Published in final edited form as:
Trends Cogn Sci. 2016 June ; 20(6): 444–455. doi:10.1016/j.tics.2016.03.011.

Decoding the Nature of Emotion in the Brain

Philip A. Kragel[1] and **Kevin S. LaBar**[1,*]
[1]Department of Psychology and Neuroscience, Duke University, Durham, NC 27708, USA

Abstract

A central, unresolved problem in affective neuroscience is understanding how emotions are represented in nervous system activity. After prior localization approaches largely failed, researchers began applying multivariate statistical tools to reconceptualize how emotion constructs might be embedded in large-scale brain networks. Findings from pattern analyses of neuroimaging data show that affective dimensions and emotion categories are uniquely represented in the activity of distributed neural systems that span cortical and subcortical regions. Results from multiple-category decoding studies are incompatible with theories postulating that specific emotions emerge from the neural coding of valence and arousal. This ‘new look’ into emotion representation promises to improve and reformulate neurobiological models of affect.

Mapping the Brain Basis of Emotion

Emotions are often experienced as discrete feelings, yet the brain basis of specific emotions remains poorly understood. The inherent challenges in localizing the neural basis of human emotions with fMRI are well illustrated by research investigating the correspondence between amygdala activity and emotional states of fear. Meta-analytic summaries of this literature [1–3] demonstrate consistent increases in blood oxygen level-dependent (BOLD) response within the amygdala during experimental manipulations eliciting states of fear. Yet, amygdala activation is observed during the elicitation of diverse **affective states** (see Glossary), including both positive and negative emotions [4], and during manipulations of broader affective dimensions, such as **arousal** and **valence** (Figure 1A). This combination of results and the limited spatiotemporal resolution of fMRI (compared with other methods discussed in Box 1) complicate specifying the role of the amygdala. Due to the variety of stimuli that engage this region, the amygdala has been proposed to play a broader role in detecting salient stimuli [5,6] and in eliciting central and autonomic arousal [7,8], of which fear is a particularly potent example. These kinds of observations have led theorists and researchers to abandon simple one-to-one mappings between a given brain structure and a given emotion [4,9–11]. In response, two divergent lines of thinking have emerged: one that abandons the notion of emotion-specific representations in the brain [4,12] and another that has refocused enquiry towards identifying distributed neural systems that underlie emotional behavior [9,13] (Box 2).

[*]Correspondence: ; Email: klabar@duke.edu (K.S. LaBar)..

Seeking new ways to characterize emotion representations, imaging researchers have begun applying multivariate techniques – namely, pattern classification and representational similarity analysis [14] – to investigate how emotions may be **decoded** from distributed patterns of brain activity. These methods, broadly termed **multivoxel pattern analysis (MVPA)** when applied to fMRI data, show much promise in other domains of cognitive neuroscience to specify how neural systems are linked to separable mental states, such as the category of perceived objects or the contents of working memory (for reviews see [15,16]). By identifying mappings between multiple measures of neural activity and single mental states, MVPA can overcome the limiting assumption that emotions are represented by dedicated modules or functionally homogeneous units [17]. Accordingly, MVPA shifts the focus of inquiry to emotion-specific patterns that emerge across locally distributed populations of neurons within a region or across neural networks at larger spatial scales, which is more closely aligned with contemporary views of mental state representations [9,13].

Multivariate approaches have several other advantages over univariate ones, which have been the mainstay of emotion imaging research historically. Multivariate approaches have high sensitivity because they incorporate more information than a single summary statistic of the most significantly activated voxel in a brain region. Their data-driven nature can reveal novel or counterintuitive insights relative to approaches that rely on testing *a priori* hypotheses derived from existing theories; rather than testing a theoretical assumption of how researchers postulate emotions are represented in the brain, the analysis technique decodes the complex fMRI data to inform the researcher how the brain organizes emotions. Outcomes from MVPA can then be compared with existing theories to help adjudicate between different perspectives. As with univariate approaches, multivariate approaches will detect any difference between conditions, even those that are not of primary interest to the researcher [18]. Thus, well-controlled experimental designs combined with additional analysis of the information content being decoded should be conducted to provide support for the interpretation of the results. Finally, pattern classification approaches can particularly benefit from inclusion of error analyses and measures from signal detection theory such sensitivity, specificity, and area under the **receiver operating characteristic (ROC) curve** [19], as accuracy measures alone are not the most informative indices of classifier performance.

This review examines recent functional neuroimaging studies that use MVPA to investigate how emotions are reflected in distributed patterns of brain activity, focusing on work that sheds light on the **representational space** that best organizes instances of emotional experience (Figure 2, Key Figure). Related work using multivariate techniques to study the perception of emotional expressions is not covered here (e.g., [20,21]; see [13] for a review). The first section reviews research classifying brain states in terms of the affective dimension of valence, or pleasantness [22]. The second part covers studies that decode fMRI activity into multiple discrete emotion categories [23,24]. Finally, we conclude by evaluating the correspondence between high-dimensional brain activity and theoretical models describing the organization of emotions.

Decoding Affectively Valenced Brain States

Classic behavioral studies in psychology using the multivariate tools of factor analysis and multidimensional scaling show that facial expressions of emotions, self-reported moods, and similarity ratings of emotion words are principally organized according to valence ([25]; but see [26] for a counterexample in autobiographical memory). Because valence is accordingly thought to be a ‘core’ affective feature of our emotional lives [27,28], researchers have begun utilizing MVPAto investigate the manner in which patterns of fMRI activity encode information according to this affective dimension.

One such study [29] employed representational similarity analysis [30] to identify brain regions whose activity reflects a continuous dimension of subjective valence spanning from negative to positive affect. Drawing on evidence from single-cell electrophysiological studies in monkeys demonstrating that different populations of neurons in the orbitofrontal cortex (OFC) code positive and negative value both independently and in an integrated fashion [31], the investigators hypothesized that MVPA would be able to detect voxel-level biases in the distribution of such neurons when measured via fMRI in humans. Accordingly, patterns of OFC response to different instances of positive affect were predicted to be similar to one another and distinct from subjectively negative experiences. To assess the specificity of valence representations, single-trial estimates of neural responses to visual scenes and gustatory stimuli were anatomically localized within the OFC. The information content of this region was characterized by constructing representational similarity matrices [32] that indexed the correlation of OFC activation between all possible pairs of stimulus valence levels. Regression models were then used to examine the relationship between the similarity of OFC response profiles and differences in subjective valence, quantified using online self-reports of positive and negative experience.

These analyses revealed that differences in subjective ratings of valence predicted the similarity of responses within the OFC when comparing responses both within and between visual and gustatory stimuli. Furthermore, classification of subjective valence based on the similarity of neural activation within this region was found to generalize across participants (**cross-validating** the classifier on data from held-out subjects), although the observed accuracy (55.6%, where chance was 50%) was considerably lower than an analogous classification of object categories on the basis of activation within the ventral temporal cortex (80.1%; Figure 3A). Such low levels of discrimination indicate that patterning within the OFC alone may not effectively serve as an objective marker of subjective valence. Nevertheless, the results demonstrate that a portion of neural activity within the OFC is consistent with a representation of subjective valence that is shared across stimulus modalities and individuals. Additionally, while external, perceptual aspects of stimuli are well characterized in modality-specific cortices [33,34], the coding of valence in the OFC is likely to involve a transformation from basic stimulus features into a more abstract, common representation (see also [35] for related work on the coding of subjective value in this region).

Another study of the valence continuum [36] classified patterns of fMRI activity to identify a neural signature that predicts differences in the subjective experience of negative emotion in response to aversive pictures. To identify patterns of BOLD response that accurately predicted negative emotional experience with a high degree of generalization, a large sample of 182 subjects was presented negative and neutral scenes from the International Affective Picture System (a set of standardized images that reliably evoke negative, neutral, and positive affective reactions [37]). Following the presentation of images on every trial, participants made behavioral ratings indicating their current emotional state, ranging from neutral (a rating of 1) to strongly negative (a rating of 5). Machine-learning models using least absolute shrinkage and selector operator principal component regression (LASSO-PCR) were then trained to predict the five levels of negative emotional experience from whole-brain estimates of BOLD response.

The resulting neural model, termed the Picture Induced Negative Emotion Signature (PINES), demonstrated high levels of sensitivity when testing in independent subjects, both between extreme ratings of negative emotion (ratings of 1 vs 5 were classified at 100% accuracy) and between adjacent ratings (90.7% and 100% accuracy for ratings of classifying trials rated 5 vs 3 and 3 vs 1, respectively). In terms of spatial localization, increased activation of several regions, including the anterior cingulate, insula, amygdala, and periaqueductal gray, contributed to the prediction of negative emotional experience. Importantly, the distributed model was found to be a better predictor of negative emotion than the average response within individual subregions or resting state networks [38], demonstrating that MVPA provided unique insight into the representation of negative emotion across activation patterns spanning the whole brain. These results clearly demonstrate that a continuous dimension of negative affect is effectively predicted by a distinct and distributed pattern of neural activation spanning multiple brain regions.

Although these results are notable in terms of signal detection capacity, it is possible that factors other than negative emotion informed the classification model. To evaluate the specificity of the PINES, the researchers applied it to fMRI data acquired during painful thermal stimulation (which is similarly negative and arousing) and compared it against a biomarker sensitive and specific to physical pain, the neurologic pain signature (NPS) (see [39] for details of its development and validation). The results of this analysis demonstrated a clear double dissociation: whereas the PINES accurately classified negative versus neutral emotional experiences, but not high- versus low-intensity pain, the NPS accurately discriminated between differences in pain reports but not emotional intensity. Although the observed specificity suggests that commonalities between experiencing pain and aversive images did not drive the results, it is still possible that other factors could inform the PINES, such as differences in attentional orienting or visual processing, which are likely to differ across negative and neutral images. Notwithstanding these limitations, the findings of this study illustrate how neural biomarkers developed using MVPA can discriminate between brain states that are similar in terms of valence and arousal at high levels of specificity.

The above studies demonstrate that valenced brain states can be differentiated on the basis of neural activity (see also [40,41] for related work), partially supporting both dimensional and **psychological construction views of emotion** that assume a prominent role of valence in the neural representation of core affect [4]. In particular, the medial OFC was found to contain valence-related information in both studies (albeit at uncorrected thresholds in [36]).

However, this emerging body of research is limited in its capacity to address how emotions are represented in a higher-dimensional space. Valence is a useful construct in part because it relates emotions to one another based on a shared feature and can accordingly be used to infer the emotional state of an individual (e.g., an individual in a positive state is less likely to report experiencing anger or sadness). However, these studies have not tested whether brain-based predictions of valence inform on-line reports of specific emotions, which would be expected if brain systems dedicated to valence and arousal form the basis of emotional experience [28,42]. Additionally, these studies do not differentiate subjective valence from arousal, which plays a prominent role in emotion and is often confounded with self-reports of valence when sampling a small number of emotional states (i.e., when classifying responses to negative and neutral images). Thus, it is not yet clear whether brain-based models of valence are concordant with dimensional theories of emotion in terms of parameterization or generalizability.

Decoding Brain States during the Experience of Discrete Emotions

Although emotions can be understood by studying features shared across emotional states, such as their valence or arousal, categorical models instead focus on differences in the antecedent events, neural circuitry, and behavioral outputs specific to each emotion [43]. These models commonly posit that emotions are experienced as independent categories in humans and are differentiated in their neurophysiological expression. Following this logic, a number of experiments have been conducted using fMRI with the goal of classifying neural activity along multiple distinct emotion categories.

In one of the first studies using MVPA methods to predict the experience ofspecific emotions on the basis of fMRI activity [44], ten method actors (eight female) were asked to experience multiple emotions (anger, disgust, envy, fear, happiness, lust, pride, sadness, and shame) through script-driven imagery when prompted by corresponding cue words. Patterns of BOLD response within the most stable 240 voxels (spanning the whole brain, but predominantly comprising orbital and lateral frontal regions) during the presentation of verbal cues were used as input to **Gaussian naïve Bayes classifiers**. The nine emotions were classified at 84% mean rank accuracy when training and testing within the same subject and at 70% mean rank accuracy when training and testing was performed on independent subjects (where chance was 50%) – establishing that emotional states can be objectively differentiated on the basis of brain activity.

To better understand the relationship between patterns of BOLD response and the affective content of the scripts (i.e., valence, arousal, control, certainty, and attention), the authors conducted an exploratory factor analysis. Although a number of associations were identified between pre-scan ratings of the scripts and factors decomposed from neural activity, none was specific. For instance, the factor explaining the most variance across patterns of neural response (which the authors interpreted to reflect valence) was not only correlated with valence ratings but was also correlated with ratings of arousal, certainty, and control. This lack of specificity may be partly due to the fact that only two positive emotions (happiness and pride) were sampled in this study, both of which are highly arousing. Additionally, the sample size was extremely small for the application of exploratory factor analysis [45] and little time was provided for participants to experience the emotions on each trial (9 s), thereby under-sampling the experiential aspects of the emotion induction. Despite these complications in relating specific emotions to broader affective constructs, this study set the stage for research classifying discrete emotions in distributed patterns of brain activity. Additionally, the study acknowledged the difficulty of evaluating psychological construction [4,28,46] and basic emotion [47,48] views with MVPA, as successful pattern classification could be the product of either cognitive constructions or emotion-specific neural systems.

A more recent study [49] utilized script-driven imagery and short movie clips to elicit the basic emotions of disgust, fear, happiness, anger, surprise, and sadness. Patterns of BOLD response from gray matter voxels spanning the whole brain were accurately classified into the six emotion categories using linear neural networks (34% and 23% accuracy for movies and imagery, where chance was 20% and 16.7%, respectively). Further, the researchers found that post-scan similarity judgments of emotion words used to cue imagery were positively correlated with the number of misclassifications made in MVPA, demonstrating that the words that were distinct in terms of their meaning were more accurately classified. Together, these findings provide evidence that subjective judgments about emotional events are consistent with the expression of distinct neural substrates probabilistically linked to their occurrence.

Whereas the previously described studies used behavioral ratings of stimuli outside the scanner, either before or after scanning, we [50] conducted an fMRI experiment in which participants reported on their emotional experience following emotion induction in the scanner using instrumental music and cinematic films (states of contentment, amusement, surprise, fear, anger, and sadness and a neutral control condition were elicited). Such on-line verification of emotional experience is critical for confirming coherence across emotional systems [51] and for isolating which affective factors contribute to classification. Participants' ratings confirmed that emotions were experienced discretely in accordance with the intended category for each stimulus and ratings on dimensional terms also discriminated among some emotions (e.g., valence ratings differentiated contentment and amusement from fear, anger, and sadness). Together, these behavioral analyses established that participants' subjective experience was concordant with theoretical models proposing dimensional and categorical representation of emotions.

Supporting the notion that emotions are represented in distributed neural systems, whole-brain patterns of BOLD response during the music and film induction were classified using partial least-squares discriminant analysis at 37.3% accuracy compared with chance levels of 14.3% (when training and testing models on independent subjects). The activity patterns of the discrete emotions predicted the induction of discrete emotional states consistently across subjects with a high degree of sensitivity and specificity. By **bootstrapping** regression coefficients of their classification models, the researchers found that activity informing the classification models for each emotion was localized in relatively non-overlapping brain regions, spanning cortical and subcortical areas (Figure 4).

To relate neural classification to emotional experience, the investigators used on-line measures of emotional experiences to construct a categorical model, with each emotion represented along an independent axis, and a 2D model organized by valence and arousal. Regression models were then constructed to assess the extent to which errors in classifying fMRI activation could be predicted on the basis of self-report. This analysis revealed that differences in categorical aspects of experience were associated with improved decoding accuracy. By contrast, instances that differed the most in terms of valence and arousal were more frequently associated with classification errors, indicating that these dimensional constructs may inefficiently discriminate among specific emotional brain states.

Although unexpected from the viewpoint of dimensional models, impaired classification of emotions that differ in terms of valence and arousal is concordant with findings from meta-analytic efforts that utilized MVPA to discriminate among basic emotion categories [52] but failed to differentiate positive and negative valence [53]. These meta-analytic findings were interpreted by the authors as supporting constructionist models of emotion [52] because the patterns of activity that predicted each emotion category spanned multiple intrinsic brain systems (see also [54]). However, direct model comparisons to rule out alternative interpretations based on categorical theories were not conducted. Further, the findings suggest that valence is not the primary driver of brain activity that distinguishes discrete emotions. Together, these emerging results suggest that patterns of brain activity indicative of affective dimensions such as arousal and valence may not account for a given brain region's contribution to a specific emotion, such as fear, as illustrated for the amygdala in Figure 1B. Given that dimensional models explain many aspects of self-reported emotions (as discussed above), neurophysiological and behavioral facets of emotion representation may not be **isomorphic** (for a similar conclusion using MVPA in the autonomic nervous system, see [55]).

Concluding Remarks

Whereas efforts focusing on functional localization have largely failed at mapping emotions onto individual brain regions, emerging research using MVPA has demonstrated that information encoded in both local neural ensembles and whole-brain activation patterns can be utilized to predict affective dimensions and discrete emotions with high levels of specificity. Findings across multiple studies demonstrate that machine learning approaches can fruitfully be used to characterize self-reports of emotional experience. Contrary to the assumption held by dimensional and psychological construction accounts that hedonic valence is at the core of emotional experience with an innate neural basis in humans [46,56], MVPA has revealed that brain representations of emotions are better characterized as discrete categories as opposed to points in a low-dimensional space parameterized along the valence continuum. However, it is not yet clear whether these category-specific, distributed activation patterns reflect evolutionarily ingrained networks, constructive processes, or a combination of factors.

Despite some broad overlap at a larger spatial scale (Figure 4), the localization of emotion-predictive patterns varies at the voxel level across the few MVPA studies of emotion induction conducted thus far [44,49,50]. Proponents of constructionist models argue that this variability indicates that multivariate classifiers do not learn the essence of emotion categories [54,56] but instead differentiate among populations of emotional instances sampled within a study. Thus, disparities in patterning across studies could be driven by differences in induction procedures, analytical methods, and inclusion of different emotions and varying numbers of emotions in each study. One important consideration is that some studies reported only the most informative voxels within their sample (e.g., [44,49]) and did not verify the extent to which emotion-predictive patterns were consistent across subjects, making it less likely that the effects will generalize to independent samples. In light of these issues, it is premature to draw any strong conclusions about the localization of emotion-specific patterning: it is possible that some aspects of classification models are idiosyncratic to particular samples or experimental manipulations, while some aspects of emotion-specific patterning could be invariant across studies, potentially linked to common functional and behavioral changes associated with specific emotions. Fully understanding which factors contribute to differences in emotion-predictive patterns across studies and whether a single, invariant neural model of emotion categories can sufficiently predict subjective experience remain questions open to future investigation.

Although categorical representation of emotions in the brain is concordant with accounts suggesting that emotions are the product of adaptive pressures [57,58], the MVPA results are compatible with a broad range of models because emotions can be considered discrete without necessitating an evolutionary stance regarding their origin [47]. For example, appraisal theories suggest that some emotions can be considered modal based on their frequency and prototypically [59] or that instances of the same emotion are similar because their antecedents share core relational themes [60]. Due to the diversity of biological, social, psychological, and computational models of emotion, there is much to be learned about the organization of affective brain states (see Outstanding Questions). Future research in this area should focus on additional analyses of the information content used by classifiers to predict emotions from neural data (such as the role of specific appraisals) and should conduct comparisons across other models of emotion for the imaging data to maximally inform theory development. For example, one recent study explored how emotions are attributed to others using verbal scenarios and found that more complex appraisal features better explained neural representations of emotional stimuli than basic or dimensional accounts [61]. Using this direct model-comparison approach to understand brain responses during emotion induction (see also [50]) is a promising avenue for future studies. As is evident from this brief review, advances in multivariate approaches to neuroimaging have reinvigorated the quest to solve one of the biggest puzzles in affective neuroscience: to identify how specific feelings emerge from complex patterns of neural activity.

Glossary

Affect	the manner in which emotional events influence behavior and subjective feelings, often operationalized in terms of valence and arousal
Arousal	the degree of activation experienced during an instance of emotion, ranging from calm to excited

Bootstrapping statistical method used to assess the accuracy of a parameter estimate through repeated resampling with replacement. The method is commonly used to estimate confidence intervals

Cross-validation a statistical technique for estimating the degree to which a model will generalize to independent data. The method involves repeatedly partitioning data into independent samples for training and evaluating models

Decoding predicting the mental state associated with a pattern of brain activity or similar dependent measure

International Affective Picture System (IAPS) a set of standardized images that includes negative, neutral, and positive visual scenes

Isomorphic two representational spaces are isomorphic if there exists a one-to-one correspondence between all elements in both spaces, such that they have the same structural properties

Gaussian naïve Bayes classifier a multivariate classification model that assumes that continuous features predictive of each class are distributed as Gaussian distributions. The mean and covariance of these distributions can be estimated on training data for each class, which can then be used to compute the probability of class membership for testing data

Least absolute shrinkage and selector operator principal component regression (LASSO-PCR) a multivariate regression procedure that combines $\ell1$ regularization and principal component analysis. High-dimensional data are reduced to a smaller number of components that are then regressed onto outcome variables while penalizing the absolute size of the regression coefficients. This approach can identify sparse models in the presence of many features, making it well suited for fMRI data

Multivoxel pattern analysis (MVPA) analysis approach that assesses the information contained in patterns of fMRI (or electrophysiological) activity, either by comparing the similarity of responses across multiple experimental conditions or by learning a mapping from multiple voxels to a categorical or continuous outcome variable

Psychological construction accounts of emotion views that stress that emotions are not biologically innate categories but are constructed from multiple processes (e.g., facial expression, somatic activity, cognitive appraisals, subjectively experienced valence) that are not specific to any emotion

Receiver operating characteristic (ROC) curve	a plot of the true-positive rate (sensitivity) against the false-positive rate (1 – specificity) indicating the performance of a binary classifier at multiple decision thresholds
Representational space	a high-dimensional space in which instances of emotion can be related to one another. The dimensionality of the space depends on the number of features sampled (e.g., voxels, self-report items)
Valence	the hedonic tone of emotional experience, ranging from bad (unpleasant) to good (pleasant)

References

1. Vytal K, Hamann S. Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. J. Cogn. Neurosci. 2010; 22:2864–2885. [PubMed: 19929758]
2. Murphy FC, et al. Functional neuroanatomy of emotions: a meta-analysis. Cogn. Affect. Behav. Neurosci. 2003; 3:207–233. [PubMed: 14672157]
3. Phan KL, et al. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. Neuroimage. 2002; 16:331–348. [PubMed: 12030820]
4. Lindquist KA, et al. The brain basis of emotion: a meta-analytic review. Behav. Brain Sci. 2012; 35:121–143. [PubMed: 22617651]
5. Cunningham WA, Brosch T. Motivational salience: amygdala tuning from traits, needs, values, and goals. Curr. Dir. Psychol. Sci. 2012; 21:54–59.
6. Sander D, et al. The human amygdala: an evolved system for relevance detection. Rev. Neurosci. 2003; 14:303–316. [PubMed: 14640318]
7. Cardinal RN, et al. Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. Neurosci. Biobehav. Rev. 2002; 26:321–352. [PubMed: 12034134]
8. Murray EA. The amygdala, reward and emotion. Trends Cogn. Sci. 2007; 11:489–497. [PubMed: 17988930]
9. Hamann S. Mapping discrete and dimensional emotions onto the brain: controversies and consensus. Trends Cogn. Sci. 2012; 16:458–466. [PubMed: 22890089]
10. Pessoa L. Beyond brain regions: network perspective of cognition–emotion interactions. Behav. Brain Sci. 2012; 35:158–159. [PubMed: 22617666]
11. Scarantino A. Functional specialization does not require a one-to-one mapping between brain regions and emotions. Behav. Brain Sci. 2012; 35:161–162. [PubMed: 22617670]
12. Barrett LF. Are emotions natural kinds? Perspect. Psychol. Sci. 2006; 1:28–58. [PubMed: 26151184]
13. Kragel PA, LaBar KS. Advancing emotion theory with multivariate pattern classification. Emot. Rev. 2014; 6:160–174.
14. Kriegeskorte N, Kievit RA. Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. 2013; 17:401–412. [PubMed: 23876494]
15. Haxby JV, et al. Decoding neural representational spaces using multivariate pattern analysis. Annu. Rev. Neurosci. 2014; 37:435–456. [PubMed: 25002277]
16. Haynes JD. A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. Neuron. 2015; 87:257–270. [PubMed: 26182413]
17. Scarantino A, Griffiths P. Don't give up on basic emotions. Emot. Rev. 2011; 3:444–454.
18. Todd MT, et al. Confounds in multivariate pattern analysis: theory and rule representation case study. Neuroimage. 2013; 77:157–165. [PubMed: 23558095]

19. Sokolova, M., et al. Proceedings of the 19th Australian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence. Springer-Verlag; 2006. Beyond accuracy, f-score and ROC: a family of discriminant measures for performance evaluation.; p. 1015-1021.
20. Ethofer T, et al. Decoding of emotional information in voice-sensitive cortices. Curr. Biol. 2009; 19:1028–1033. [PubMed: 19446457]
21. Peelen MV, et al. Supramodal representations of perceived emotions in the human brain. J. Neurosci. 2010; 30:10127–10134. [PubMed: 20668196]
22. Russell JA. A circumplex model of affect. J. Pers. Soc. Psychol. 1980; 39:1161–1178.
23. Ekman P, Cordaro D. What is meant by calling emotions basic? Emot. Rev. 2011; 3:364–370.
24. Oatley K, Johnson-Laird PN. Cognitive approaches to emotions. Trends Cogn. Sci. 2014; 18:134–140. [PubMed: 24389368]
25. Feldman Barrett L, Russell JA. The structure of current affect: controversies and emerging consensus. Curr. Dir. Psychol. Sci. 1999; 8:10–14.
26. Talarico JM, et al. Emotional intensity predicts autobiographical memory experience. Mem. Cogn. 2004; 32:1118–1132.
27. Barrett LF. Solving the emotion paradox: categorization and the experience of emotion. Pers. Soc. Psychol. Rev. 2006; 10:20–46. [PubMed: 16430327]
28. Russell JA. Core affect and the psychological construction of emotion. Psychol. Rev. 2003; 110:145–172. [PubMed: 12529060]
29. Chikazoe J, et al. Population coding of affect across stimuli, modalities and individuals. Nat. Neurosci. 2014; 17:1114–1122. [PubMed: 24952643]
30. Kriegeskorte N, et al. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U.S.A. 2006; 103:3863–3868. [PubMed: 16537458]
31. Morrison SE, Salzman CD. The convergence of information about rewarding and aversive stimuli in single neurons. J. Neurosci. 2009; 29:11471–11483. [PubMed: 19759296]
32. Kriegeskorte N, et al. Representational similarity analysis – connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2008; 2:4. [PubMed: 19104670]
33. Brouwer GJ, Heeger DJ. Decoding and reconstructing color from responses in human visual cortex. J. Neurosci. 2009; 29:13992–14003. [PubMed: 19890009]
34. Kamitani Y, Tong F. Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 2005; 8:679–685. [PubMed: 15852014]
35. McNamee D, et al. Category-dependent and category-independent goal-value codes in human ventromedial prefrontal cortex. Nat. Neurosci. 2013; 16:479–485. [PubMed: 23416449]
36. Chang LJ, et al. A sensitive and specific neural signature for picture-induced negative affect. PLoS Biol. 2015; 13:e1002180. [PubMed: 26098873]
37. Lang, PJ., et al. International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual. University of Florida; 2008.
38. Yeo BT, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. 2011; 106:1125–1165. [PubMed: 21653723]
39. Wager TD, et al. An fMRI-based neurologic signature of physical pain. N. Engl. J. Med. 2013; 368:1388–1397. [PubMed: 23574118]
40. Baucom LB, et al. Decoding the neural representation of affective states. Neuroimage. 2012; 59:718–727. [PubMed: 21801839]
41. Shinkareva SV, et al. Representations of modality-specific affective processing for visual and auditory stimuli derived from functional magnetic resonance imaging data. Hum. Brain Mapp. 2014; 35:3558–3568. [PubMed: 24302696]
42. Barrett LF. Valence is a basic building block of emotional life. J. Res. Person. 2006; 40:35–55.
43. Tracy JL, Randles D. Four models of basic emotions: a review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. Emot. Rev. 2011; 3:397–405.
44. Kassam KS, et al. Identifying emotions on the basis of neural activation. PLoS ONE. 2013; 8:e66032. [PubMed: 23840392]
45. MacCallum RC, et al. Sample size in factor analysis. Psychol. Methods. 1999; 4:84.

46. Lindquist KA. Emotions emerge from more basic psychological ingredients: a modern psychological constructionist model. Emot. Rev. 2013; 5:356–368.

47. Ekman P. An argument for basic emotions. Cogn. Emot. 1992; 6:169–200.

48. Johnson-Laird PN, Oatley K. Basic emotions, rationality, and folk theory. Cogn. Emot. 1992; 6:201–223.

49. Saarimäki, H., et al. Discrete neural signatures of basic emotions.. Cereb. Cortex. 2015. Published online April 29, 2015. http://dx.doi.org/10.1093/cercor/bhv086

50. Kragel PA, LaBar KS. Multivariate neural biomarkers of emotional states are categorically distinct. Soc. Cogn. Affect. Neurosci. 2015; 10:1437–1448. [PubMed: 25813790]

51. Mauss IB, et al. The tie that binds? Coherence among emotion experience, behavior, and physiology. Emotion. 2005; 5:175. [PubMed: 15982083]

52. Wager TD, et al. A Bayesian model of category-specific emotional brain responses. PLoS Comput. Biol. 2015; 11:e1004066. [PubMed: 25853490]

53. Lindquist, KA., et al. The brain basis of positive and negative affect: evidence from a meta-analysis of the human neuroimaging literature.. Cereb Cortex. 2015. Published online January 28, 2015. http://dx.doi.org/10.1093/cercor/bhv001

54. Clark-Polner E, et al. Multivoxel pattern analysis does not provide evidence to support the existence of basic emotions. Cereb. Cortex. Published online February. 2016; 29:2016. http://dx.doi.org/10.1093/cercor/bhw028.

55. Kragel PA, LaBar KS. Multivariate pattern classification reveals autonomic and experiential representations of discrete emotions. Emotion. 2013; 13:681–690. [PubMed: 23527508]

56. Barrett LF. The conceptual act theory: a précis. Emot. Rev. 2014; 6:292–297.

57. Damasio A, Carvalho GB. The nature of feelings: evolutionary and neurobiological origins. Nat. Rev. Neurosci. 2013; 14:143–152. [PubMed: 23329161]

58. Anderson DJ, Adolphs R. A framework for studying emotions across species. Cell. 2014; 157:187–200. [PubMed: 24679535]

59. Scherer KR. What are emotions? And how can they be measured? Soc. Sci. Inform. 2005; 44:695–729.

60. Lazarus RS. Progress on a cognitive–motivational–relational theory of emotion. Am. Psychol. 1991; 46:819. [PubMed: 1928936]

61. Skerry AE, Saxe R. Neural representations of emotion are organized around abstract event features. Curr. Biol. 2015; 25:1945–1954. [PubMed: 26212878]

62. Yarkoni T, et al. Large-scale automated synthesis of human functional neuroimaging data. Nat. Methods. 2011; 8:665–670. [PubMed: 21706013]

63. Tzourio-Mazoyer N, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage. 2002; 15:273–289. [PubMed: 11771995]

64. Logothetis NK. What we can do and what we cannot do with fMRI. Nature. 2008; 453:869–878. [PubMed: 18548064]

65. Logothetis NK, et al. Neurophysiological investigation of the basis of the fMRI signal. Nature. 2001; 412:150–157. [PubMed: 11449264]

66. Kriegeskorte N, et al. How does an fMRI voxel sample the neuronal activity pattern: compact-kernel or complex spatiotemporal filter? Neuroimage. 2010; 49:1965–1976. [PubMed: 19800408]

67. Rutishauser U, et al. The primate amygdala in social perception – insights from electrophysiological recordings and stimulation. Trends Neurosci. 2015; 38:295–306. [PubMed: 25847686]

68. Murray RJ, et al. The functional profile of the human amygdala in affective processing: insights from intracranial recordings. Cortex. 2014; 60:10–33. [PubMed: 25043736]

69. Salzman CD, Fusi S. Emotion, cognition, and mental state representation in amygdala and prefrontal cortex. Annu. Rev. Neurosci. 2010; 33:173–202. [PubMed: 20331363]

70. Oya H, et al. Electrophysiological responses in the human amygdala discriminate emotion categories of complex visual stimuli. J. Neurosci. 2002; 22:9502–9512. [PubMed: 12417674]

71. Lachaux JP, et al. Relationship between task-related gamma oscillations and BOLD signal: new insights from combined fMRI and intracranial EEG. Hum. Brain Mapp. 2007; 28:1368–1375. [PubMed: 17274021]

72. Guillory SA, Bujarski KA. Exploring emotions using invasive methods: review of 60 years of human intracranial electro-physiology. Soc. Cogn. Affect. Neurosci. 2014; 9:1880–1889. [PubMed: 24509492]

73. Borchers S, et al. Direct electrical stimulation of human cortex – the gold standard for mapping brain functions? Nat. Rev. Neurosci. 2012; 13:63–70. [PubMed: 22127300]

Box 1. Methodological Considerations for Studying the Neural Basis of Emotion

Although fMRI is advantageous because it can noninvasively sample neural activation spanning the whole brain, this neuroimaging method has several limitations when used to map emotion onto the brain. In terms of spatial resolution, voxels typically have an in-plane resolution on the order of 10 mm^2 and have been estimated to contain, on average, over 5 million neurons [64]. Moreover, simultaneous electrophysiological recording and fMRI acquisition in macaques has shown that the BOLD response best correlates with local field potentials, indicative of inputs and local processing of a brain region as opposed to its spiking output [65]. Consequently, the activation ofasingle voxel can be driven by diverse neural populations and can be regarded as a complex spatiotemporal filter [66] rather than a simple summation of neuronal activity over space and time. This mismatch between the spatiotemporal resolution of fMRI and the neural substrates underlying emotional behavior makes it unlikely that a single voxel will demonstrate emotion-specific activation (i.e., consistently exhibit increased activation for one emotion but notfor other emotions [4]), even ifspecialized neurons reside within a voxel.

Electrophysiological recording methods such as measuring local field potentials or single-unit recordings provide a means to quantify emotion-related neuronal activity at the cellular level (e.g., see [67–69] for reviews of affective processing in the human and nonhuman primate amygdala). One such study [70] measured local field potentials from depth electrodes in the amygdala and found that aversive stimuli – threatening images in particular – showed elevated gamma-band power, which is most closely linked to the BOLD fMRI response in humans [71]. Although there are relatively few studies relating electrophysiological measures to the subjective experience of distinct emotions (compared with functional neuroimaging), data collected using these invasive methods are generally consistent with evidence from functional neuroimaging [72], demonstrating that neuronal representations ofemotion categories are distributed across a number ofcortical and subcortical brain regions. Future work manipulating neuronal activity and measuring the impact on the experience of emotion (e.g., using electrical stimulation mapping [73]) will inform causal brain-behavior relationships that can facilitate functional interpretations.

Box 2. Representational Spaces for Emotional States

A major advantage of MVPA over subtraction-based methods is its capacity to efficiently relate representational spaces to one another [15]. By treating each multivariate pattern of BOLD response as a point in high-dimensional space, theoretical models can be directly related to distributed patterns of neural activity, either within local regions or across the whole brain. Thus, MVPA offers a frame work for constraining conceptual or computational theories of emotion with neural data.

The usefulness of representational spaces is evident in contrasting different cognitive models ofemotion, which predict different relationships among emotions. In one account of basic emotions [48], happiness, sadness, disgust, fear, and anger are considered functionally independent because they are associated with specific antecedent events: making progress towards a goal, the loss of a goal, perceiving something to reject, perceiving a threat to survival, or blocking of a goal. Alternatively, a circumplex model of emotions derived from judgments about emotion concepts and self-reported emotional experience [22] suggests that these emotions are fundamentally represented along affective dimensions of valence (pleasantness) and arousal (activation).

To quantitatively map these models onto distributed patterns of neural activity, representational spaces can be constructed based on their core assumptions (Figure 2). In one formulation of a dimensional model, emotions are represented in a 2D space characterized by valence and arousal axes. This model assumes that instances offear, anger, and disgust are less distinct from one another than sadness and happiness because they share negative valence and high arousal. For this model, emotions tend to cluster in a low-dimensional space. In the basic emotions view, instances of different emotions are approximately equidistant from one another and form distinct categories. In this model, emotions are sparse, relatively independent, and span a higher-dimensional space. With these model-based representational spaces in hand, MVPA can be applied (e.g., through decoding or representational similarity analyses) to identify brain regions with consistent representational geometries. Model comparison can be directly conducted to identify the theoretical perspectivethat best explainsthe activity patterns; for instance, by determining whetherapattern classifier's errors conform more to one model or another [14].

Trends

Due to limitations of univariate approaches, scientists have begun to apply multivariate statistical tools to decode how emotion constructs are represented in high-dimensional patterns of human brain activity.

Recent studies show that functional neuroimaging data can be accurately classified along affective dimensions and discrete emotion categories.

Data from studies classifying brain states into multiple emotion categories suggest that dimensions of valence and arousal do not principally organize neural representations of specific emotions.

Outstanding Questions

Do patterns of emotion-specific neural activation play a causal role in the experience of emotion?

How does emotion regulation affect neural biomarkers of specific emotional states?

Can multivariate neural biomarkers be used to track the dynamics of emotions over time?

Given that basic emotions are thought to be driven by innate circuitry, can high-resolution functional imaging provide novel insight into their representation in the local activity of subcortical neural networks?

Do emotion-specific patterns of cognitive appraisal have distinct neural bases?

How do neural representations of emotion change across development or across cultures?

Do neural biomarkers of emotional states have practical utility in diagnosis or predicting clinical outcomes?

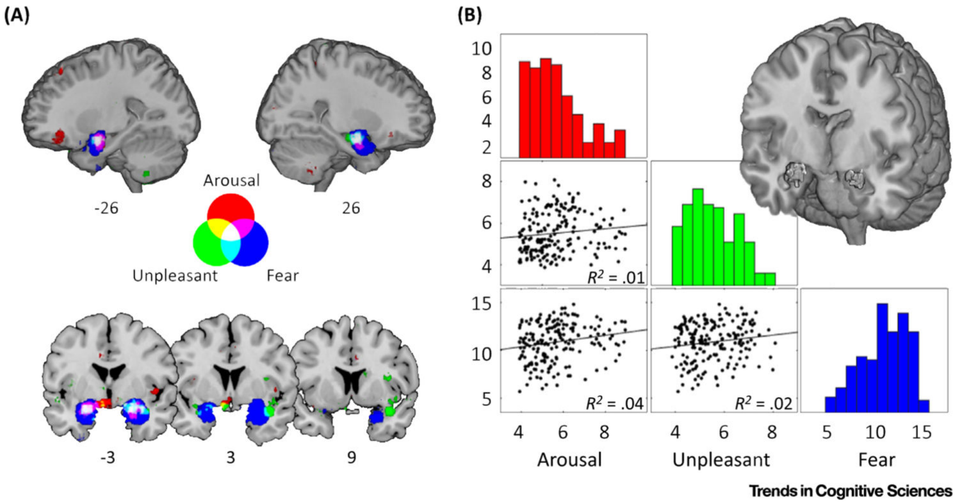


Figure 1. Overlapping Yet Distinct Profiles of Amygdala Activation Predict Experimental Manipulations of Arousal, Unpleasantness, and Fear

(**A**) Probabilistic reverse-inference maps from an automated meta-analysis of the neuroimaging literature [62] indicate the probability of a study including the terms 'arousal' (227 studies), 'unpleasant' (106 studies), or 'fear' (298 studies) given the observed activation. Color maps reflect *z*-scores and are additive as indicated by the legend; the white region indicates voxels predictive of all three processes. (**B**) Spatial cross-correlations of amygdala voxels [63] that commonly predict arousal, unpleasantness, and fear (displayed in white). Each point corresponds to a single voxel and solid lines indicate the best least-squares fit. Despite shared localization, patterns of predictive scores for arousal or unpleasantness explain relatively little variance across voxels predictive of fear.

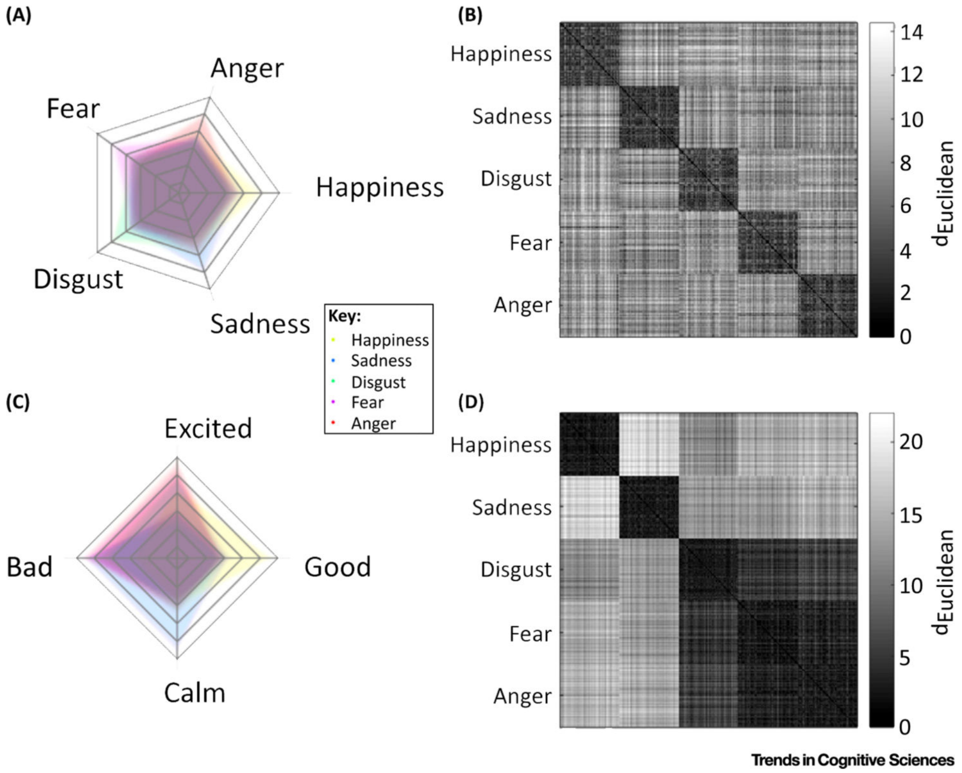
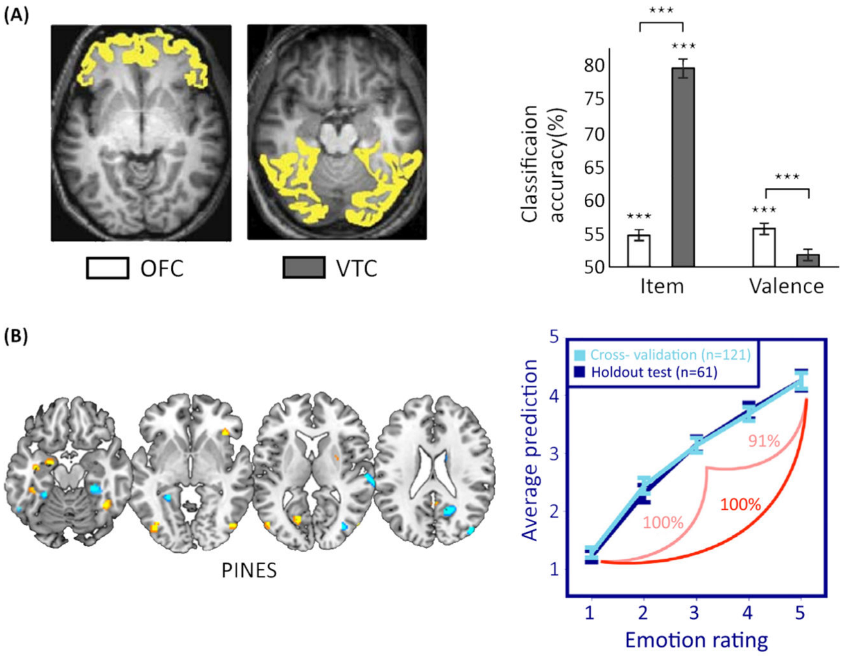


Figure 2.

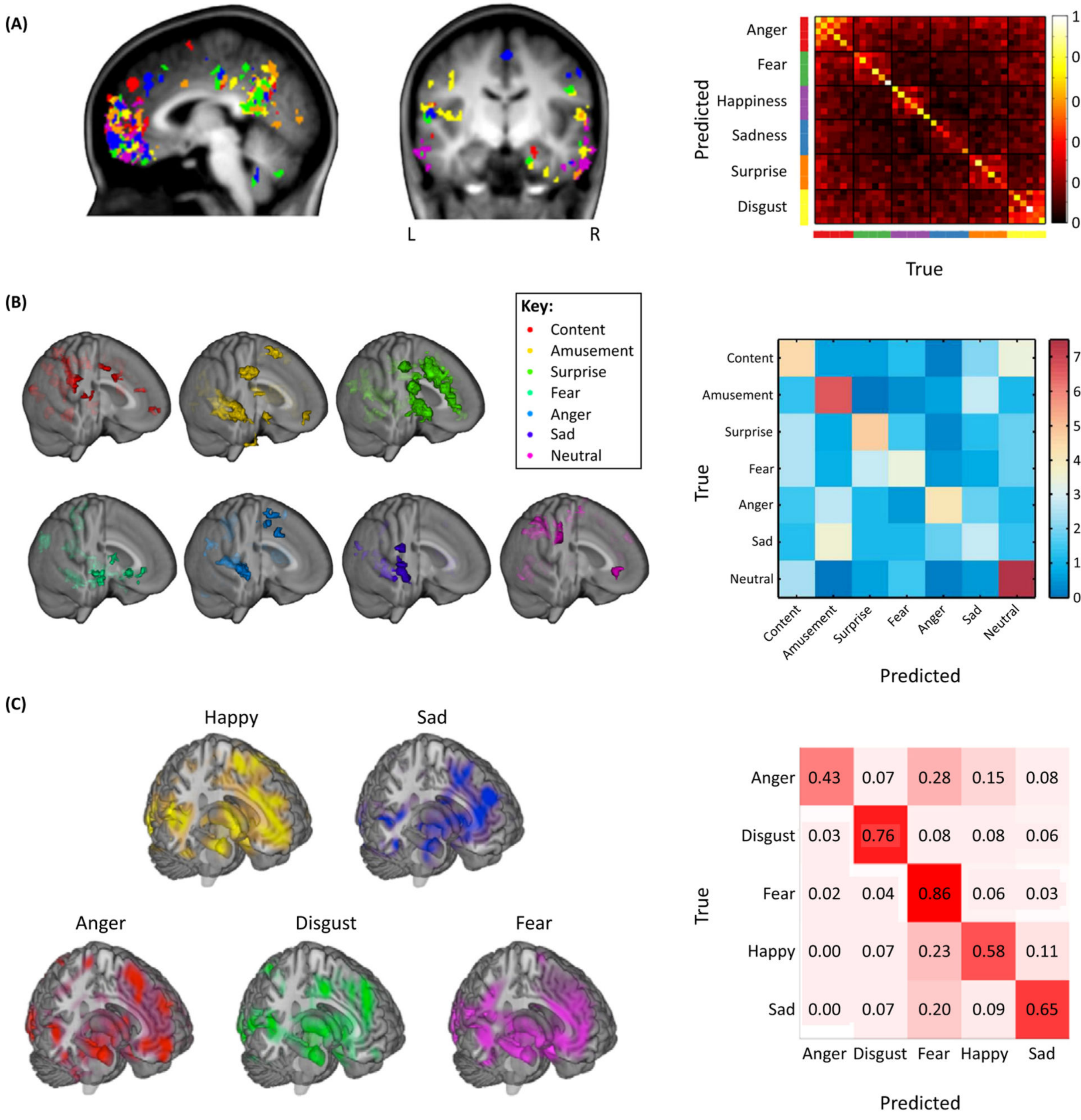
(A) Representational space characterized by basic emotion models. Instances of emotion are drawn from multivariate Gaussian distributions with a simple structure (each distribution is centered along one of five independent dimensions). (**B**) Euclidean distance matrix illustrates how emotions are discrete and equidistant in this representational space. (**C**) Radar plot depicting locations of affective concepts based on a 2D model. Each point corresponds to a single instance of emotion, drawn from one of five multivariate Gaussian distributions with unique locations in valence–arousal space. (**D**) Euclidean distance matrix illustrates how instances of disgust, fear, and anger are proximal to one another in this representational space.



Trends in Cognitive Sciences

Figure 3. Decoding Local and Global Brain Representations along a Continuous Dimension of Valence

(**A**) Double dissociation from [29], wherein patterns of the orbitofrontal cortex (OFC) response to visual stimuli were found to predict differences in subjective valence in novel subjects whereas ventral temporal cortex (VTC) activation predicted the items conveyed in the images. Brain regions highlighted in yellow depict regions of interest used for classification. (**B**) Peak classification weights from the Picture Induced Negative Emotion Signature [36], which maps whole-brain activation patterns to a continuous prediction of negative emotional experience. The model performed in excess of 90% accuracy when testing in independent subjects (A) reproduced, with permission, from [29]; (B) reproduced, with permission, from [36].



Trends in Cognitive Sciences

Figure 4. Brain-Based Models of Discrete Emotion Categories

(**A**) Importance maps (indicating thetop 1% offeatures)computed forwithin-subject classification of six basic emotions in the imagery experiment from [49]. (**B**) Partial least-squares regression coefficients indicate voxels in which activation reliably predicts the music-and film-evoked emotional states in independent subjects from [50]. (**C**) Intensity maps from the Bayesian Spatial Point Process model developed from the peak coordinates of 148 neuroimaging studies of emotion [52]. The intensity maps indicate the expected number of activations from studies assigned to each emotion category. The confusion matrices indicate the correspondence between ground truth and predicted labels. In general, most entries fall along the diagonal indicating good performance, with few errors between similarly valenced emotions (e.g., fear, anger, and sadness). (A) reproduced, with permission, from [49]; right panel of (B) reproduced, with permission, from [50]; (C) reproduced, with permission, from [52].