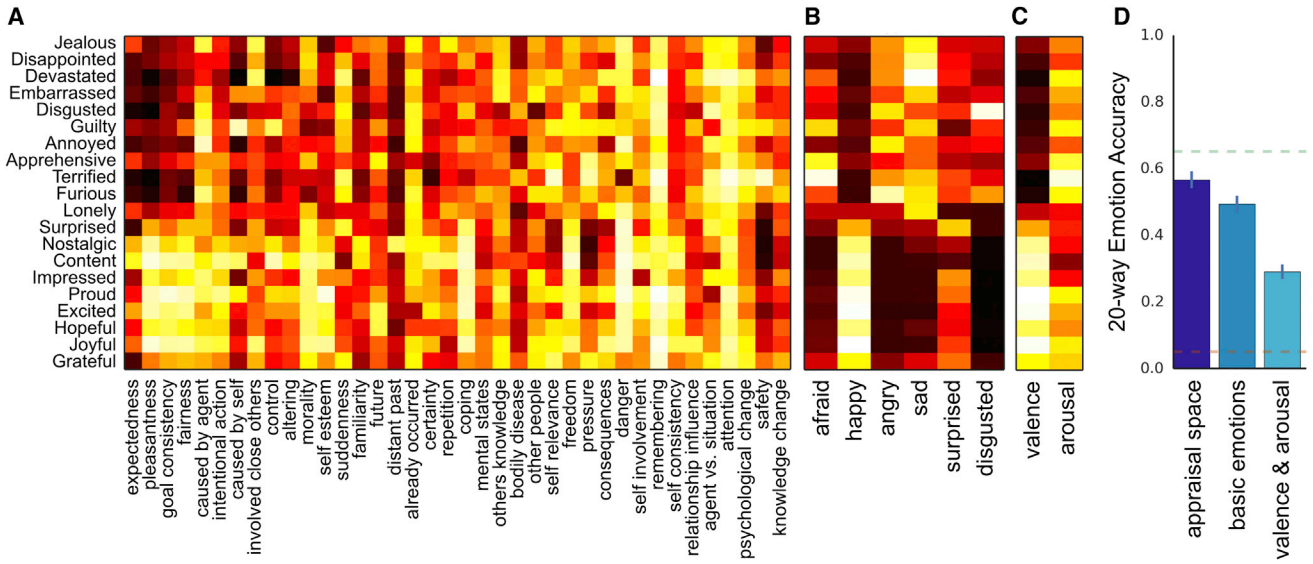

Figure 2. Competing Behavioral Feature Spaces Derived from MTurk Ratings

(A–C) Matrix of emotions × average dimension scores for the appraisal space (A), the six basic emotion space (B), and the circumplex space (C).

(D) Classification of 20 emotions (across stimulus exemplars) using information from each of the three competing spaces (±SEM across exemplars). Orange dotted line reflects chance (.05); blue dotted line reflects behavioral performance (.65).

accuracies of these neural populations (see Figure S1B 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 2A) 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. Specifically, we tested whether models trained on each of the feature vectors for a subset of stimuli could reliably classify the emotion label of untrained stimuli (see Supplemental Experimental Procedures). 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 classified well above a chance level of 5%, the appraisal feature space outperformed the other lower-dimensional spaces (57%, compared to behavioral benchmark of 65%; see Figure 2B; note because we used cross-validated accuracy, this analysis is not biased by the dimensionality of the models). 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 ($t(199) = 2.176$, $p = 0.031$).

RDMs derived from these three feature spaces were then compared to neural RDMs in each region to identify the space that best accounts for the similarity of the emotion conditions in their neural patterns. Because the appraisal RDM could perform best simply because it better discriminates the 20 emotions, we compared its performance to that of a pure categorical model and an RDM defined from the behavioral confusion matrix (see Supplemental Experimental Procedures), both of which also successfully discriminate the emotions (Figure 3). We also tested a model in which condition similarity is defined in terms of similarity of word-frequency vectors, a representation frequently used in fully automated approaches to emotional text classification such as sentiment analysis of reviews or other social media [34, 35]. Does the appraisal space outperform a raw word-level representation of the stimuli? Finally, we tested three control spaces capturing possible lower-level dimensions: reading ease, syntactic complexity, and rated intensity (confounded with motor response) (see Supplemental Experimental Procedures).

For each region, we correlated 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 emotion conditions in voxel-level patterns was positively correlated with similarity in the space of 38 appraisal dimensions (group-level Kendall's tau, DMPFC: 0.28; MMPFC: 0.21). Correlations with individual subject neural RDMs revealed a reliable relationship between the neural and model RDMs (see Table 2; Figure 4). In both DMPFC and MMPFC, the neural similarities were more correlated with the appraisal space than with either basic or circumplex spaces (see Table 3). In both regions, the appraisal RDM also outperformed the categorical and confusion spaces, suggesting that the superior performance of this model cannot be fully explained by its ability to better differentiate the 20 emotions. The appraisal space also outperformed the RDM defined from word-token frequencies and the control spaces for reading ease, syntactic complexity, and intensity (see Table 3).

We conducted the same analyses in the remaining ToM regions (RTPJ, LTPJ, PC, RSTS, and VMPFC): neural representations in these ROIs were also reliably correlated with the

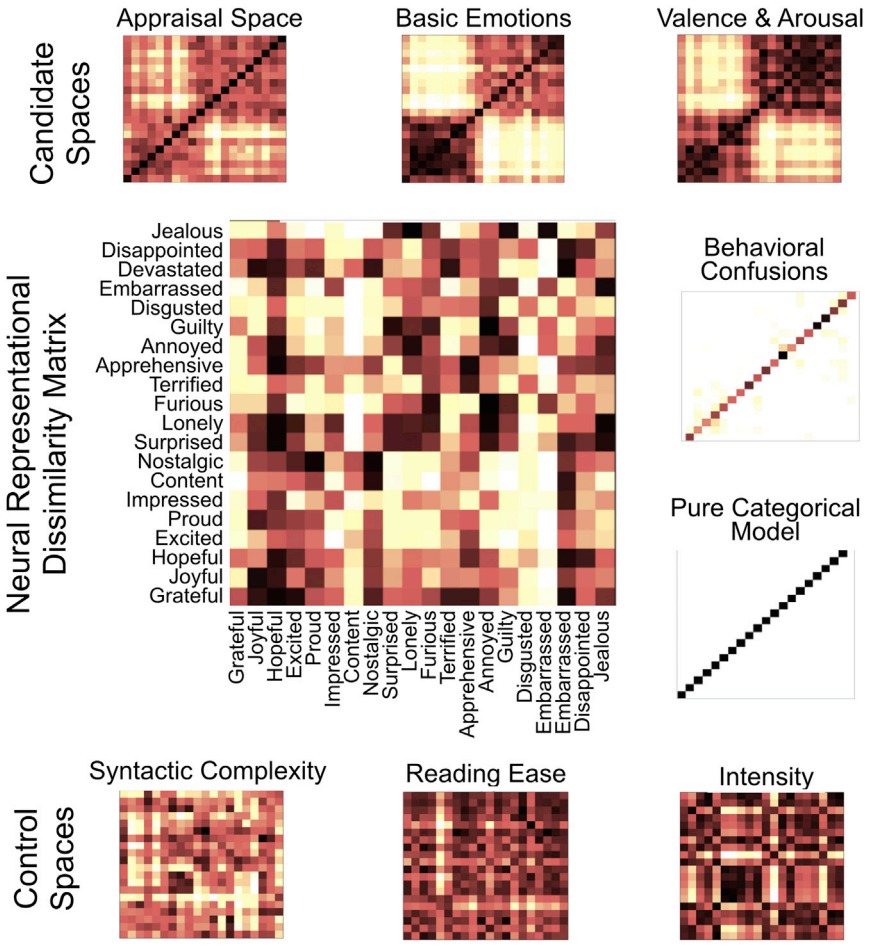

Figure 3. RSA Methods

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. In addition to the three candidate theories, we also test confusion and categorical spaces. Given that the appraisal space best captures the distinctions between the 20 emotions, it could outperform simpler models simply by virtue of its superior emotion discrimination. To test this possibility, we compare the appraisal space to a pure categorical RDM, which assumes that all emotions are perfectly and equally discriminable. As a more conservative test, we compute the correlation between neural RDMs and the raw behavioral confusion matrix. Like the categorical model, this confusion RDM captures the distinctions between the 20 emotions but also encodes similarity between different emotions as reflected in the behavioral confusions. If the appraisal space outperforms these two models, it suggests that the appraisal space fits the neural data in virtue of the features rather than emotion discriminability alone.

appraisal space RDM (see Figure 4 for RTPJ; see Figure S3 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). However, DMPFC and MMPFC were the only regions in which the high-dimensional space significantly outperformed all models.

Region Contributions

We could reliably decode emotions in all ToM ROIs, and the appraisal space did the best job of capturing the neural similarity space in most regions. Is the same information represented redundantly, or might these regions contribute differently to the representation of emotions? When classifying only valence, a model trained with voxels from all ToM ROIs (M(SEM) = 0.581(0.016), $t(21) = 4.942$, $p < 0.001$) performed less well than a model trained only with voxels in DMPFC (58.1% relative to 61%). However, when classifying the full set of 20 emotions, a model trained with voxels from all regions outperformed any of the individual ROIs, raising the possibility of non-redundant information across ToM regions.

To test for possible representational differences across the ROIs, we first used an iterative split-half reliability analysis (Supplemental Experimental Procedures). We found that neural RDMs in DMPFC and RTPJ were more correlated with themselves than with the other ROI ($M_{within} = 0.178$, $M_{between} = 0.164$, $p < 0.001$) and that this effect was not observed between MMPFC and RTPJ ($M_{within} = 0.121$, $M_{between} = 0.123$, $p < 0.922$). To further characterize potential non-redundancy, we explored whether the regions differed in the particular situation features they represent. Rather than compute separate RDMs for each of 38 appraisal features, we identified a reduced set of ten features that captured the most unique variance in behavioral ratings across items, using a stepwise regression approach (see Figures S2 and S4; Supplemental Experimental Procedures). We then computed the RDMs for this ten-dimensional space and also for each of the ten features individually and correlated each with the neural RDMs in different regions. The neural RDMs in all regions were reliably correlated with the RDM of the ten-feature space (see Table 2; Figure S4), which appears to capture much of the representational structure of the initial 38-dimensional space (Figure S2). Consistent with the results above, a repeated-measures ANOVA on the neural-model correlations for each feature (with ROI and feature as within-subjects factors) revealed a significant ROI × feature interaction for the comparison of DMPFC and RTPJ ($F(9,171) = 2.06$, $p = 0.036$), but not between MMPFC and RTPJ ($F(9,171) = 1.036$, $p = 0.414$). Together, these results provide evidence that multiple ToM regions are involved in the attribution of emotion and that some of these regions may contribute unique information to the final representational space that governs behavior.

Table 2. Neural RDM Results

ROI	Model	M	SEM	Z	df	Significance
DMPFC	appraisals	.08	.02	3.32	19	<0.001
	ten features	.08	.02	3.21	19	<0.001
MMPFC	appraisals	.06	.02	2.95	19	<0.002
	ten features	.05	.02	2.61	19	<0.004
RTPJ	appraisals	.07	.02	3.59	21	<0.001
	ten features	.06	.01	3.55	21	<0.001
ToM network	appraisals	.09	.02	3.68	21	<0.001
	ten features	.08	.02	3.68	21	<0.001

Model-neural correlations for 38-dimensional abstract event space and reduced space of ten features. df, degrees of freedom.

DISCUSSION

Decades of research in the science of emotion have aimed to characterize emotions in terms of some low-dimensional space of basic affective primitives [1, 23, 27, 36]. Behaviorally, we find that a space of abstract event features, derived from work in appraisal theory [33], reliably outperforms these simpler spaces in discriminating the 20 different emotions in our stimuli. Consistent with previous reports [13, 14], we find that neural representations in MPFC contain information about attributed emotions. Whereas prior studies focused on coarse distinctions (e.g., valence), we classify a set of nuanced emotions at above-chance levels. Moreover, by expanding to a rich space of eliciting situations, we are able to decode attributed emotions in all regions of the ToM network, and the searchlight results suggest that this information is largely restricted to these regions (particularly MPFC, RTPJ, and LTPJ).

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

The present work also probes the underlying representational structure that supports emotion discrimination. Previous literature [12–14] is consistent with the possibility that MPFC codes a limited space of affective dimensions such as valence and/or arousal. Moreover, even in our neural classification 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 not only that brain regions involved in ToM reasoning contain information about attributed emotions but also that this information is best captured by the high-dimensional space of event features.

In the majority of ToM regions, the similarity of emotion conditions in their voxel response patterns is most correlated with the similarity of the emotions in the space of appraisals. This result suggests a neural code that 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 data suggest that human emotion attribution is organized around abstract features of the causal context in which different emotions occur rather than the affective primitives that have dominated prior research.

A challenge for future work will be characterizing the scope and specificity of the neural representations in ToM regions. Do these neural populations contain representations specific to attributed emotion, coded within a space of emotion-relevant event features, or contain information in the form of domain-general semantic representations used in the service of emotion attribution? It is quite possible that these event representations function as intermediate features in the service of diverse inferential processes in addition to emotion attribution. Ultimately, successfully inferring emotions depends on a rich body of world knowledge, and neural populations specific to social cognition must interface with more general-purpose semantic systems. Characterizing information flow within and between these different networks will be an important avenue for future research.

Characterizing Representational Spaces

To characterize the feature space that governs representation of attributed emotion in the human brain, we draw on methods 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 [18, 37–39]. In one study, Mitchell and colleagues [38] coded object words in terms of co-occurrence with a set of verbs hypothesized to pick out relevant semantic dimensions (e.g., "manipulate," "taste"), a representation that was sufficient to support neural classification of untrained stimuli. Later work showed 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?) [40]. 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 ToM brain regions.

The study of object representation has also made headway on understanding differences across regions and temporal stages [37, 39], with RSA in particular providing a flexible framework for comparing the structure of the representations along the ventral pathway [41]. Interestingly, the present results provide preliminary evidence that ToM regions (particularly DMPFC versus RTPJ) may differ in their contributions to emotion inference. 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 stimulus representations that outperform the 38-dimensional space explored here. Many early approaches to modeling neural object representations involved hand-picked features (e.g., 25 verbs)

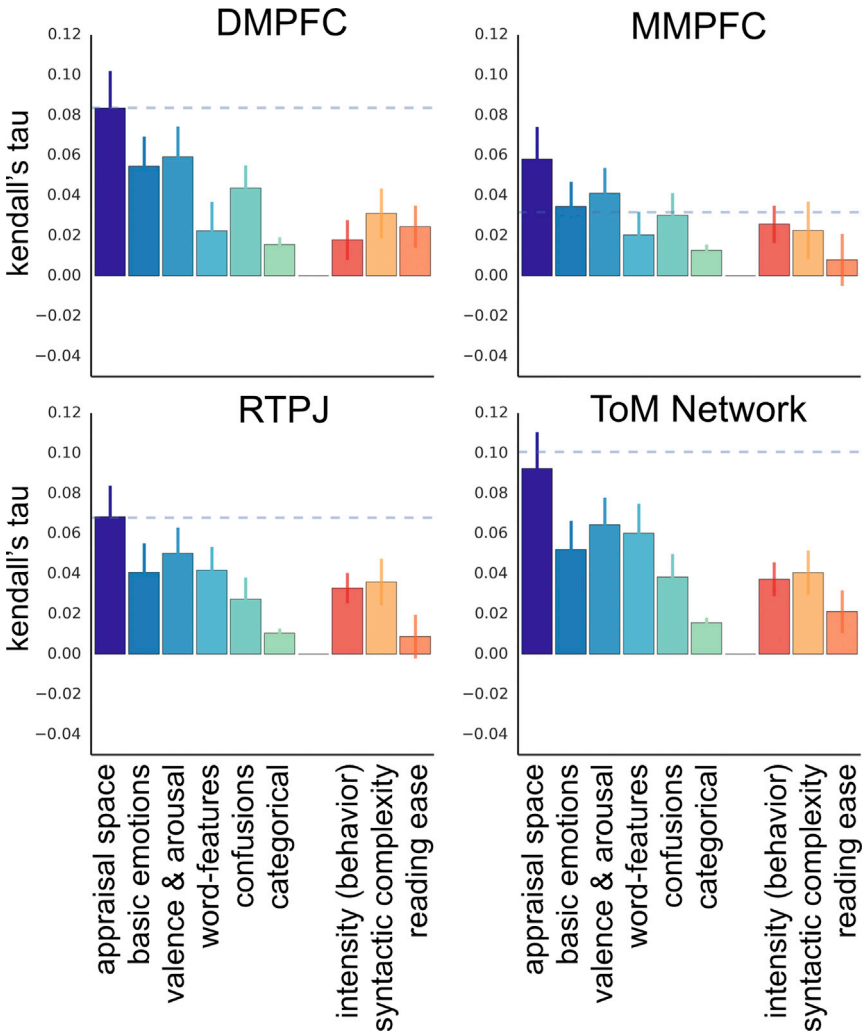

Figure 4. RSA Results

Mean correlation (Kendall's tau) between candidate model RDMs and individual subject neural RDMs (±SEM across subjects). Dotted line shows the noise ceiling (see Table S3).

[38] and often manual coding of stimuli within those spaces [37, 42]. However, recent research using high-throughput, data-driven approaches has yielded computational models that can be applied to raw stimuli (i.e., images) and achieve high quantitative fit to neural patterns [43]. Here, candidate features were selected based on prior theories without subsequent optimization (this list may therefore contain redundant or uninformative features, and some additional features are likely necessary), and the stimuli required manual annotation (MTurk ratings). In fact, the model using 38 abstract event features falls short of human behavioral performance when labeling stimuli (57% versus 65% accurate), indicating that this collection of features does not completely capture intuitive emotion knowledge. A data-driven discovery method might be better able to capture the full range of relevant dimensions; future research would ideally identify new sets of optimized features (either event features or some other candidate basis) and ways to infer these features directly from text.

A second, more fundamental limitation is that this approach aims to encode human emotion knowledge in terms of lists of appraisal checks applied to each stimulus. While this flat feature vector approach has been productive in other domains [44] and proved useful in the present paradigm, it is unlikely that representations in domains of high-level cognition such as ToM can be reduced to operations over lists of associated features [45]. For example, an attributed emotion 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 causal and compositional nature of emotion inference [4], future research may need to incorporate structured, generative knowledge representations from other areas of cognitive science [46]. The present findings lay groundwork for such research by providing an initial sketch of specific dimensions that might structure human emotion concepts and a framework for evaluating competing models of this knowledge.

Conclusions

Despite important open questions, the present data provide novel insight into the representations underlying human emotion inference and the neural populations that support them. We show that it is possible to decode attributed emotions from neural patterns in regions involved in mental state reasoning and provide quantitative insight into the underlying representational structure that supports this inferential ability. 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 are not reducible to more primitive affective primitives such as valence and arousal.

EXPERIMENTAL PROCEDURES

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

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 (Supplemental Experimental Procedures). A given subject rated stimuli on either features from the abstract event space (e.g., "Did someone cause this situation intentionally, or did it occur by accident?"; see Feature Table in Supplemental Information) or dimensions corresponding to the basic emotion space (e.g., "Was <character> happy in this situation?")

Table 3. Neural RDM Results

Comparison	ROI	M1	M2	z	Significance
Appraisals versus basic emotions	DMPFC	.08	.05	3.02	.002
	MMPFC	.06	.03	2.31	.021
Appraisals versus circumplex	DMPFC	.08	.06	2.84	.005
	MMPFC	.06	.04	2.80	.005
Appraisals versus word frequency	DMPFC	.08	.02	2.99	.003
	MMPFC	.06	.02	2.17	.030
Appraisals versus confusions	DMPFC	.08	.04	2.54	.011
	MMPFC	.06	.03	2.20	.028
Appraisals versus categorical	DMPFC	.08	.02	3.17	.002
	MMPFC	.06	.01	2.61	.009
Appraisals versus reading ease	DMPFC	.08	.02	2.39	.017
	MMPFC	.06	.01	2.02	.044
Appraisals versus syntax	DMPFC	.08	.03	2.50	.012
	MMPFC	.06	.02	1.98	.048
Appraisals versus intensity	DMPFC	.08	.02	3.21	.001
	MMPFC	.06	.03	2.05	.040

Statistical comparisons (Wilcoxon signed-rank test) of neural-model correlations for the appraisal space compared to all other candidate models, using neural RDMs in MMPFC (df = 19) and DMPFC (df = 19).

and the circumplex space (e.g., "Did <character> find this situation to be positive or negative?").

Feature-Based Classification of Behavioral Data

To test whether any of the three 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 (linear support vector machine [SVM]) trained on these features could classify the 20 distinct emotions (see Supplemental Experimental Procedures). We tested whether each feature space provided a basis for emotion discrimination that generalized across the different exemplars by using item-based cross-validation folds and computing the average cross-item classification accuracy for each feature space (comparing to the behavioral benchmark: 65%).

fMRI Emotion Attribution Task

In the emotion attribution task, subjects viewed 200 emotion stimuli, along with ten stories describing physical pain [47]. The experiment consisted of ten runs (7.37 min/run), each containing one exemplar for each of 21 trial types (20 emotion conditions, 1 pain). Each story was presented at fixation for 13 s, followed by a 2 s window for 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 s. The order of conditions was counterbalanced across runs and participants, and order of individual stories for each condition was randomized.

fMRI Analyses

Acquisition and preprocessing details are provided in Supplemental Experimental Procedures.

Classification Analyses

We first aimed to replicate previous valence decoding in MPFC [14] by choosing a subset of conditions that most closely matched the happy versus sad emotions used in that study ("excited," "joyful," "proud" versus "devastated," "disappointed," "annoyed") and testing whether voxel patterns in MPFC could reliably classify the valence of these stimuli. We then tested whether voxel patterns in MPFC or other ToM regions could reliably classify the set of 20 emotions.

We conducted MVPA within ROIs that were functionally defined based on individual subject localizer scans (including a ToM network ROI defined as the union of each subject's individually localized ROIs). We computed a single voxel pattern for each individual trial by averaging the preprocessed bold images for the trial and *Z* scoring relative to the mean across all trial responses in each 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 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-versus-one method [48]. Classification accuracy was averaged across ten cross-validation folds to yield a score for each subject per ROI, assessed with a one-sample t test (one tailed) over individual accuracies (comparing to chance: 0.5 for positive versus negative; 0.05 for 20-way classification). See Supplemental Experimental Procedures for further details.

RSA

To create RDMs for the competing representational spaces, we first averaged the feature vectors (from MTurk ratings) for each emotion condition (across stimuli), yielding the emotion-by-feature matrices shown in Figure 2. For each matrix, we then computed the Euclidean distance of feature vectors for each pair of emotions. We conducted this analysis iteratively (n = 1,000) across split halves of the data (five items per condition in each half), such that the self-distances along the diagonal are meaningful. In addition to the five candidate feature spaces (circumplex, basic emotions, appraisals, confusions, and categorical), we generated an additional space defined in terms of the similarity in word occurrences across stimuli, as well as 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 (Supplemental Experimental Procedures).

Neural RDMs were computed separately for each region in each subject with the same procedure as for feature space RDMs, except that features were voxel-wise neural responses rather than behavioral ratings (see Supplemental Information). We computed similarity of the conditions (Euclidean distance) in their voxel patterns (conducted across even and odd subsets so that the diagonal is interpretable), yielding an RDM for each region. To compare neural and model similarity spaces, we then computed the rank correlation (Kendall's tau-a) between the model and neural RDMs for each region in each subject and compared these correlations to chance (average Kendall's tau = 0) with a Wilcoxon test. We also compared the fit of different models by conducting a one-tailed Wilcoxon signed-rank test on the correlations for different pairs of models.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2015.06.009.

AUTHOR CONTRIBUTIONS

A.E.S. and R.S. designed and performed the research, analyzed the data, and wrote the paper.

ACKNOWLEDGMENTS

This research was supported by a National Science Foundation Graduate Research Fellowship (A.E.S.) and NIH Grant 1R01 MH096914-01A1 (R.S.). We thank L. Schulz, J. Tenenbaum, N. Kanwisher, J. Koster-Hale, B. Deen, E. Nook, and H. Richardson for helpful comments and discussion.

Received: April 4, 2015
Revised: May 12, 2015
Accepted: June 3, 2015
Published: July 23, 2015

REFERENCES

1. Ekman, P. (1992). Are there basic emotions? Psychol. Rev. *99*, 550–553.
2. Bachorowski, J.-A., and Owren, M.J. (2003). Sounds of emotion: production and perception of affect-related vocal acoustics. Ann. N Y Acad. Sci. *1000*, 244–265.
3. Dael, N., Mortillaro, M., and Scherer, K.R. (2012). Emotion expression in body action and posture. Emotion *12*, 1085–1101.
4. Ortony, A. (1990). The Cognitive Structure of Emotions (Cambridge University Press).
5. Clore, G.L., and Ortony, A. (2013). Psychological construction in the OCC model of emotion. Emot. Rev. *5*, 335–343.
6. 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.
7. Abelson, R.P., and Sermat, V. (1962). Multidimensional scaling of facial expressions. J. Exp. Psychol. *63*, 546–554.
8. 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).
9. Russell, J.A., and Bullock, M. (1986). On the dimensions preschoolers use to interpret facial expressions of emotion. Dev. Psychol. *22*, 97–102.
10. 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*, 692.
11. 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.
12. 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.
13. Peelen, M.V., Atkinson, A.P., and Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. J. Neurosci. *30*, 10127–10134.
14. Skerry, A.E., and Saxe, R. (2014). A common neural code for perceived and inferred emotion. J. Neurosci. *34*, 15997–16008.
15. Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. Proc. Natl. Acad. Sci. USA *103*, 3863–3868.
16. 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.
17. Kriegeskorte, N., and Kievit, R.A. (2013). Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. *17*, 401–412.
18. Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. *2*, 4.
19. Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol. *10*, e1003915.
20. Levenson, R.W. (2011). Basic emotion questions. Emot. Rev. *3*, 379–386.
21. Du, S., Tao, Y., and Martinez, A.M. (2014). Compound facial expressions of emotion. Proc. Natl. Acad. Sci. USA *111*, E1454–E1462.
22. Barrett, L.F. (2006). Valence is a basic building block of emotional life. J. Res. Pers. *40*, 35–55.
23. Russell, J.A. (1980). A circumplex model of affect. J. Pers. Soc. Psychol. *39*, 1161–1178.
24. Barrett, L.F., and Bliss-Moreau, E. (2009). Affect as a psychological primitive. In Advances in Experimental Social Psychology, M.P. Zanna, ed. (Academic Press), pp. 167–218.
25. Barrett, L.F., and Wager, T.D. (2006). The structure of emotion evidence from neuroimaging studies. Curr. Dir. Psychol. Sci. *15*, 79–83.
26. 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. Published online January 28, 2015. http://dx.doi.org/10.1093/cercor/bhv001.
27. 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.
28. Barrett, L.F. (2006). Are emotions natural kinds? Perspect. Psychol. Sci. *1*, 28–58.
29. 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. *17*, 715–734.
30. Ellsworth, P.C. (2013). Appraisal theory: old and new questions. Emot. Rev. *5*, 125–131.
31. Scherer, K.R. (1999). Appraisal theory. In Handbook of Cognition and Emotion, T. Dalgleish, and M.J. Power, eds. (New York: John Wiley & Sons), pp. 637–663.
32. Meuleman, B., and Scherer, K. (2013). Nonlinear appraisal modeling: an application of machine learning to the study of emotion production. IEEE Trans. Affect. Comput. *4*, 398–411.
33. Scherer, K.R., and Meuleman, B. (2013). Human emotion experiences can be predicted on theoretical grounds: evidence from verbal labeling. PLoS ONE *8*, e58166.
34. Pang, B., and Lee, L. (2004). A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics ACL '04. (Stroudsburg: Association for Computational Linguistics).
35. 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.
36. Lindquist, K.A. (2013). Emotions emerge from more basic psychological ingredients: a modern psychological constructionist model. Emot. Rev. *5*, 356–368.
37. 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. *26*, 120–131.
38. 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.
39. 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.
40. 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.
41. Cichy, R.M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. Nat. Neurosci. *17*, 455–462.
42. 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.
43. 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. USA *111*, 8619–8624.

44. 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.

45. Laurence, S., and Margolis, E. (1999). Concepts and cognitive science. In Concepts: Core Readings, E. Margolis, and S. Laurence, eds. (MIT), pp. 3–81.

46. 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.

47. 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.

48. 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.

