Representational Similarity and Pattern Classification of Fifteen Emotional States Induced by Movie Clips and Text Scenarios

Yaohui Ding[1], Nathan M. Muncy[2], John L. Graner[1], Joel S. White[1], Alisa C. Schutz[3], Leonard Faul[4], John M. Pearson[1], and Kevin S. LaBar[1]

1. Duke University
2. University of Notre Dame
3. University of Minnesota
4. Boston College

Keywords: Representational Similarity Analysis, Machine Learning, Hierarchical Clustering, Brain Imaging, Affect

Abstract

Theoretical models emphasize that categorical factors, dimensional factors, or their combination may define the semantic space organization of emotion representations. While recent behavioral work has applied innovative multivariate methods for testing these theories, neuroscientific assessments remain limited due to a focus on a small number of emotions, single theoretical perspectives, univariate methods, or small sample sizes. We overcame these limitations in a comprehensive functional magnetic resonance imaging (fMRI) study in which participants (N = 136) viewed 150 movie clips and 150 text scenarios that reliably induced 15 different emotional states spanning positive, negative, and neutral valence. For the movie inductions, representational similarity analysis yielded a correspondence between the categorical behavioral responses and the brain activity patterns, and partial least squares discriminant analysis achieved strong decoding performance for all 15 emotions from whole-brain fMRI data, with importance maps encompassing cortical, limbic, and subcortical regions. Classification error analyses and a Bayesian model comparison supported the categorical nature of the emotion representations relative to a 2-dimensional arousal-valence model. Hierarchical clustering of the representational dissimilarity matrices revealed that the 15 emotions were organized into similarly meaningful clusters at both the subjective and neural levels. Results from the scenario inductions, while demonstrating similar behavioral effects and behavioral hierarchical structures, were more difficult to decode from the fMRI data. Overall, these findings provide novel insights into how emotions are organized and represented in the human brain and evidence a relationship between our subjective experience of and brain responses to emotional inductions.

Significance Statement

We found a robust relationship between participants' categorical endorsement of and their neural responses to emotionally evocative movie clips using an unsupervised machine learning approach. Subsequently, we were also able to decode the emotional content of these movie clips using a supervised classification technique. Finally, we found that these emotions are organized in a very similar ways both in terms of participants' self-report and their brain responses to these stimuli. Together, these data-driven and computational modeling based findings from our study significantly advance our understanding of how emotions are organized and represented in our mind and brain.

Introduction

How are emotions organized in our mind and brain? This is one of the central questions in the field of affective neuroscience with deep scientific and vast clinical implications. Different theoretical orientations make different predictions about which facets of emotional experience are important for distinguishing emotions from one another. Discrete emotion theories predict that emotions tend to be experienced in distinct universal categories that have unique evolutionary trajectories and brain systems, at least for a few basic emotions (Ekman, 1992; Panksepp, 1982). Dimensional theories postulate instead that specific emotions emerge from a small set of intersecting emotional dimensions like arousal and valence (Russell, 1980). Appraisal theories include other dimensions of experience that are important for cognitively evaluating the meaning of an emotional situation as it dynamically unfolds, some of which are more universal like novelty and some of which are more culturally specific, like moral appraisals that contribute to normative significance evaluation (Scherer, 2001). Finally, constructionist theories postulate that each instance of emotion is constructed uniquely in the service of allostatic goals based on contextual demands, predictive coding, internal models of the body in the world, the learning history of the individual, and sociocultural factors (Barrett, 2017a, 2017b).

Each of these theories makes different predictions about how emotions might be represented in the brain. Unfortunately, there is no current consensus to guide neuroscientific investigations other than that there is no one-to-one correspondence between a given emotion and univariate activity in a given brain region (Lindquist et al., 2012). With the rise and maturation of multivariate pattern classification (MVPA) techniques (Kragel & LaBar, 2014), researchers started to investigate whether emotional states could be decoded or classified from multi-voxel patterns of blood oxygenation level dependent (BOLD) activity in a data-driven way. For example, Kassam et al. (2013) decoded nine emotional states based on the neural responses from the same participant, other participants, and a different emotion induction modality. Kragel and LaBar (2015) decoded seven emotional states induced from movie and music and showed that the neural biomarkers of these emotional states are categorically distinct and widely distributed across the entire brain. In this work, they also found that classification errors, i.e., the extent to which one emotion was confused with another emotion, were best explained by the pairwise distances of these emotions in a 7-dimensional space (spanned by the seven categories). Similarly, Saarimäki et al. (2016) found specific neural signatures of six basic emotions (disgust, fear, happiness, sadness, anger, and surprise) induced by two stimulus modalities, short movies and mental imagery. Saarimäki et al. (2018) showed that there was a robust correspondence between the behavioral and neural similarities of fourteen different emotional states. Additionally, they also found that the neural bases of these fourteen emotions included brain regions that were widely distributed across the cortical and subcortical areas. These studies overcame some limitations of earlier studies, e.g., the static nature of the emotional stimuli, a small set of emotions, and the univariate approach and answered a wide array of questions regarding how emotions are represented in the brain. However, they were still limited by their small sample sizes (Kragel & LaBar, 2015; Saarimäki et al., 2016), lack of involvement of subcortical regions (Kragel & LaBar, 2015), and a single theoretical perspective (Saarimäki et al., 2018).

Semantic space theory

The recently developed computational and data-driven approach to emotion research—the semantic space theory (Cowen & Keltner, 2021; Keltner et al., 2023)-- provides a novel theoretical framework for studying emotion. According to this theory, emotions can be understood in a semantic space that is characterized by three major aspects or questions of our emotional experiences: *dimensionality*, *distribution*, and *conceptualization*. The dimensionality aspect is

concerned with determining the number of meaningful experiences or expressions that are distinct (Cowen & Keltner, 2021) within the semantic space. We may think of it as the number of orthogonal dimensions needed to adequately describe human emotions. For example, according to the circumplex model of affect (Russell, 1980), only two dimensions (valence and arousal) are needed to describe human emotion. The distribution aspect of semantic space theory captures the clustering or blending of emotional states. It questions the relative position of emotions within the semantic space, including overlaps and boundaries. Finally, the conceptualization aspect of semantic space theory is interested in how emotions are perceived by those experiencing them. It attempts to identify the mapping of subjective experience onto emotion models. In Cowen and Keltner (2021), the authors showed that the semantic space theory was able to capture much more complex and nuanced aspects of human subjective experiences to various types of emotional evocative stimuli, e.g., ancient art, face and voice, short video clips, music, and prosody. Critically, these studies emphasized that discrete emotions were represented with fuzzy categorical boundaries that were foundational to the semantic spaces, while dimensions of emotional experience like valence and arousal provided means to bridge these fuzzy categories. In addition, the semantic spaces differed by induction modality, with different modes of inducing emotions eliciting semantic spaces that differed somewhat in the dimensionality and distribution of the semantic space. We note that these studies primarily focused on the subjective experience of emotions based on self-report. To date, the only study that has assessed the neuroscientific evidence of the semantic space theory was Horikawa et al. (2020). In this study, they found that the neural representations of 27 emotions were categorical and involved distributed and transmodal brain regions. Though this study was impressive in terms of the number of emotions that they induced, its generalizability was severely limited because of a small sample size (N=5) and a within-subject approach to representational similarity and decoding analyses.

Research Questions

In the current study, we experimentally induced fifteen emotional states using two modalities, short movie clips (N=150) and text scenarios (N=150). We expand upon previous work by including a large number of emotions, multiple induction methods (movies and scenarios) each with a large number of stimuli, supervised and unsupervised data analytic methods, direct tests of two competing theoretical models using error analyses and a Bayesian model comparison, and finally a large sample size. This design provides a greater number of sample data sets to power our analyses. Guided by the semantic space theory (Cowen & Keltner, 2021), we investigated the following three sets of questions using three different but complementary computational approaches, namely representational similarity analysis, partial least squares discriminant analysis (a type of supervised machine learning technique), and hierarchical clustering analysis.

First set of questions

1. Do participants' categorical endorsement agree (Spearman correlation) with their brain response to the emotional stimuli?
2. Do participants' valence/arousal ratings agree (Spearman correlation) with their brain response to the emotional stimuli?
3. Do participants' categorical endorsement agree (Spearman correlation) with their valence/arousal ratings of the emotional stimuli?

Second set of questions

1. Can we use participants' brain responses to decode the emotion labels of the emotional stimuli they experienced in the scanner?
2. Can we use the categorical endorsement to explain the classification errors?

3. Can we use the valence arousal ratings to explain the classification errors?
4. Do categorical or dimensional theories better explain the data?
5. Which voxels or brain regions are important for decoding each emotion?

Third set of questions

1. Is there an intuitive and meaningful organization in terms of participants' categorical endorsement of the emotional stimuli?
2. Is there a meaningful organization in terms of participants' valence/arousal rating of the emotional stimuli?
3. Is there a meaningful organization in terms of participants' brain response to the emotional stimuli?
4. Finally, do we see similar structures from these three sources of measurement of these emotional stimuli?

Materials and Methods

Participants

The study was approved by the Institutional Review Board (Approval Number: Pro00105943) at Duke University Health System. All participants provided verbal and written informed consent prior to taking part in the study. One hundred sixty six healthy English-speaking participants between 18-49 years old (93 female; mean age=30.44 ± 8.55) enrolled in the current study. Of these participants, one hundred thirty six (77 female; mean age=30.35±8.46) of them completed at least one session of MRI. Additionally, five pilot participants were run before the experimental and scan protocol were finalized. Therefore, their data were not included in the final analyses. All participants must have obtained at least a high school diploma or equivalent, had no history of neurological illness, and had no current psychiatric diagnoses or psychotropic medication use. Participants were excluded if they had a substance abuse disorder, history of breathing or heart-related conditions, intracranial or medical implants, transdermal delivery systems, non-removable piercings, were pregnant, or claustrophobic. Each participant was compensated at a rate equivalent to $25/hour.

Experimental design and stimuli

The study consisted of three sessions, named Visit 1, Visit 2, and Visit 3, respectively. Visit 1 and Visit 2 were separated by at least 7 days, Visit 2 and Visit 3 by at least 5 days. Visit 1 was entirely online and consisted of participants filling out seven surveys: the Affect Intensity Measure (Bryant et al., 1996), Affect Liability Scale (Harvey et al., 1989), Emotion Regulation Questionnaire (Gross & John, 2012), Penn State Worry Questionnaire (Meyer et al., 1990), Ruminative Response Style (Nolen-Hoeksema & Morrow, 1991), State-Trait Anxiety Inventory (Spielberger et al., 1971), and the Toronto Alexithymia Scale (Bagby et al., 1994). We note that because we did between-subject (limited by the number of samples from each participant) representational similarity and decoding analyses, we were not able to correlate any metrics from these analyses with individual differences from these psychometric surveys. Visit 2 and 3 were MRI experiments where participants took part in an emotion induction task while they were in the scanner. We used 150 validated, emotionally-evocative, short (3 to 8 s) movie clips (Cowen and Keltner (2017) and 150 validated, one- or two-sentence scenarios (Faul et al. (2023) for the emotion induction task. Both the movie clips and scenarios were divided into 15 emotion categories based on normative ratings (10 stimuli per category per induction method). Five consecutive stimuli of the same emotion category were grouped into a block (two blocks per category per induction method) and presentation order was randomized across emotion blocks. The movie clips and scenarios were presented to the

participants in two separate MRI sessions. Participants were first presented with either the movie clips or text scenarios on a TV screen for 8 to 13 seconds. Then they were instructed to replay these stimuli in their mind for 5 seconds. Afterwards, they responded to a question about where the events of each stimulus took place (indoor/outdoor) via button press. Next, they were asked to endorse an emotion that they subjectively experienced the most while watching and replaying each emotional block, and to what degree they experienced it on a scale from 1 to 10 (Figure 1-A). Finally, the participants were presented with a scrambled black and white image for 8 to 12 seconds, which were referred to as "washout". See Supplementary Materials--S1 for more detailed descriptions of these stimuli and S2 for a more detailed description of the emotion induction task.

MRI acquisition

We collected T1, BOLD, and resting state fMRI data from each participant. The MRI protocol on each session consisted of a localizer acquisition, a high-resolution T1-weighted acquisition, a reverse-phase-encode-direction (RPED) echo planar imaging (EPI) acquisition, nine EPI acquisitions (8 task runs and a resting-state run), and another RPED EPI acquisition. The T1-weighted acquisition had the following parameters: acquisition matrix = 256 x 256, repetition time = 2250 ms, echo time = 3.12 ms, field of view = 256 mm, in-plane voxel size = 1.0 x 1.0 mm, slice thickness = 1.0 mm, spacing between slices = 0 mm, 192 axial AC-PC aligned slices. EPI acquisitions had the following parameters: acquisition matrix = 128 x 128, repetition time = 2000 ms, echo time = 30 ms, field of view = 256 x 256 mm, in-plane voxel size = 2.0 x 2.0 mm, slice thickness = 2.0 mm, spacing between slices = 0 mm, phase-encode direction = anterior-to-posterior, 69 axial slices angled 30 degrees positive of AC-PC, multiband acceleration factor = 3, in-plane acceleration factor =2. Runs 1, 2, 3, 5, 6, and 7 of the functional task had 262 time-points and were 8 minutes and 44 seconds long. Runs 4 and 8 had 200 time-points and were 6 minutes and 40 seconds long. The resting state acquisition had 240 volumes and were 8 minutes long. The RPED EPI acquisitions used the same parameters as the functional task EPI acquisitions except a phase-encode direction of posterior-to-anterior and 2 time-points.

Functional MRI data preprocessing

Preprocessing and modeling of functional MRI (fMRI) data were conducted with the following software suites: dcm2niix v1.0.20211006 (Li et al., 2016), Convert 3D v1.1.0 (Yushkevich et al., 2006), FreeSurfer v7.2.0 (Fischl, 2012), fMRIPrep v23.0.2 (Esteban et al., 2019), MRIQC v23.1.0 (Esteban et al., 2017), FMRIB Software Library v6.0.6.4 (Jenkinson et al., 2012), and Analysis of Functional NeuroImages v24.2.01 (Cox, 2012).

Preprocessing. Anatomic T1-weighted and functional T2*-weighted DICOMs were converted into NIfTI files, and FreeSurfer was then pre-run via the recon-all -all command to generate a set of anatomic parcellations. FMRIPrep next utilized the FreeSurfer output, T2*w run, and T2*w field map files to conduct a workflow which produced preprocessed fMRI data warped into MNI152NLin6Asym template space; see Supplemental Section [S1] for full details. Finally, additional preprocessing was conducted on fMRIPrep output to prepare the fMRI data for first-level modeling. These extra preprocessing steps consisted of bandpass filtering the functional data and then scaling the output such that all fMRI data having a common y-axis range. Additionally, we note that while the ICA-AROMA procedures were used during the fMRIPrep.

Feature extraction

First-level models of the functional data were conducted via FSL's feat command (Woolrich et al., 2001), which fits a general linear model (GLM) to each run of the fMRI data. The main regressors were the general task events (in/outdoor judgments, washout screens, emotion label selection, and

emotion intensity selection), and emotion events (stimulus and replay, modeled separately, for each emotion). Nuisance regressors in the GLM included averaged CSF and WM signals, frame-wise motion derivative, and translation and rotation parameters with their derivative), and a confounds file (volumes with a framewise displacement > 0.5 mm were censored). The emotion stimuli blocks vs. washout were the GLM's parameters of interest. All of these regressors were convolved with a double-gamma hemodynamic response function (HRF) before entering into the GLM. Following first-level modeling, the voxel-wise emotion vs. washout parameter was extracted from the respective cope file, resulting in >900K voxel values. This large number of parameters was due both to the spatial resolution of the data as well as the inclusion of extra cerebral, WM, and CSF voxels. To reduce the number of values, then, we masked the data with a gray matter mask (cortical and subcortical) constructed from the template prior tpl-MNI152NLin6Asym_res-02_atlas-HOSPA_desc-th25_dseg. Masking reduced the number of values to approximately 150K. Masks of seven canonical brain networks (Thomas Yeo et al., 2011), control, default, dorsal attention, limbic, Salience/Ventral Attention, Somatomotor, and Visual, were created from ICA decompositions of the functional resting-state images collected as part of this study.

Representational Similarity Analysis (RSA)

Neural dissimilarity matrix

Whole-brain voxel-wise beta weights from mass univariate general linear models were used to calculate the neural dissimilarity matrix among the fifteen experimentally induced emotions. In total, we had 3143 blocks of beta weights from the movie stimuli and 3247 blocks of beta weights from the scenario stimuli. Each emotion condition had roughly 240 blocks. We first averaged the beta weights within each emotion condition to get an averaged activation pattern for that emotion. Then, we computed pairwise cosine similarities of these averaged activation patterns and used one minus these pairwise cosine similarities to form the neural dissimilarity matrix. Separate neural dissimilarity matrices are calculated for the neural responses to movie and scenario stimuli.

Dissimilarity matrix based on participants' self-report

We defined two types of behavioral dissimilarity matrices from participants' behavioral responses to the emotional stimuli. For each of the emotion blocks that participants experienced in the scanner, we have their categorical endorsement of these blocks. From these categorical endorsements, we computed a confusion matrix by comparing participants' endorsements with the ground truth labels from previous studies (Cowen & Keltner, 2017; Faul et al., 2023). Then we computed pairwise distances (1-cosine similarity) among all the emotions. Again, these pairwise distances were Box-Cox transformed (Box & Cox, 1964) and then used to form the categorical dissimilarity matrix. Because we also had post scan valence and arousal ratings of the stimuli, we calculated another behavioral dissimilarity matrix based on these valence and arousal ratings. More specifically, we first averaged the valence and arousal ratings of within each block (5 stimuli per block) and then within each emotion (roughly 240 blocks per emotion). Then we computed pairwise Euclidean distances among all the fifteen emotions and used these pairwise distances (Box-Cox transformed) to form the valence-arousal (continuous) dissimilarity matrix.

Bootstrap confidence interval and permutation testing

We used a nonparametric bootstrap procedure (Tibshirani & Efron, 1993) to estimate the confidence intervals of the representational similarity between neural and behavioral responses. We resampled the upper triangular entries of the neural RDM, with replacement, 5000 times to generate the bootstrap samples. We then computed the Spearman rank correlation between the original behavioral RDM and the resampled neural RDM. We used a bootstrap procedure (number

of bootstraps = 5000) to get the confidence intervals of the Spearman correlations between the neural and behavioral dissimilarity matrices. This procedure gave us a distribution of the Spearman correlation coefficients from these bootstrap resamples. From this distribution, we calculated the 95% confidence interval using the percentile method (i.e., the 2.5th and 97.5th percentiles).We also performed permutation testing (number of permutations = 5000) to derive the p values of the Spearman correlation coefficients. More specifically, we permuted the labels of the neural data before computing the averaged BOLD activity within each emotion condition. The rationale is that if there is no meaningful association between patterns of neural activity of these emotion blocks and the behavioral ratings of the emotion blocks, then we could randomly permute the labels of the emotion blocks and get a similar association between the neural and behavioral dissimilarity matrices. However, if the associations between the neural and behavioral dissimilarity matrices that we get from the permuted labels are markedly different from the association that we get from the un-permuted emotion blocks, then it suggests that the association is robust. We used the following formula to calculate permutation p values. All analyses were conducted in MATLAB using custom scripts.

$$p = \frac{number(\rho_{permuted} > \rho_{observed}) + 1}{n_{perm} + 1}$$

Multi-Voxel Pattern Classification

Multiclass to binary classification

We used a one versus all coding scheme to convert the multiclass classification problem into fifteen binary classifications. This means that we trained a different classifier to distinguish each emotion (e.g., amusement) against all the other emotions. In addition to the computational efficiency of this approach, one of the main reasons that we took this approach was because we wanted to identify the voxels (emotion map) that differentiated a given emotion against all the other emotions.

Nested cross validation

Overfitting is a common problem in supervised machine learning, especially when the number of predictors is much larger than the number of samples. We used a nested cross validation scheme to improve the generalizability of our classifiers (Lewis et al., 2023). More specifically, the data were first divided into eight subject independent folds (outer loop). The samples from each subject are only in the same fold, thus preventing information leakage. Each seven folds of the data were used as the training data and the remaining eighth fold as the test data. Each training data set is then divided into five folds (inner loop), again in a subject independent manner. Four of these inner folds were used as training set and the remaining fifth fold as the validation set.

Anomaly detection

Outliers are hard to detect in high-dimensional spaces due to their geometric properties. In this study, we used the isolation forest algorithm (Liu et al., 2008) for anomaly detection. Crucially, the algorithm was applied to each training fold from the outer loop after the data have been partitioned according to the nested cross validation scheme described in the previous section. This procedure prevents information leakage. The algorithm outputs an anomaly score for each sample in the data set. We used 0.1% (1 out of 1000) as the threshold for removing potential outliers; if a sample's anomaly score were in the top 0.1 percentile, then it would be considered as an outlier and removed from the data set.

Feature transformation

Voxel-wise beta weights were z-scored before being fed into the classifiers to ensure that all voxels were weighted equally before the classifiers were trained. The z scoring was done separately on the training and test data sets to prevent information leakage.

Partial least squares discriminant analysis

There are many techniques available when it comes to multi-voxel pattern classification (Kragel et al., 2018). In this study, we chose partial least squares discriminant analysis (Li et al., 2018) to be consistent with our prior fMRI decoding study using a smaller number of emotions (Kragel & LaBar, 2015). The input to the classifier consists of beta weights from all emotion blocks. The outcome variable is the normed label of each emotion block. Partial least squares discriminant analysis (PLS-DA) is a linear supervised classification approach that consists of partial least squares (PLS) and linear discriminant analyses (LDA). During the PLS step, samples were first projected onto a lower dimensional (e.g., 2-dimensional) latent space. Crucially, during this dimensionality reduction step, the labels of the samples were used to weight (or constrain) the covariance matrix. As a result, the successive components extracted from PLS make the samples from different classes as separable as possible in the latent space, instead of extracting as much variance from the data set as possible. Thus, PLS can be understood as supervised principle component analysis (Ruiz-Perez et al., 2020). Once the samples were projected onto a latent dimensional space, linear discriminant analysis is done to classify samples into different classes. Because PLS is linear, the coefficients (or weights) associated with the voxels are straightforward to interpret. The projection to latent space approach allowed us to investigate the question of how many latent dimensions are needed to achieve maximum classification accuracy. Another advantage of the PLS-DA approach is that there is only one hyperparameter that needs to be tuned, namely the number of latent dimensions.

Multiclass classification performance metrics

The predicted scores from each classifier were combined into a single matrix. With fifteen classifiers, we get a matrix of 15 columns of predicted scores. Each row of the matrix represents the predicted scores of the sample from the 15 classifiers. These predicted scores are then transformed into probability scores using a Sigmod function. To decide which emotion category the sample came from, we used a winner-takes-all approach, i.e., the sample is classified into the class from the classifier with the highest predicted score. We repeated this process for each row of the matrix to obtain a vector of predicted emotion labels for all samples. A confusion matrix was then computed by comparing the predicted emotion labels of all the samples with their ground truth labels. From this multiclass confusion matrix, we first counted the following class-specific quantities: true positive, true negative, false positive and false negative. Then we computed the following class specific metrics: accuracy, sensitivity, specificity, and precision. Finally, we computed these macro-averaged multiclass performance metrics, balanced accuracy, macro-averaged sensitivity, macro-averaged specificity, macro-averaged precision, and macro-averaged F1-score. All of the formulas used are defined in this paper (Rainio et al., 2024). They are also tabulated in Table 13 in S7-Tables for easy reference.

Bootstrap confidence intervals and permutation testing

Similar to the RSA analysis, we used a bootstrap procedure (number of bootstraps = 5000) to get the confidence intervals of these multiclass classification metrics defined in the previous section. For permutation testing, we randomly shuffled the ground truth labels of all the samples before computing the confusion matrix and subsequently the classification metrics. We repeated this

process 5000 times to get the null distributions of these metrics and calculated their p values using the same equation presented in the RSA section.

Maps of PLS-DA coefficients

The PLS-DA coefficients were first z-scored and then averaged across the eight outer folds. This process is repeated for the PLS-DA coefficients from each classifier. Hence, we obtained fifteen vectors of PLS-DA coefficients, one for classifying each emotion. Next, we applied a threshold of 1.96 (corresponding to a two-tailed p value of 0.05; uncorrected) to these vectors and only retained the voxels whose PLS-DA coefficients were greater than this threshold. This ensemble averaging approach gives a good estimate of relative importance of each voxel in classifying each emotion. These supra-threshold voxels were subsequently clustered using FSL's clustering algorithm (Jenkinson et al., 2012). We used a cluster-forming threshold of 30 voxels to reduce clusters that could potentially be false positive. To find the anatomical labels of these clusters, we used the default atlases in FSL and the Harvard Ascending Arousal Network Atlas (Edlow et al., 2023). Finally, the maps of these coefficients were visualized using MRIcroGL (Rorden, 2025).

Post-hoc classification error analysis and Bayesian model comparison

Similar to Kragel and LaBar (2015), we conducted a post-hoc informational content analysis to examine whether the classification errors can be explained by theoretical predictions of the participants' behavioral ratings of the emotional stimuli. More specifically, we represented each row of the confusion matrix from participant in-scanner endorsement ratings as a vector in a 15-dimensional space, where each emotion category is a unique dimension. Then, we computed pairwise distances of these 15 vectors and operationalized these pairwise distances as the distances among these fifteen emotions in a categorical 15-dimensional space. We also defined pairwise distances using the averaged valence and arousal rating of each emotion in a 2-dimensional valence and arousal space (circumplex model). Finally, we examined the Spearman correlation between the classification errors and pairwise distances among these emotions in both the categorical and dimensional spaces. Additionally, we fit four Poisson regression models to the classification errors. The first model only had a constant intercept term. The second model had a constant intercept plus the pairwise distances from the categorical rating as a regressor. The third model had a constant intercept and the pairwise distances from the dimensional rating as an additional regressor. Finally, the fourth model had a constant intercept term and pairwise distances from both the categorical and dimensional model as regressors. Finally, we did Bayesian model comparison which model explained the classification errors the best.

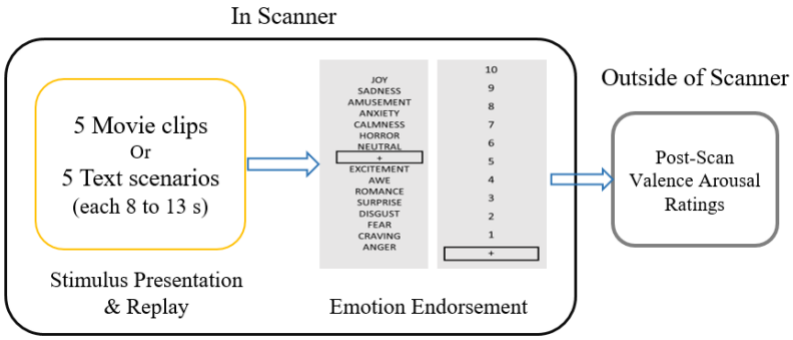
Results

Self-reported emotion endorsement and valence-arousal ratings

In total, we gathered participants' subjective emotion endorsement ratings from 3556 blocks of movie and 3788 blocks of scenario stimuli. Performance metrics computed from participants' emotion endorsement of the **movie** blocks were as follows: *balanced accuracy* (0.7395, CI = [0.7255, 0.7534]), macro-averaged *sensitivity* (0.7395, CI = [0.7255, 0.7534]), macro-averaged *specificity* (0.9814, CI = [0.9803, 0.9824]), macro-averaged *precision* (0.7488, CI = [0.7347, 0.7627]), and macro-averaged *F1-score* (0.7441, CI = [0.7303, 0.7580]). All p values are less than 0.001 (see Table 1 in SI-6: Tables for more information). Emotion-specific performance metrics of participants' endorsement of **movie** blocks are tabulated in Table 2 in SI-6: Tables. Additionally, see SI-5: Figure A-1 for information about the endorsed and expected frequencies of each emotion. See SI-5: Figure B-1 for the averaged intensity rating of each emotion from the movie blocks. Similarly, performance metrics computed from participants' subjective emotion endorsement of the **scenario** blocks were as follows: *balanced accuracy* (0.7031, CI = [0.6882, 0.7174]), macro-

averaged *sensitivity* (0.7031, CI = [0.6882, 0.7174]), macro-averaged *specificity* (0.9788, CI = [0.9777, 0.9798]), macro-averaged *precision* (0.7238, CI = [0.7100, 0.7371]), and macro-averaged *F1-score* (0.7133, CI = [0.6992, 0.7268]). All p values are less than 0.001 (see Table 3 in SI-6: Tables for more information). Emotion-specific performance metrics of participants' endorsement of **scenario** blocks are tabulated in Table 4 in SI-6: Tables. Additionally, see SI-5: Figure A-2 for information about the endorsed and expected frequencies of each emotion. See SI-5: Figure B-2 for the averaged intensity rating of each emotion from the scenario blocks. Taken together, these results suggest that we successfully induced all fifteen emotions in our participants using both movie and scenario stimuli, based on participants' self-reported categorical emotion endorsement and dimensional (valence and arousal) ratings.

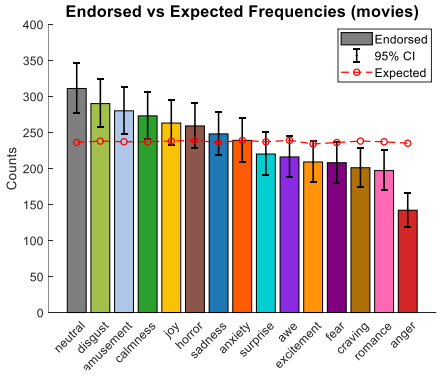
A



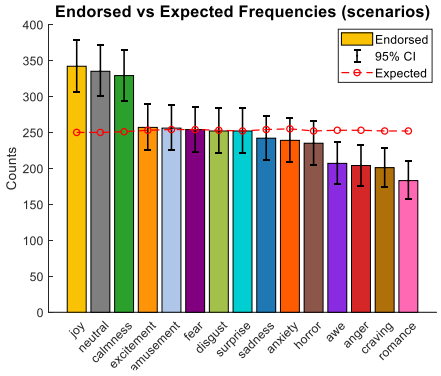
Movie blocks

Scenario blocks

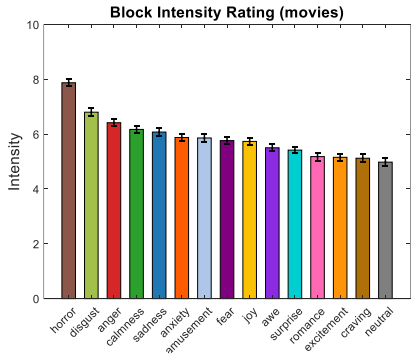
B-1



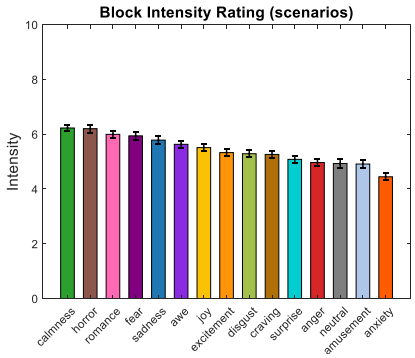
B-2



C-1



C-2



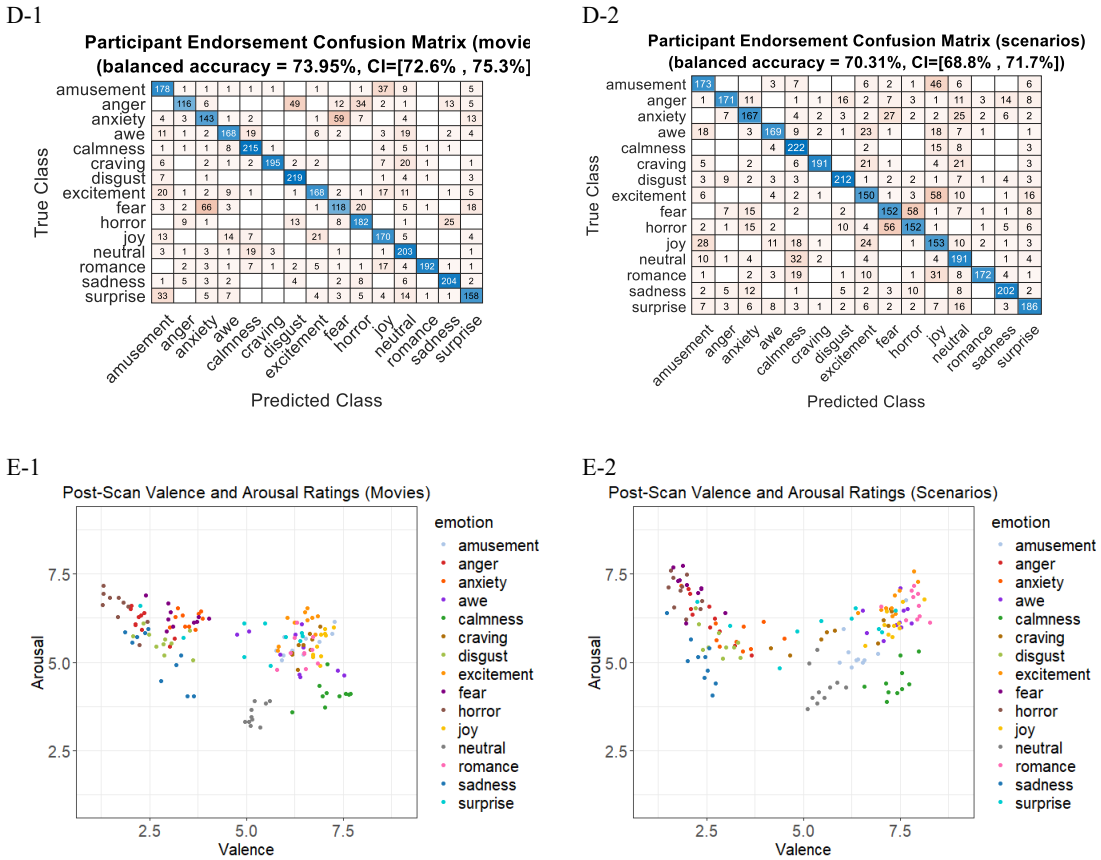
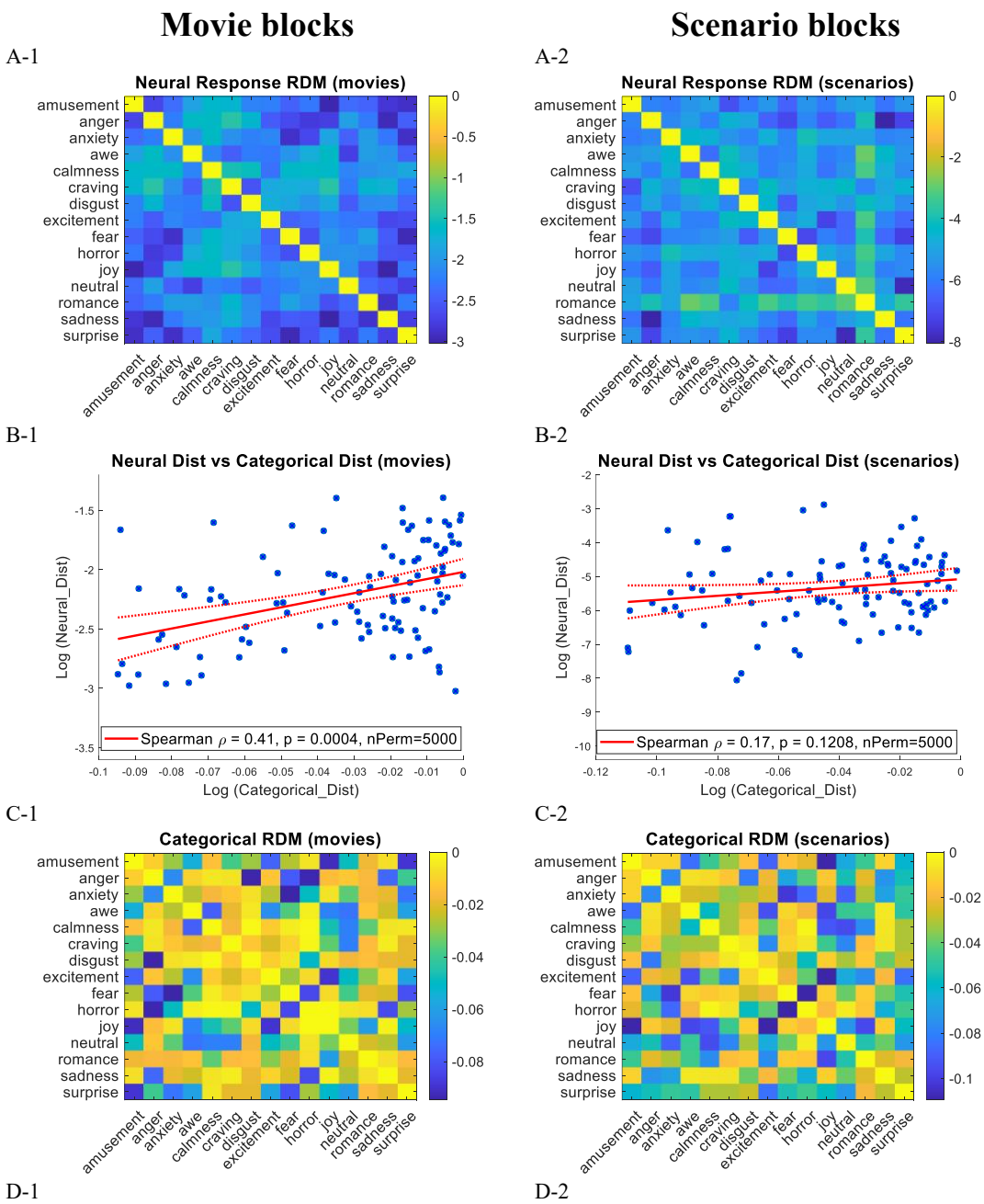


Figure 1. Task Design and Results of Participants' In Scanner Emotion Endorsement and Post-Scan Valence-Arousal Ratings. **A**. An illustration of the task design. Movie clips and scenarios were presented in separate scanning sessions, and a single block of 5 stimuli all came from the same emotion category. **B-1**: Bar plots of endorsed vs expected frequencies of each emotional state (movies). **B-2**: Bar plots of endorsed vs expected frequencies of each emotional state (scenarios). **C-1**: Bar plots of intensity rating of each emotional state (movies). **C-2**: Bar plots of intensity rating of each emotional state (scenarios).**D-1**. Confusion matrix of participant emotion endorsement of movie blocks: balanced accuracy (73.95%, CI= [72.6%, 75.35%]). **D-2**. Confusion matrix of participant emotion endorsement of movie blocks: balanced accuracy (70.31%, CI= [68.8%, 71.7%]). **E-1**. Scatter plot of participants' valence arousal ratings of movie stimuli. **E-2**. Scatter plot of participants' valence arousal ratings of scenario stimuli.

Representational similarity between self-report and brain responses

There is a strong and positive correlation (see Figure 2: D-1 & D-2) between the pairwise categorical distances and pairwise dimensional distances for both the **movie** (Spearman rho = 0.5932, CI = [0.4563, 0.7001], p = 0.0004) and **scenario** (Spearman rho = 0.7509, CI = [0.6713, 0.8057], p = 0.0002) stimuli. See the Behavioral Dissimilarity Matrix subsection of the Representation Similarity Analysis section in Materials and Methods for definitions of categorical and dimensional distances. This result suggests that emotions that are more distant from each other in the valence-arousal space tend to be categorized in different ways. However, the relationship between participants' self-report and brain responses to these stimuli is more nuanced. Specifically, there is a moderate and positive correlation (see Figure 2: B-1) between pairwise categorical distances and pairwise neural distance, but only for the movie stimuli (Spearman rho = 0.4113, CI

= [0.2132, 0.5902], p = 0.0004) and not for the scenario stimuli (Spearman rho = 0.1673, CI = [-0.02761, 0.3518], p = 0.1208). In terms of the relationship between pairwise neural distance and pairwise dimensional distance, we did not find a statistically significant correlation for movie (Spearman rho =0.0846, CI = [-0.1193, 0.2859], p = 0.3004) or scenario (Spearman rho = 0.1465, CI= [-0.06299, 0.344], p = 0.0562) stimuli (See Figure 2: D-1 & D-2). Overall, these results suggest that participants' categorical and dimensional ratings of the emotional stimuli are highly consistent with each other, whereas the correspondence between participants' self-report and brain responses to the stimuli is stronger for the categorical ratings, at least for movie inductions.



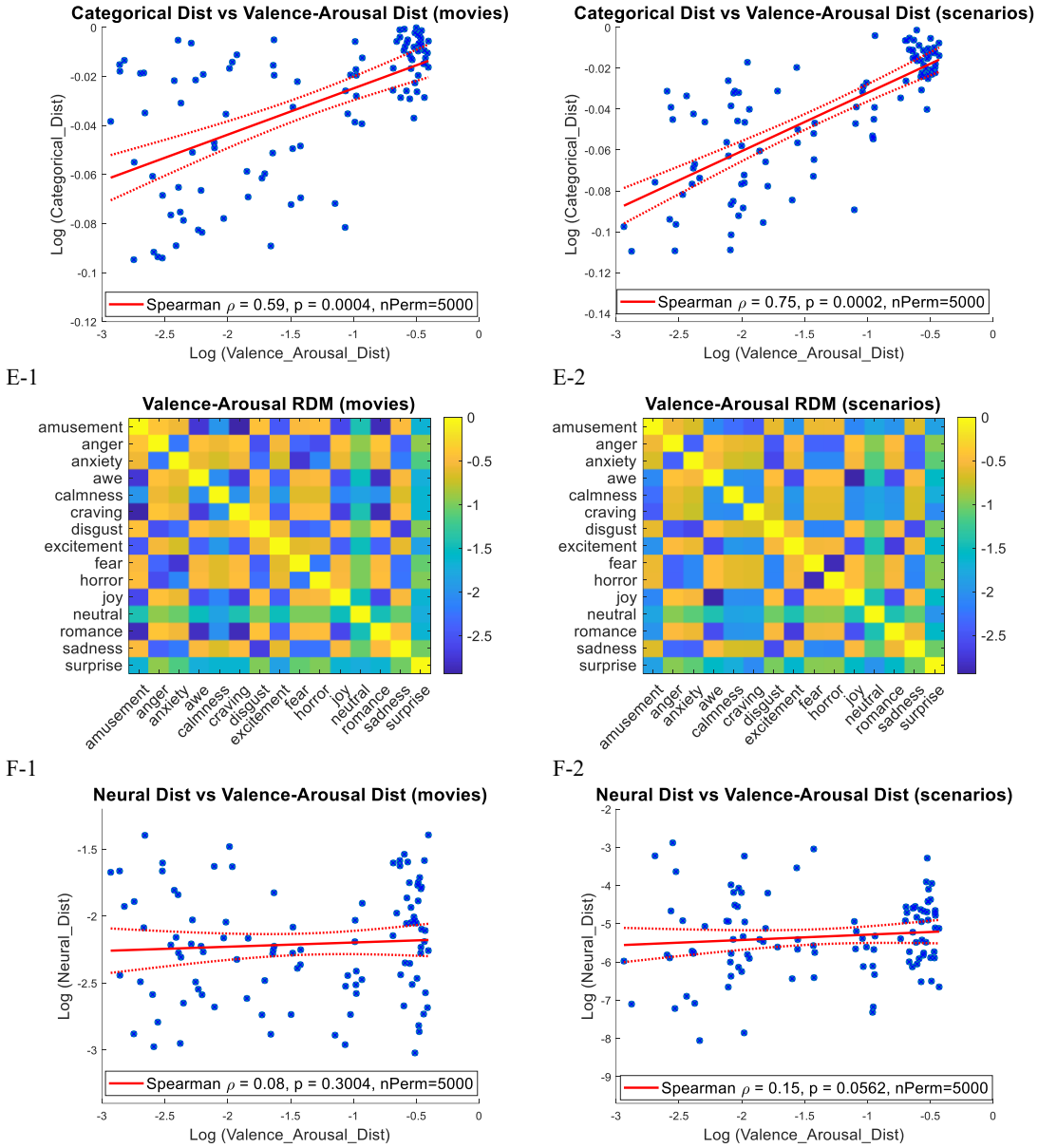


Figure 2. Representational Similarity Analysis between Self-Report and Brain Responses to Movie and Scenario Blocks. **A-1**. Representational dissimilarity matrix (RDM) of patterns of BOLD activity from movie blocks. **A-2**. RDM of patterns of BOLD activity from scenario blocks. **B-1**. Correlation (Spearman rho = 0.4113, p = 0.0004) between pairwise neural distance and pairwise categorical distance (movie blocks) **B-2**. Correlation (Spearman rho = 0.1673, p = 0.1208) between pairwise neural distance and pairwise categorical distance (scenario blocks). **C-1**. RDM of categorical endorsement of the movie blocks. **C-2**. RDM of categorical endorsement of the scenario blocks. **D-1**. Correlation (Spearman rho = 0.5932, p = 0.0004) between pairwise categorical distance and pairwise dimensional distance (movie blocks). **D-2**. Correlation (Spearman rho = 0.7509, p = 0.0002) between pairwise categorical distance and pairwise dimensional distance (scenario blocks). **E-1**. RDM of valence arousal ratings of the movie stimuli. **E-2**. RDM of valence arousal ratings of the scenario stimuli. **F-1**. Correlation (Spearman rho = 0.0846, p = 0.3004) between pairwise neural distance and pairwise dimensional distance (scenario blocks). **F-2**. Correlation (Spearman

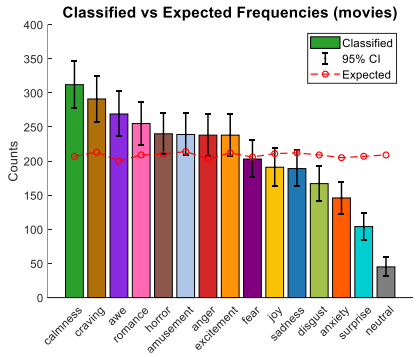
rho = 0.1465, p = 0.0562) between pairwise neural distance and pairwise dimensional distance (scenario blocks). nPerm = number of permutations

Decoding emotional states from movie and scenario blocks

After anomaly detection and outlier removal, 3127 blocks of whole-brain voxel-wise beta weights were retained from the movie stimuli and 3232 blocks from the scenario stimuli. The PLS-DA classifiers trained on beta weights from **movie** blocks achieved a strong performance for decoding the emotional states from the movie stimuli (see Figure 3: A-1). More specifically, the PLS-DA classifier performance metrics from **movie** blocks were as follows: *balanced accuracy* (0.5371, CI = [0.5208, 0.5533]), macro-averaged *sensitivity* (0.5371, CI = [0.5208, 0.5533]), macro-averaged *specificity* (0.967, CI = [0.9657, 0.9682]), macro-averaged *precision* (0.5274, CI = [0.5076, 0.5469]), and macro-averaged *F1-score* (0.5322, CI = [0.5146, 0.5493]). All p values are less than 0.001 (see Table 7 in SI-6: Tables for more information). Emotion-specific performance metrics of PLS-DA classification of **movie** blocks are tabulated in Table 8 in S6--Tables. Macro-averaged AUC (area under the curve) value of the PLS-DA movie classifiers (see Figure 3: B-1) is 0.91, CI= [0.89, 0.94]. In contrast, the PLS-DA classifiers trained on beta weights from the **scenario** blocks performed more poorly (see Figure 3. A-2), though still better than chance (see Figure 3: C-2). The performance metrics from **scenario** blocks were as follows: *balanced accuracy* (0.1862, CI = [0.1733, 0.199]), macro-averaged *sensitivity* (0.1862, CI = [0.1733, 0.199]), macro-averaged *specificity* (0.9418, CI = [0.9409, 0.9427]), macro-averaged *precision* (0.1812, CI = [0.1663, 0.1963]), and macro-averaged *F1-score* (0.1837, CI = [0.1701, 0.1973]). All p values are less than 0.001 (see Table 9 in SI-6: Tables for more information). Emotion-specific performance metrics of PLS-DA classification of **scenario** blocks are tabulated in Table 10 in SI-6: Tables. Macro-averaged AUC value of the PLS-DA scenario classifiers (see Figure 3: B-2) is 0.66, CI= [0.89, 0.94]. When comparing the four types of Poisson regression models described in Methods: Post-hoc classification error analysis and Bayesian model comparison, we found that the Poisson regression model with the categorical rating regressor explained the errors the best for both movies (log Bayes factor=21.89; see Figure 3: E-1) and scenarios (log Bayes factor=20.5241; see Figure 3: E-2). The beta coefficients from the categorical Poisson models are the following: intercept (3.7192, CI= [3.4976, 3.9408], t=16.781, p<0.0001), categorical pairwise distance regressor (-1.6995, CI= [-1.9108, -1.4882], t=-8.0411, p<0.0001) for movies (see Figure 3: E-3). Similarly, we have the following estimates, intercept (4.0192, CI= [3.8180, 4.2204], t= 19.977, p<0.0001), categorical pairwise distance regressor (-1.4124, CI= [-1.6136, -1.2112], t= -7.3913, p<0.0001) for scenarios (see Figure 3: E-4). The fact the estimates of the categorical pairwise distance regressor are negative tells us that an increase in categorical pairwise distance between two emotions leads to lower confusion between those two emotions.

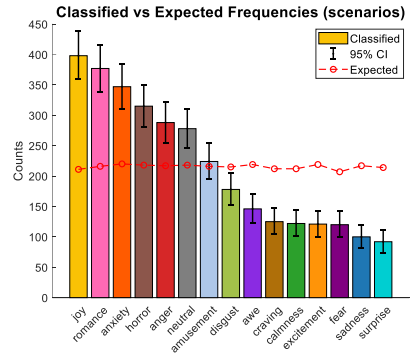
Movie blocks

A-1

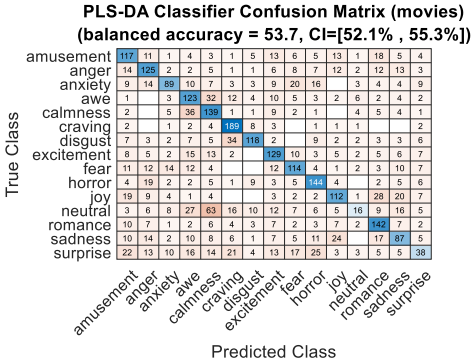


Scenario blocks

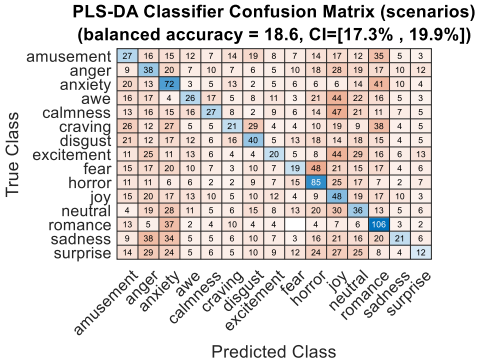
A-2



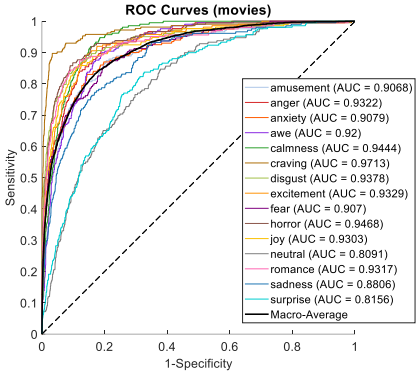
B-1



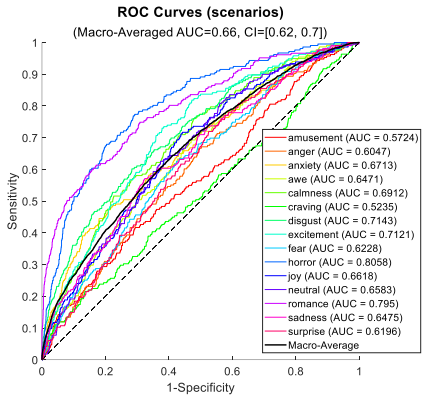
B-2



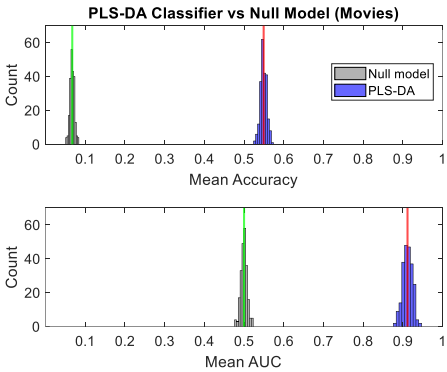
C-1



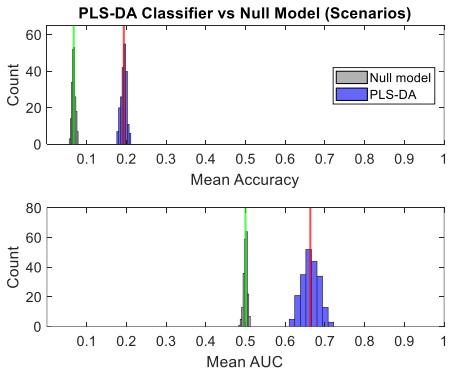
C-2



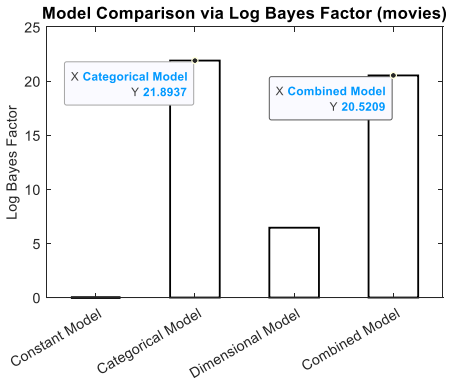
D-1



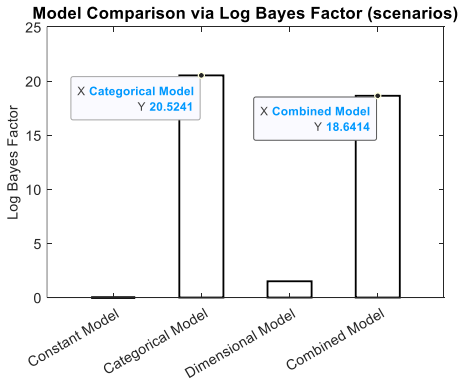
D-2



E-1



E-2



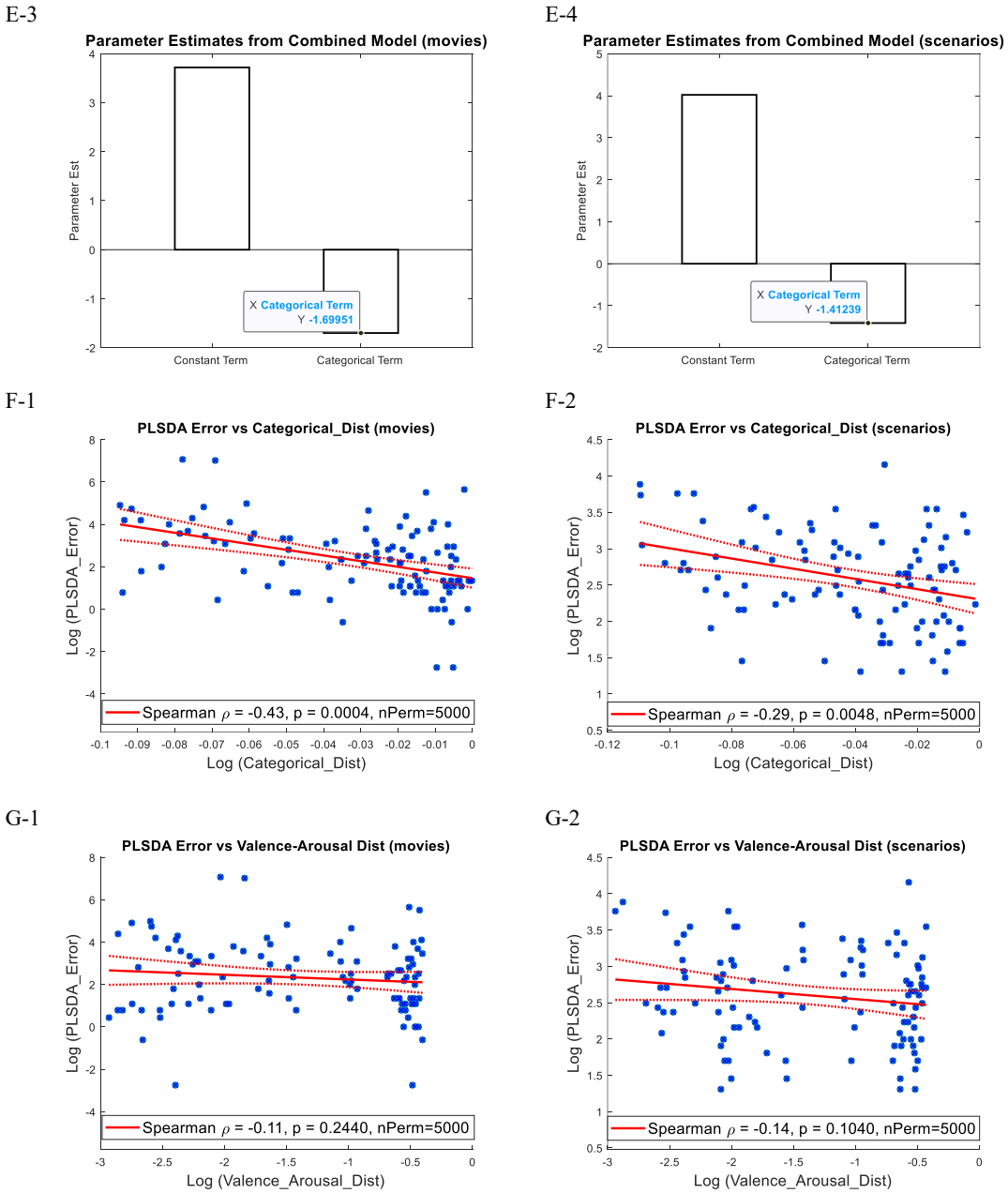
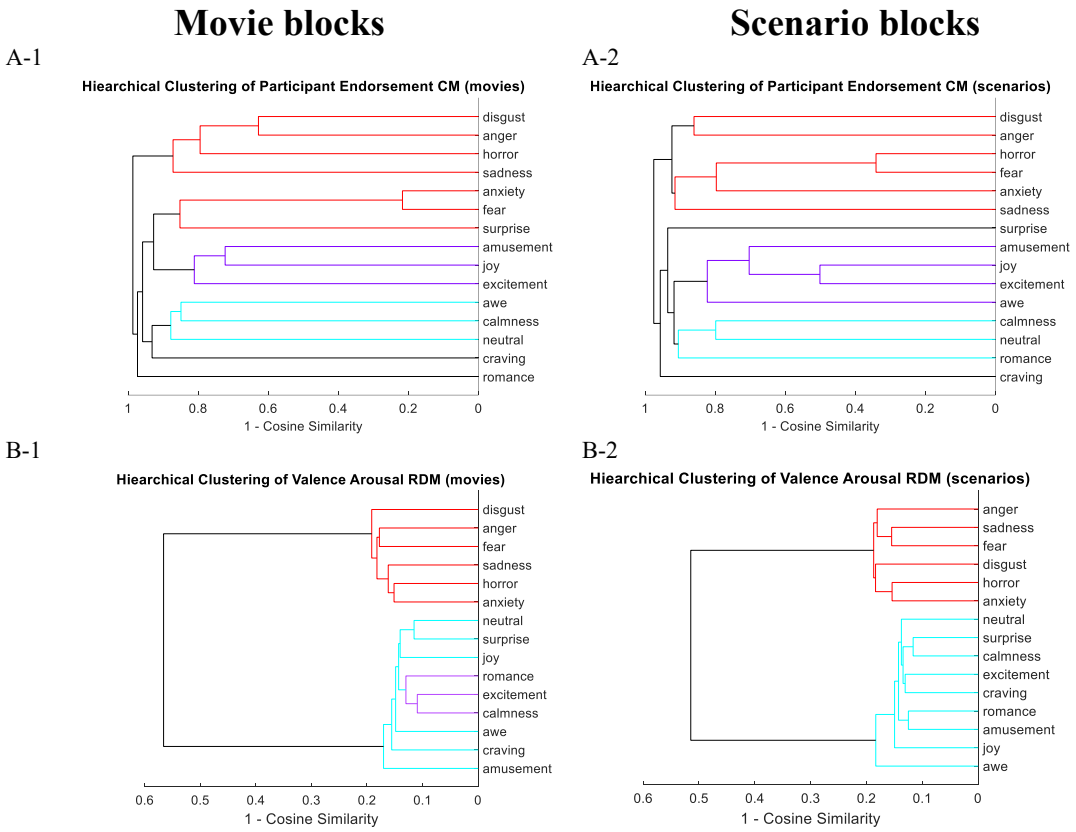


Figure 3. PLS-DA Classification Results of Whole-Brain Patterns of BOLD Activity. **A-1**. Confusion matrix from PLS-DA classification of movie blocks: balanced accuracy (0.5371, CI = [0.5208, 0.5533], p = 0.0002). **A-2**. Confusion matrix from PLS-DA classification of scenario blocks: balanced accuracy (0.1862, CI = [0.1733, 0.199], p = 0.0002). **B-1**. Receiver operating characteristic (ROC) curves of PLS-DA classification of movie blocks: macro-averaged AUC (0.91, CI= [0.89, 0.94]). **B-2**. ROC curves of PLS-DA classification of scenario blocks: macro-averaged AUC (0.66, CI= [0.62, 0.70]). **C-1**. Comparing performance metrics from the PLS-DA classifier and the null model (movie blocks). **C-2**. Comparing performance metrics from the PLS-DA classifier and the null model (scenario blocks). **D-1**. Hierarchical clustering of the PLS-DA confusion matrix in from movie blocks. **D-2**. Hierarchical clustering of the PLS-DA confusion matrix from scenario blocks. **E-1:** Bar plots of the log Bayes factors from four Poisson regression models (movies). **E-2**:

Bar plots of the log Bayes factors from four Poisson regression models (scenarios). **E-3:** Bar plots of the parameter estimates from the categorical Poisson regression model (movies). **E-4**: Bar plots of the parameter estimates from the categorical Poisson regression model (scenarios). **F-1**. Correlation (Spearman rho = -0.4308, p = 0.0004) between PLS-DA classification error and pairwise categorical distance (movie blocks). **F-2**. Correlation (Spearman rho = -0.2926, p = 0.0048) between PLS-DA classification error and pairwise categorical distance (scenario blocks). **G-1**. Correlation (Spearman rho = -0.108, p=0.2412) between PLS-DA classification error and pairwise dimensional distance (movie blocks). **G-2**. Correlation (Spearman rho = -0.1432, p=0.104) between PLS-DA classification error and pairwise dimensional distance (scenario blocks).

Hierarchical clustering of representational dissimilarity matrices

Here we present the results from the hierarchical clustering analysis of various representational dissimilarity matrices from the previous section. Several interesting and meaningful clusters of emotions emerged from hierarchical clustering of participant's subjective categorical endorsement of the movie (see Figure 4: A-1) and scenario (see Figure 4: A-2) inductions. For example, for both movies and scenarios, we see a disgust-anger cluster for negative emotions and an amusement-joy-excitement cluster for positive emotions. In contrast, the emotion clusters from participants' valence/arousal ratings of the emotional stimuli (see Figure 4: B-1 & B-2) are more dichotomous. Additionally, we observe that the clusters from the movie and scenario subjective endorsement data are similar. However, this correspondence between movie and scenario stimuli does not hold for the clusters computed from participants' brain responses to the emotional stimuli, for which the clusters from the movie stimuli (see Figure 4: C-1) are quite different from the clusters from the scenario stimuli (see Figure 4: C-2). Finally, clusters computed from the PLS-DA confusion matrix from movie stimuli (see Figure 4: D-1) are also quite different from those computed from the PLS-DA confusion matrix from the scenario stimuli (see Figure 4: D-2).



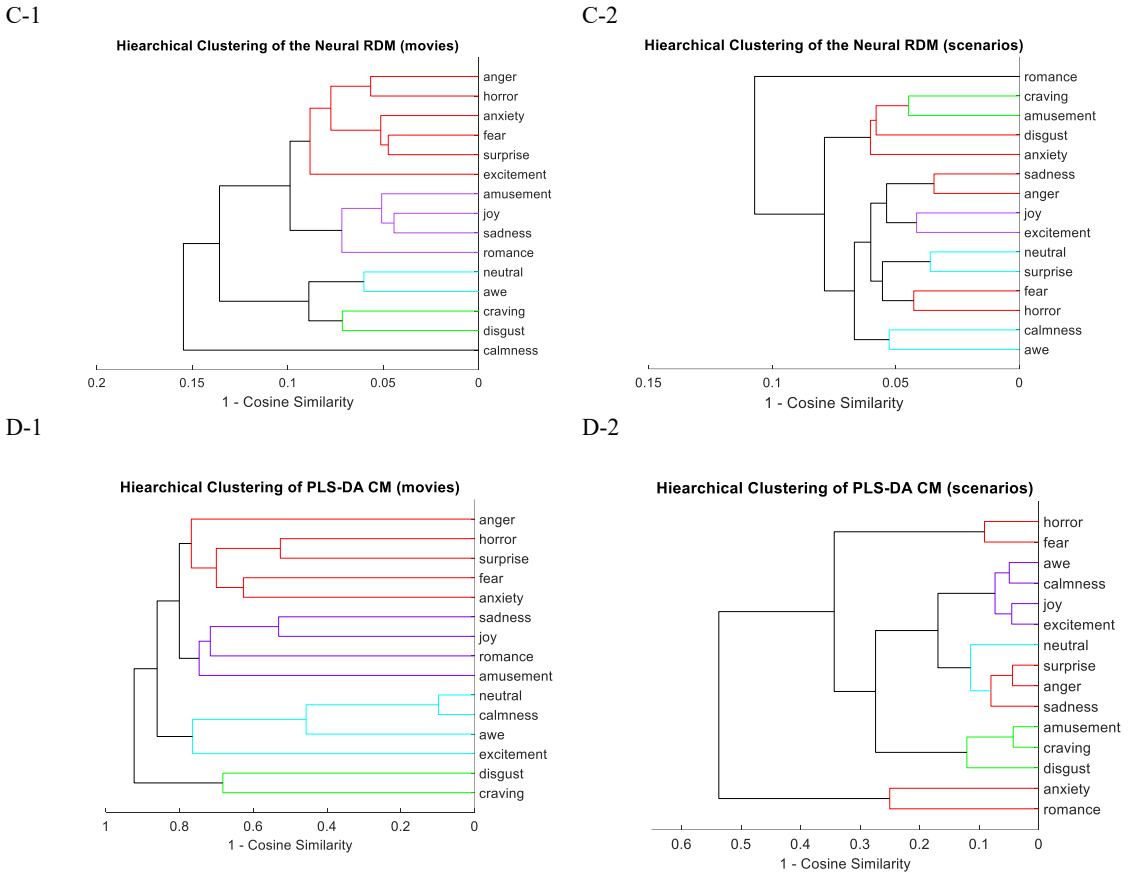
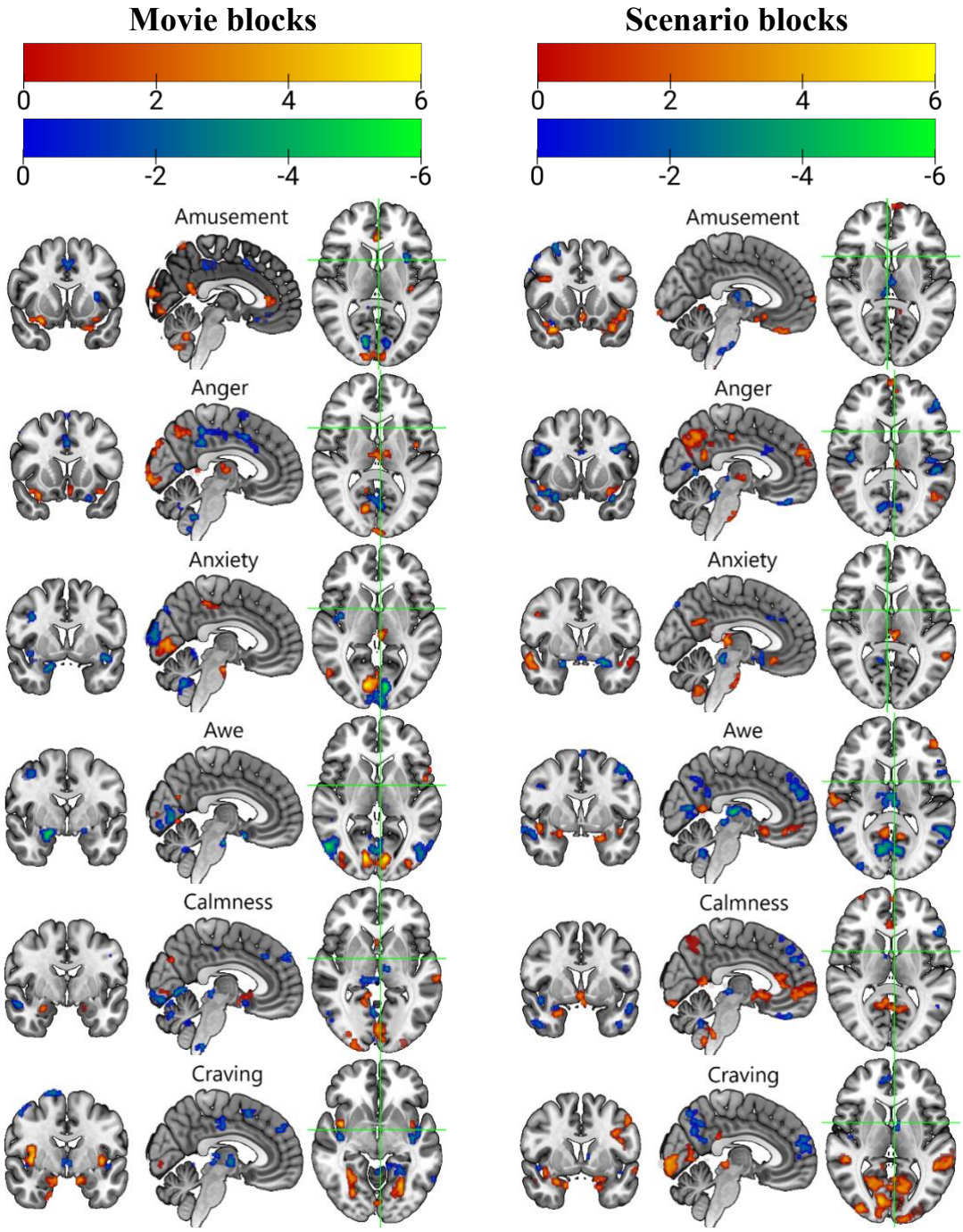


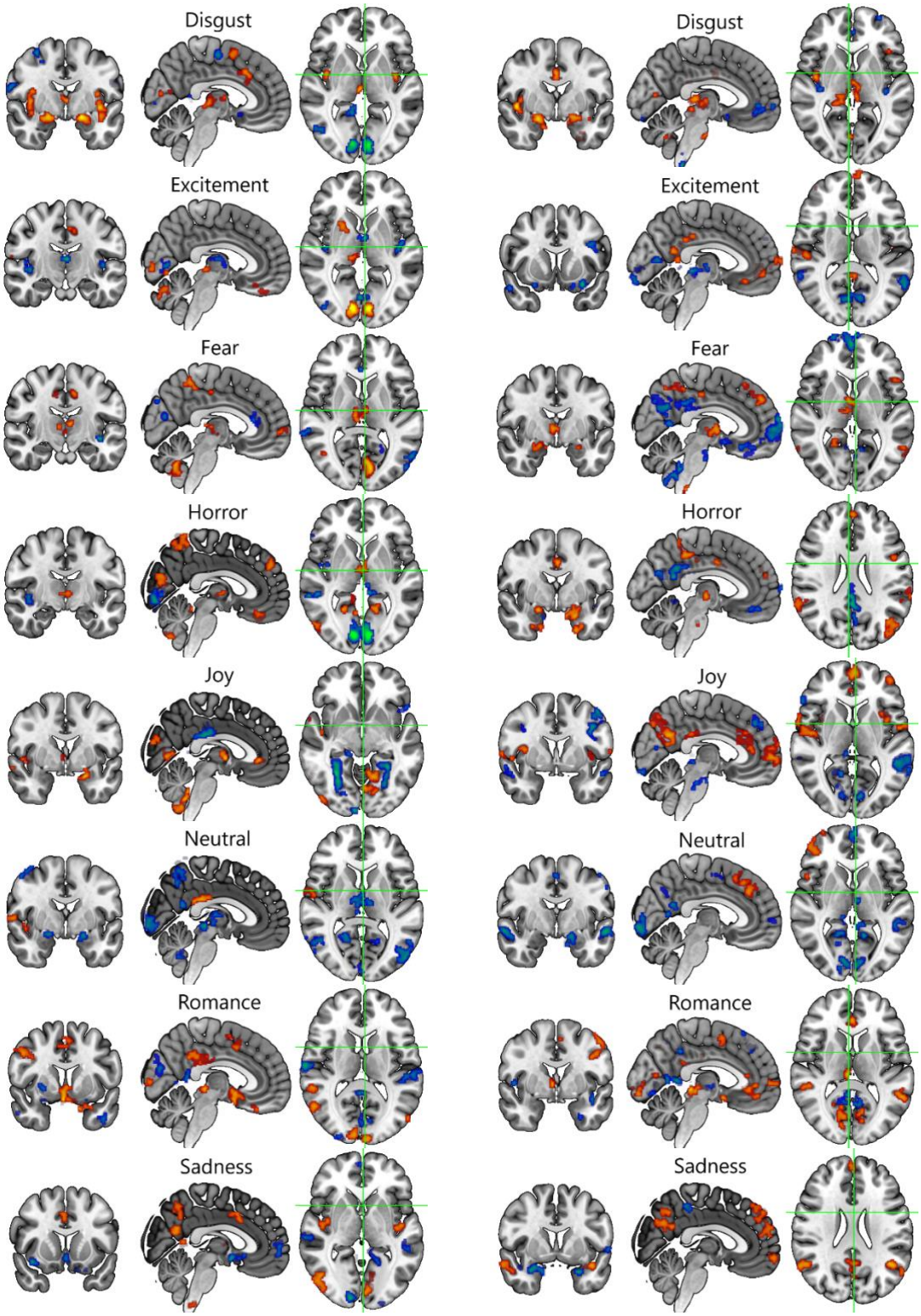
Figure 4. Dendrograms Showing the Emotion Clusters from Participants' Categorical Endorsement, Valence/Arousal Ratings and Brain Responses to the Emotion Inductions, and the PLS-DA Classification of the Stimuli. **A-1**. Hierarchical clustering of the categorical Representation Dissimilarity Matrix (RDM) (movie blocks) in Figure 2: C-1. **A-2**. Hierarchical clustering of the Categorical RDM (scenario blocks) in Figure 2: C-2. **B-1**. Hierarchical clustering of the Valence/Arousal RDM (movie blocks) in Figure 2: E-1. **B-2**. Hierarchical clustering of the Valence/Arousal RDM (scenario blocks) in Figure 2: E-2. **C-1**. Hierarchical clustering of the Neural RDM (movie blocks) in Figure 2: A-1. **C-2**. Hierarchical clustering of the Neural RDM (scenario blocks) in Figure 2: A-2. **D-1**. Hierarchical clustering of the confusion matrix from PLS-DA classification of movie blocks in Figure 3: B-1. **D-2**. Hierarchical clustering of the confusion matrix from PLS-DA classification of scenario blocks in Figure 3: B-2.

Maps of PLS-DA coefficients

When inspecting the voxel maps of the informative PLS-DA coefficients, the first observation that we can make is that brain regions that are important for decoding each emotion are widely distributed across cortical, limbic, and subcortical regions. Notably, the PLS-DA classifiers identified several subcortical regions, including in the brainstem, as well as in the cerebellum, which were missing in prior studies. Our preliminary analysis also showed that the classification accuracy from whole-brain patterns of activity exceeds the classification accuracies obtained with activity from each of the seven canonical resting-state networks (see Figure 1: A-1 & A-2 in Supplemental Materials—S5). This distributed pattern of PLS-DA coefficients suggests that our emotional experience is an integrative process, where neural information from the sensory, limbic, and cortical regions all contribute to classifier's ability to differentiate one emotion from another. Additionally, we see that visual and other sensory areas are more involved for decoding the movie stimuli, suggesting that the representation could be more perceptually driven. In contrast, default

mode and other associative regions play more important roles in decoding the hypothetical scenario stimuli. Therefore, it is possible that the neural representation of emotion elicited from text scenarios is more abstract and internally generated. These differences suggest that neural representation of emotion could be partly modality-dependent, especially during the initial stage of emotional processing. Taken together, the results from our study support the idea that neural representation of emotional states is distributed, high dimensional, and that regions (or networks) engaged may depend in part on the induction modality and the perceptual richness of the stimuli.




Disgust
Excitement
Fear
Horror
Joy
Neutral
Romance
Sadness
Disgust
Excitement
Fear
Horror
Joy
Neutral
Romance
Sadness

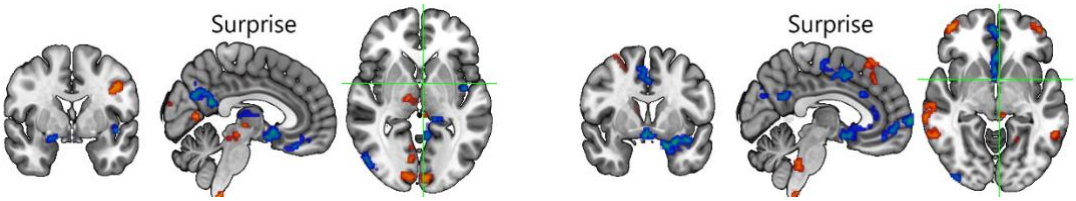


Figure 5. Maps of Informative Coefficients from PLS-DA Classification of Movie and Scenario Inductions

Discussion

There is a relationship between self-report and brain responses to emotion induction

The results from the RSA analyses suggest that there is a relationship between participants' self-reported experiences in response to the emotion inductions and their brain responses. However, this relationship was only significant for the movie stimuli (see Figure 2: B-1), but not for the scenario stimuli (see Figure 2: B-2). Multivariate pattern classification of the emotional content from these stimuli further corroborates this finding. The PLS-DA classifiers trained on movie stimuli achieved a strong performance (see Figure 3: B-1, C-1 & D-1), whereas the classifiers trained on scenario stimuli achieved a weak performance (see Figure 3: B-2, C-2 & D-2), though still above-chance-level accuracy. These results suggest that contextually richer movie stimuli may elicit more robust and consistent patterns of brain activity for each emotion, compared to the hypothetical text scenarios. If we assume the scenario stimuli successfully elicited all of the fifteen target emotions in our participants (see Fig. C-2 and E-2), then there is a possibility that the neural representation of emotions might be modality-dependent and less distinct with text-based scenarios. We note that these scenarios required the participants to be more motivated to imagine themselves in the textual statements and to self-generate the resulting emotional state, which may have been tedious given that the task took 1.5 hours in the scanner. Inspection of the canonical resting-state network analyses supports the idea that the movies relied relatively more on visual networks whereas the scenarios relied relatively more on the default-mode network.

Neural pattern classification performance tracked predictions of categorical but not dimensional theories

A major theoretical question about emotion representation in the brain is whether emotions can be better predicted using categorical or dimensional factors or a hybrid combination of them (Kragel et al., 2024). The Bayesian model comparison directly compared evidence for these models in predicting the classifier performance on the neural data. For both movies and scenarios, the analysis found strong evidence for the categorical model (or the hybrid model) but relatively weaker evidence for the dimensional model. Although the hybrid model performed nearly as well as the categorical model alone, the parameter estimate for the dimensional model was close to zero, meaning that performance was largely driven by the categorical factor. Furthermore, our PLS-DA decoding error analyses on the movies indicated that the pairwise neural distance between emotions is positively correlated with their pairwise distance in categorical space (see Figure 2: B-1), which implies that if the (mean) patterns of BOLD activity associated with two emotions are more distinct (as measured by the metric of 1-cosine similarity), then these two emotions tend to be categorized in different ways. However, pairwise neural distance was neither positively nor negatively correlated with pairwise distance in the valence/arousal space (see Figure 2: F-1). These findings provide insight into the putative mechanisms that the pattern classification algorithm was relying

on to make its predictions, with evidence favoring a categorical architecture. Because our supplemental analyses (see Supplemental Figure 5B-1) showed that maximal PLS-DA classifier performance required a minimum of a 7-dimensional latent space to optimally decode the movies, it is likely that a circumplex model of arousal/valence is simply too low-dimensional. The poor decoding performance of the scenarios precludes strong conclusions from similar error analyses; yet, we note that the Bayesian model comparison still favored the categorical model in this case, despite the fact that a much lower-dimensional space was needed for optimal decoding performance (Supplemental Figure 5B-2).

Emotions are organized into intuitive and meaningful clusters at both the subjective and neural levels of representation

The results from hierarchical clustering analysis provide complementary evidence to suggest that emotions are organized into both larger-scale and finer-scale clusters that are meaningful, intuitive, yet nuanced, at both the subjective and neural levels of representation. First, we see roughly four clusters based on participants' categorical endorsement of the emotional stimuli. These clusters are roughly the same for movie (see Figure 4: A-1) and scenario stimuli (see Figure 4: A-2), for example the disgust–anger-horror cluster for negative emotions and the amusement–joy–excitement cluster for positive emotions. In terms of participants' valence and arousal ratings of these stimuli, we see two large clusters that are quite distant from each other (see Figure 4: B-1 & B-2), one for positively valenced emotions and one for negatively valenced emotions. Within each cluster, there are, however, smaller clusters. Comparing these clusters from the categorical endorsement and valence/arousal ratings, we see that the clusters from categorical endorsement are more equidistant from each other and fine-grained, whereas the clusters from valence/arousal ratings are more dichotomous. This difference suggests that participants are able to make finer distinctions and categorizations of their emotional experiences when presented with emotion labels that may help them tap into the emotion concepts that they have already acquired. Other studies in the literature have shown the effectiveness of developing finer emotional granularity through learning a more diverse and fine-grained sets of emotion concepts (Winkielman et al., 2023). In contrast, participants seem less able to differentiate their emotional experiences when asked to rate the valence/arousal of the emotional stimuli that they encountered in this study. It also seems that participants tend to rely more on the valence property of the stimuli to make such coarse-grain evaluation, i.e., an emotion is either good or bad, as evidenced by the two large-scale clusters from participants' valence/arousal ratings that are split along the valence dimension (see Figure 4: B-1 & B-2). This bias is consistent with the finding in the literature that people find it harder to report level of arousal than valence when it comes to their affective experiences, perhaps in part due to the multiple varieties of arousal experiences (Satpute et al., 2019). When examining the clusters obtained from participants' brain activity, we see that they resemble more closely the clusters from participants' categorical endorsement of the emotional stimuli, i.e., the clusters are also less dichotomous and more fined-grained. This suggests that the brain representation of emotional experiences may be more categorically distinct. However, this is only the case for movie stimuli (see Figure 4: C-1). For scenario stimuli, the clusters from participants' brain activity are more intermixed and less intuitive (see Figure 4: C-2). Not surprisingly, the clusters derived from the confusion matrix of the PLS-DA classifier (see Figure 4: D-1) resemble closely the clusters computed from participants' brain activity. Consequently, these clusters also resemble those computed from participants' categorical endorsement of the movie stimuli. These results from hierarchical clustering analysis corroborate and provide additional details to the main finding from the RSA analysis, which is that there is a robust relationship between participants' subjective experience (self-report) of and their brain responses to the emotional stimuli in the current study. Taken together, these data-driven and computational modeling based findings from our study

highlight the categorical, distributed, conceptually distinctive, and high-dimensional nature of our emotional experiences, consistent with parallel observations in support of Semantic Space Theory (Cowen & Keltner, 2021).

Limitations and future directions

The readers should keep in mind several limitations when interpreting the results from the current study. First, the results from the current study may not generalize to clinical populations, as emotions could be organized differently both at the subjective and/or neural level for participants with certain psychiatric conditions such as mood and anxiety disorders. Second, the valence/arousal ratings were obtained outside of the scanner. This decision was for practical considerations due to the length of the study in the scanner (1.5 hours). We note that the arousal/valence ratings obtained did correlate with those in the normative database for the stimuli, and that our prior study on a subset of 7 emotions did obtain these ratings in the scanner (Kragel & LaBar, 2015) and yielded similar findings as we did here. Because our supplemental analysis showed that roughly seven dimensions are needed for the PLS-DA classifiers to achieve maximum classification accuracies for decoding movies, future studies could investigate whether a more complex dimensional model may account better for neural classification performance. Finally, future studies should investigate how emotions are organized in the body and examine the representational similarity among self-report, brain response, and physiological responses to emotion induction. Overall, the results from the current study advance our understanding of how a relatively large set of emotions are organized and represented in the brain, with evidence that there is a relationship between our subjective experience of and brain responses to emotional inductions that highlight the richness and complexity of human affective experiences.

Acknowledgments

The current study was funded by the U.S. National Institutes of Health grant RO1 MH124112-01. L. F. was supported by an NSF Graduate Fellowship. The authors also wish to thank Mary Baumann, Shannan Chen, and Sebastian Sanchez for their contribution to participant recruitment and data collection for the current project.

Supplemental Materials

S1- Movie and Scenario Stimuli

Movie stimuli (10 per emotion category) were selected based on normative ratings acquired from (Cowen & Keltner, 2017), as well as additional ratings we collected for new movie clips that were not in the Cowen and Keltner (2017) dataset. New movie clips were especially needed to fill categories such as anger, calmness, excitement, and neutral, which were underrepresented in Cowen and Keltner (2017). Following similar procedures as Cowen and Keltner (2017) and those detailed in Faul et al. (2023), we presented movie clips to participants recruited from Amazon's Mechanical Turk (MTurk) and asked them to select among 34 categories the emotion(s) that best described how they felt when watching each movie (we replaced the boredom category label with neutral). Movie clips were rated by at least 9 individuals in Cowen and Keltner (2017) and at least 11 individuals in our own normative data. We then determined the agreement among participants when choosing category labels (i.e., the concordance rate). Movie clips were only selected for the current study if the concordance rate for the category of interest was greater than 0.5 (reflecting a true majority among raters), and we also required this concordance rate to be greater than any of the other emotion categories in the study. All movie clips lasted between 3-12 seconds (M = 6.8 s, SD = 2.3 s). Although a one-way ANOVA test indicated a significant difference in average movie length across the 15 emotion categories ($F_{14,135} = 1.99$, $p = 0.023$), none of the follow-up pairwise comparisons with a Tukey HSD test were significant (all ps > 0.05). Follow-up tests only indicated a marginally shorter duration for the romance category compared to the horror ($p = 0.062$), excitement ($p = 0.075$), and surprise ($p = 0.089$) categories. Thus, movie clips were generally comparable in duration across the 15 emotion categories. The average normative concordance rate for the movie stimuli was 0.73 (SD = 0.13). The category with the highest mean concordance rate was disgust (0.97), and the lowest categories were anxiety and excitement (0.62). We observed a significant difference in concordance rates between categories ($F_{14,135} = 16.89$, $p < .001$). Follow-up pairwise comparisons with a Tukey HSD test revealed that the mean concordance rate for disgust was significantly higher than all other categories except amusement (all ps < .05). The mean concordance rates for amusement, sadness, craving, fear, and calmness were also significantly higher when compared to some of the other categories, whereas anxiety, excitement, and awe had comparatively lower concordance rates.

Text scenarios (10 per emotion category) were selected based on data from Faul et al. (2023), who normed custom scenarios as well as scenarios originally created by Fields and Kuperberg (2012). Comprehensive details on these scenarios and their normative data are further detailed in Faul et al. (2023). In brief, all text scenarios were rated by at least 11 MTurk participants who selected emotion categories that best described how they felt when imagining each scenario. As with the movie stimuli, we selected a text scenario for the current study only if the concordance rate was greater than 0.5 for the intended category, and we also ensured that this rate was greater compared to all other categories of interest. All scenarios were written in the second-person perspective and consisted of two sentences (mean words = 19.31, SD = 4.52). With a one-way ANOVA test, we confirmed that the 15 emotion categories did not differ in the number of words per scenario ($F_{14,135} = 1.53$, $p = .108$), average text concreteness ($F_{14,135} = 1.25$, $p = .245$), and average text imageability ($F_{14,135} = 0.79$, $p = .682$). Concreteness and imageability values were obtained with TAALES version 2.0 (Kyle et al., 2018), using metrics from Brysbaert et al. (2014) and the MRC Psycholinguistic Database (Coltheart, 1981), respectively. The average normative concordance rate for the text scenarios was 0.75 (SD = 0.09). The category with the highest mean concordance rate was fear (0.90), and the lowest was neutral (0.66). As with the movie stimuli, we observed a significant difference in concordance rates between categories ($F_{14,135} = 8.45$, $p < .001$). Follow-up comparisons with a Tukey HSD test revealed that the mean concordance rate for fear was

significantly higher than all other categories except disgust (all *p*s < .05). The mean concordance rate for disgust was also significantly higher than most other categories, whereas neutral and awe had comparatively lower concordance rates compared to some of the other categories.

S2-Task Design

The task was completed over the course of two MRI sessions, separated by at least seven days. The two imaging sessions were identical except for the type of stimulus used: on one day, participants were presented with movie stimuli and on the other day they were presented with text scenarios (order of stimulus type was balanced across participants). Each imaging session contained eight functional task runs. A total of 30 task blocks were presented over the eight runs: runs 1, 2, 3, 5, 6, and 7 contained 4 blocks; runs 4 and 8 contained 3 blocks. Every run began and ended with an 8-second presentation of a fixation cross (small black cross presented in the middle of the screen). Each task block consisted of 5 induction trials followed by emotion endorsement questions and passive viewing of a grayscale washout picture. Each trial contained three parts: passively viewing a movie or reading a text scenario, mentally replaying the movie or scenario, and responding to a question about where the events of the movie or scenario took place. Each movie was presented without sound for its full duration (see below for average movie lengths by emotion) and was immediately followed by the instruction "Replay the movie in your mind". The "replay" instruction time for each movie was set such that the total movie and "replay" time was always 13 seconds. Each text scenario was displayed for an amount of time equal to the average movie length of its associated emotional induction category (e.g. the average length of anxiety-inducing movie clips was eight seconds, so all anxiety-inducing scenarios were displayed for eight seconds). Text scenario display was immediately followed by the instruction "Replay the scenario in your mind". As with the movie stimuli, the duration of the "replay" instruction time following each text scenario was set to make the total scenario and "replay" time combine to 13 seconds. Following each "replay" instruction period, a fixation cross was presented for 2 seconds (average, jittered between 1 and 3 seconds). Participants were then presented with the question "Where did the [movie/scenario] take place?" and the possible responses: "1: Indoors", "2:Outdoors", "3:Both", "4:Uncertain". Participants selected one of these options by pressing the appropriate button on the button box held in their right hand. Participants had 4 seconds to select a response. This question was included to verify participant attention to movie and scenario stimuli while avoiding affective characteristics of the stimuli. After the final trial of each block, participants were asked to report the emotion they experienced the most during that block. A selection screen was presented with a list of the 15 emotion options (amusement, anger, anxiety, awe, calmness, craving, disgust, excitement, fear, horror, joy, neutral, romance, sadness, surprise) in a column down the center of the screen and the question "Which of these did you experience most during the last group of movies?". The order of emotions on the screen was randomized for each presentation. Participants selected an emotion using a selection box that began on a small plus symbol set between the 7th and 8th emotions in the list. Participants used the index finger on the button box to move the selection cursor up the list and the middle finger button to more the selection cursor down the list. Once moved away from the plus symbol in the middle of the list, the cursor would never go back to that position, instead jumping between the 7th and 8th emotions. Participants had 11 seconds to leave the selection cursor on the emotion they wished to select. Following the emotion endorsement screen, participants were asked to rate to what degree they experienced the emotion they selected. The message "You selected [emotion]. How intensely did you experience it?" ([emotion] replaced with the actual emotion chosen) was displayed on the left-hand portion of the screen and the integers 1 through 10 were displayed in a column down the center of the screen in descending order. A plus symbol was displayed below the number 1 and was again the starting position of a selection cursor.

Participants used the same button controls to move the cursor up or down to highlight a response. They had 7 seconds to leave the selection cursor on the integer they wished to select. Following emotion intensity endorsement, one of three grayscale "washout" pictures was randomly chosen and displayed for 10 seconds (average, jittered between 8 and 12 seconds). Three neutral pictures (of flowers and trains) were first converted to grayscale and then phase-scrambled. This procedure created pictures which retained the low-level image characteristics of the originals, but contained no discernable objects.

S3-MRI Scanner

Magnetic resonance imaging was performed on a Siemens Magnetom 3.0T scanner. The functional task was presented on a MRI-compatible 4K monitor (InroomViewingDevice, NordicNeurolabs, Bergen, Norway) placed at the rear of the scanner. Participants viewed the monitor with a mirror attached to the head coil. Participant responses were collected via a 4-button response box held in the right hand (Current Designs, Inc., Philadelphia, PA).

S4 -MRI Preprocessing

Results included in this manuscript come from preprocessing performed using *fMRIPrep* 23.0.2 (Esteban et al., 2018; Esteban et al., 2019), which is based on *Nipype* 1.8.6 (Gorgolewski et al., 2011; Gorgolewski et al., 2018).

"The following boilerplate text was automatically generated by fMRIPrep with the express intention that users should directly copy it into their manuscripts unchanged. It is released under the CC0 license."

Preprocessing of B_0 inhomogeneity mappings

A total of 9 fieldmaps were found available within the input BIDS structure for this particular subject. A B_0-nonuniformity map (or *fieldmap*) was estimated based on two (or more) echo-planar imaging (EPI) references with topup in FSL 6.0.5.1 (Andersson et al., 2003).

Anatomical data preprocessing

A total of 1 T1-weighted (T1w) images were found within the input BIDS dataset.The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al., 2010), distributed with ANTs 2.3.3 (Avants et al., 2008), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a *Nipype* implementation of the antsBrainExtraction.sh workflow (from ANTs), using MNI152NLin6Asym as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast FSL 6.0.5.1 (Zhang et al., 2001). Brain surfaces were reconstructed using recon-all in FreeSurfer 7.3.2 (Dale et al., 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (Klein et al., 2017). Volume-based spatial normalization to two standard spaces (MNI152NLin6Asym, MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the T1w template. The following templates were were selected for spatial normalization and accessed with *TemplateFlow* (23.0.0, (Ciric et al., 2022)): *FSL's MNI ICBM 152 non-linear 6th Generation Asymmetric Average Brain Stereotaxic Registration Model (Evans et al., 2012)*; TemplateFlow ID:

MNI152NLin6Asym, *ICBM 152 Nonlinear Asymmetrical template version 2009c* (Fonov et al., 2009); TemplateFlow ID: MNI152NLin2009cAsym.

Functional data preprocessing

For each of the 9 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 6.0.5.1:57b01774, (Jenkinson et al., 2002)). The estimated *fieldmap* was then aligned with rigid-registration to the target EPI (echo-planar imaging) reference run. The field coefficients were mapped on to the reference EPI using the transform. BOLD runs were slice-time corrected to 0.946s (0.5 of slice acquisition range 0s-1.89s) using 3dTshift from AFNI ((Cox & Hyde, 1997),RRID:SCR_005927). The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration (Greve & Fischl, 2009). Co-registration was configured with six degrees of freedom. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), DVARS and three region-wise global signals. FD was computed using two formulations following (Power et al., 2014) (absolute sum of relative motions, [Power2014]) and Jenkinson et al. (2002) (relative root mean square displacement between affines. FD and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by (Power et al., 2014)). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (*CompCor*, (Behzadi et al., 2007)). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for the two *CompCor* variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 2% variable voxels within the brain mask. For aCompCor, three probabilistic masks (CSF, WM and combined CSF+WM) are generated in anatomical space. The implementation differs from that of Behzadi et al. in that instead of eroding the masks by 2 pixels on BOLD space, a mask of pixels that likely contain a volume fraction of GM is subtracted from the aCompCor masks. This mask is obtained by dilating a GM mask extracted from the FreeSurfer's *aseg* segmentation, and it ensures components are not extracted from voxels containing a minimal fraction of GM. Finally, these masks are resampled into BOLD space and binarized by thresholding at 0.99 (as in the original implementation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the *k* components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al., 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers. Additional nuisance timeseries are calculated by means of principal components analysis of the signal found within a thin band (*crown*) of voxels around the edge of the brain, as proposed by Patriat et al. (2017). The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD run in MNI152NLin6Asym space*. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Automatic removal of motion artifacts using independent component analysis (ICA-AROMA,(Pruim et al., 2015)) was

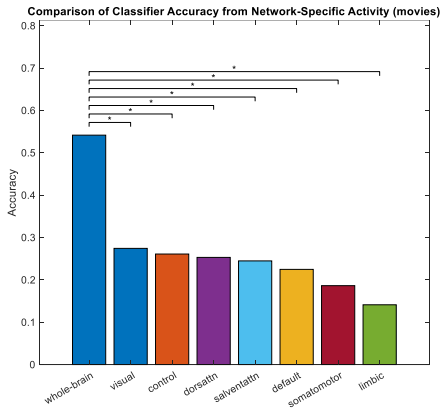
performed on the *preprocessed BOLD on MNI space* time-series after removal of non-steady state volumes and spatial smoothing with an isotropic, Gaussian kernel of 6mm FWHM (full-width half-maximum). Corresponding "non-aggresively" denoised runs were produced after such smoothing. Additionally, the "aggressive" noise-regressors were collected and placed in the corresponding confounds file. All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

Study-specific canonical network mask generation

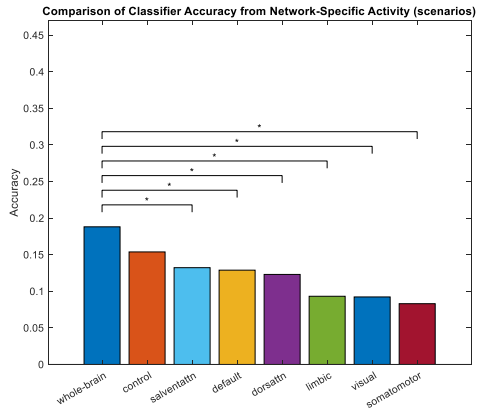
Masks of seven canonical brain networks (Thomas Yeo et al., 2011), control, default, dorsal attention, limbic, Salience/Ventral Attention, Somatomotor, and Visual, were created from ICA decompositions of the functional resting-state images collected as part of this study. Two sets of network masks were created, one from resting state data collected during movie sessions and one from resting state data collected during scenario sessions. All resting state data were first put through quality assurance. An image was excluded if any visual artifacts were present, more than 10% of the image time-points had FD > 0.5mm, or the maximum volume-to-volume motion exceeded 3.0mm (Birn, 2023). This lead to 30 out of 114 (26%) of movie session resting-state images being excluded and 38 out of 121 (31%) of scenario session resting-state images being excluded. These exclusion percentages are somewhat high. We suspect this is due to the resting-state scan occurring at the end of a relatively long (90-min) imaging session, which may have made it difficult for participants to remain still during acquisition. However, there were still 84 good resting state images from the movie sessions and 83 good resting state images from the scenario sessions. Only participants who contributed task-based data to the classification analysis were included in this resting-state analysis. Following QA, the remaining data sets from the movie sessions were put through FSL's MELODIC group ICA (Beckmann et al., 2005) using temporal concatenation with automatic dimensionality estimation. The threshold for IC maps was set to 0.5. This resulted in a group ICA decomposition comprised of 71 components. The network correspondence toolbox (Kong et al., 2025) was then used to calculate the correspondence of each component with each of the seven canonical networks of interest. For each network, the component with the lowest p-value (or several components if they shared the lowest) was visually inspected against the network mask. For six of the networks (default, dorsal attention, limbic, salience/ventral attention, somatomotor, and visual), a single component was found which well-matched the spatial distribution of the network. The thresholded z-stat images produced by MELODIC for these components were binarized to serve as the participant-derived canonical network masks. For the control network, one component was found which contained the right-lateralized portions of the network and another was found which contained the left-lateralized portions of the network. A mask was created from the union of these two component masks as the participant-derived control network mask. The same process was carried out using the resting state data sets from the scenario sessions. This resulted in an ICA decomposition comprised of 88 components. Five of the canonical networks (default, dorsal attention, salience/ventral attention, somatomotor, and visual) were well represented by single components. Masks for the limbic and control network were each created by taking the union of two component masks.

S5 - Figures from Additional Analyses

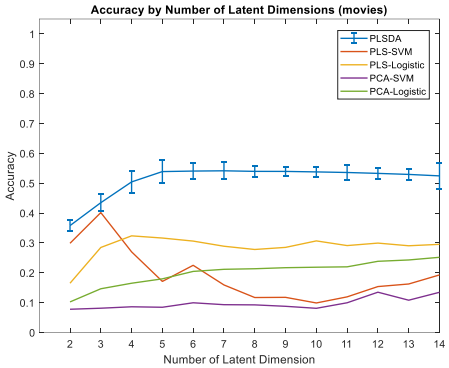
A-1



A-2



B-1



B-2

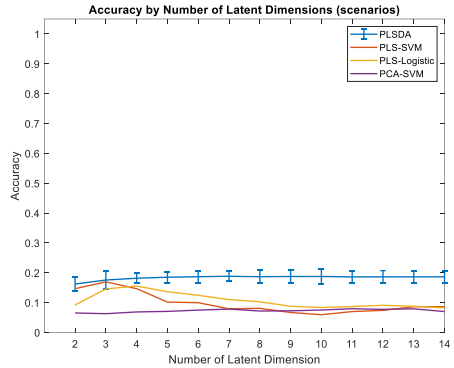


Figure 1: Figures from additional analyses. **A-1**: Comparison of classification accuracies from whole-brain patterns of activity vs seven canonical resting state networks (**movies**). **A-2**: Comparison of classification accuracies from whole-brain patterns of activity vs seven canonical resting state networks (**scenarios**). **B-1**: Classification accuracies as a function of number of latent dimensions for PLS-DA, PLS-SVM, PLS-Logistic, PCA-SVM, and PCA-Logistic (**movies**). **B-2**: B-1: Classification accuracies as a function of number of latent dimensions for PLS-DA, PLS-SVM, PLS-Logistic, PCA-SVM, and PCA-Logistic (**scenarios**).

S6 - Tables

Table 1. Performance Metrics of Participant In-Scanner Emotion Endorsement of **Movie** Blocks

Movies				
Metrics	Estimate	Bootstrap CI	Permutation P value	Chance (Null Distribution)
Balanced Accuracy	0.7395	[0.7255, 0.7534]	**0.0002***	0.0667
Sensitivity (macro-average)	0.7395	[0.7255, 0.7534]	**0.0002***	0.0667
Specificity (macro-average)	0.9814	[0.9803, 0.9824]	**0.0002***	0.9333
Precision (macro-average)	0.7488	[0.7347, 0.7627]	**0.0002***	0.0677
F1-Score (macro-average)	0.7441	[0.7303, 0.758]	**0.0002***	0.0677

Table 2. Emotion Specific Metrics of Participant In-Scanner Endorsement of **Movie** Blocks

Movies				
	Sensitivity	Specificity	Precision	F-1 Score
Amusement	0.7542	0.9693	0.6357	0.6899
Anger	0.4874	0.9922	0.8169	0.6105
Anxiety	0.6034	0.9711	0.5983	0.6008
Awe	0.7089	0.9855	0.7778	0.7417
Calmness	0.9034	0.9825	0.7875	0.8415
Craving	0.8159	0.9982	0.9701	0.8864
Disgust	0.9280	0.9786	0.7552	0.8327
Excitement	0.7029	0.9876	0.8038	0.7500
Fear	0.4979	0.9729	0.5673	0.5303
Horror	0.7615	0.9768	0.7027	0.7309
Joy	0.7265	0.9720	0.6464	0.6841
Neutral	0.8602	0.9675	0.6527	0.7422
Romance	0.8067	0.9985	0.9746	0.8828
Sadness	0.8608	0.9867	0.8226	0.8412
Surprise	0.6723	0.9813	0.7182	0.6945

Table 3. Performance Metrics of Participant In-Scanner Emotion Endorsement of **Scenario** Blocks

Scenarios				
Metrics	Estimate	Bootstrap CI	Permutation P value	Chance (Null Distribution)
Balanced Accuracy	0.7031	[0.6882, 0.7174]	**0.0002***	0.0667
Sensitivity (macro-average)	0.7031	[0.6882, 0.7174]	**0.0002***	0.0667
Specificity (macro-average)	0.9788	[0.9777, 0.9798]	**0.0002***	0.9333
Precision (macro-average)	0.7238	[0.71, 0.7371]	**0.0002***	0.0677
F1-Score (macro-average)	0.7133	[0.6992, 0.7268]	**0.0002***	0.0677

Table 4. Emotion Specific Metrics of Participant In-Scanner Endorsement of **Scenario** Blocks

Scenarios				
	Sensitivity	Specificity	Precision	F-1 Score
Amusement	0.6920	0.9765	0.6758	0.6838
Anger	0.6840	0.9907	0.8382	0.7533
Anxiety	0.6653	0.9796	0.6987	0.6816
Awe	0.6680	0.9893	0.8164	0.7348
Calmness	0.8740	0.9697	0.6748	0.7616
Craving	0.7520	0.9972	0.9502	0.8396
Disgust	0.8379	0.9887	0.8413	0.8396
Excitement	0.5952	0.9697	0.5837	0.5894
Fear	0.5984	0.9711	0.5984	0.5984
Horror	0.5961	0.9765	0.6468	0.6204
Joy	0.6071	0.9465	0.4474	0.5152
Neutral	0.7549	0.9593	0.5701	0.6497
Romance	0.6798	0.9969	0.9399	0.7890
Sadness	0.8016	0.9887	0.8347	0.8178
Surprise	0.7381	0.9813	0.7381	0.7381

Table 5. Representational Similarity Between Self-Report and Brain Responses to **Movie** Blocks

Movies				
RSA (Spearman's rho)	Estimate	Bootstrap CI	Permutation P value	Chance
Neural_vs_Categorical	**0.4113**	[0.2132, 0.5902]	**0.0004**	0
Neural_vs_Dimensional	0.0846	[-0.1193, 0.2859]	0.3004	0
Categorical _vs_Dimensional	**0.5932**	[0.4563, 0.7001]	**0.0004**	0

Table 6. Representational Similarity Between Self-Report and Brain Responses to **Scenario** Blocks

Scenarios				
RSA (Spearman's rho)	Estimate	Bootstrap CI	Permutation P value	Chance
Neural_vs_Categorical	0.1673	[-0.02761, 0.3518]	0.1208	0
Neural_vs_Dimensional	0.1465	[-0.06299, 0.344]	0.0562	0
Categorical _vs_Dimensional	**0.7509**	[0.6713, 0.8057]	**0.0002**	0

Table 7. Performance Metrics of PLSDA Classification of **Movie** Blocks

Movies				
Metrics	*Estimate*	*Bootstrap CI*	*Permutation P value*	*Chance*
Balanced Accuracy	0.5371	[0.5208, 0.5533]	**0.0002***	0.0667
Sensitivity (macro-average)	0.5371	[0.5208, 0.5533]	**0.0002***	0.0667
Specificity (macro-average)	0.967	[0.9657, 0.9682]	**0.0002***	0.9333
Precision (macro-average)	0.5274	[0.5076, 0.5469]	**0.0002***	0.0677
F1-Score (macro-average)	0.5322	[0.5146, 0.5493]	**0.0002***	0.0677

Table 8. Emotion Specific Metrics of PLSDA Classification of **Movie** Blocks

Movies				
	Sensitivity	Specificity	Precision	F-1 Score
Amusement	0.5652	0.9582	0.4895	0.5247
Anger	0.5869	0.9612	0.5252	0.5543
Anxiety	0.4450	0.9805	0.6096	0.5145
Awe	0.5885	0.9500	0.4572	0.5146
Calmness	0.6619	0.9407	0.4455	0.5326
Craving	0.8832	0.9650	0.6495	0.7485
Disgust	0.5813	0.9832	0.7066	0.6378
Excitement	0.6085	0.9626	0.5420	0.5733
Fear	0.5534	0.9695	0.5616	0.5575
Horror	0.6825	0.9671	0.6000	0.6386
Joy	0.5283	0.9729	0.5864	0.5558
Neutral	0.0766	0.9901	0.3556	0.1260
Romance	0.6927	0.9613	0.5569	0.6174
Sadness	0.4203	0.9651	0.4603	0.4394
Surprise	0.1818	0.9774	0.3654	0.2428

Table 9. Performance Metrics of PLSDA Classification of **Scenario** Blocks

Scenarios				
Metrics	*Estimate*	*Bootstrap CI*	*Permutation P value*	*Chance*
Balanced Accuracy	0.1862	[0.1733, 0.199]	**0.0002***	0.0667
Sensitivity (macro-average)	0.1862	[0.1733, 0.199]	**0.0002***	0.0667
Specificity (macro-average)	0.9418	[0.9409, 0.9427]	**0.0002***	0.9333
Precision (macro-average)	0.1812	[0.1663, 0.1963]	**0.0002***	0.0677
F1-Score (macro-average)	0.1837	[0.1701, 0.1973]	**0.0002***	0.0677

Table 10. Emotion Specific Metrics of PLSDA Classification of **Scenario** Blocks

Scenarios				
	Sensitivity	Specificity	Precision	F-1 Score
Amusement	0.1280	0.9348	0.1205	0.1241
Anger	0.1759	0.9171	0.1319	0.1508
Anxiety	0.3273	0.9087	0.2075	0.2540
Awe	0.1193	0.9602	0.1781	0.1429
Calmness	0.1244	0.9685	0.2213	0.1593
Craving	0.0963	0.9655	0.1680	0.1224
Disgust	0.1852	0.9542	0.2247	0.2030
Excitement	0.0930	0.9665	0.1653	0.1190
Fear	0.0868	0.9665	0.1583	0.1121
Horror	0.4009	0.9238	0.2698	0.3226
Joy	0.2264	0.8841	0.1206	0.1574
Neutral	0.1644	0.9197	0.1295	0.1449
Romance	0.5121	0.9104	0.2812	0.3630
Sadness	0.0968	0.9738	0.2100	0.1325
Surprise	0.0561	0.9735	0.1304	0.0784

Table 11. Correlations between PLSDA Errors and Various Distance Measures (**Movies**)

Movies				
Spearman's rho	Estimate	Bootstrap CI	Permutation P value	Chance
PLSDA-Error_vs_Categorical	**-0.4308**	[-0.5964, -0.2449]	**0.0004**	0
PLSDA-Error_vs_ Dimensional	-0.108	[-0.3352, 0.1087]	0.2412	0
PLSDA-Error _vs_Neural	**-0.7424**	[-0.8187, -0.641]	**0.0002**	0

Table 12. Correlations between PLSDA Errors and Various Distance Measures (**Scenarios**)

Scenarios				
Spearman's rho	Estimate	Bootstrap CI	Permutation P value	Chance
PLSDA-Error_vs_Categorical	**-0.2926**	[-0.4709, -0.09984]	**0.0048**	0
PLSDA-Error_vs_ Dimensional	-0.1432	[-0.321, 0.04194]	0.104	0
PLSDA-Error _vs_Neural	**-0.5168**	[-0.6489, -0.3629]	**0.0002**	0

Table 13. Formulas for Multiclass Classification Performance Metrics

true positive (TP)	$TP_i = n_{ii}$	true negative (TN)	$TN_i = \sum_{j \neq i} \sum_{h \neq i} n_{jh}$
false positive (FP)	$FP_i = \sum_{j \neq i} n_{ji}$	false negative (FN)	$FN_i = \sum_{j \neq i} n_{ij}$
accuracy (acc)	$acc_i = \frac{TP_i}{TP_i + FN_i}$	precision (prec)	$prec_i = \frac{TP_i}{TP_i + FP_i}$
sensitivity (se)	$se_i = \frac{TP_i}{TP_i + FN_i}$	specificity (sp)	$sp_i = \frac{TN_i}{TN_i + FP_i}$
Balanced Accuracy	$acc_{Balanced} = \sum_{i=1}^{k} acc_i \Big/ k$	Macro-Averaged Sensitivity	$se_{Macro} = \sum_{i=1}^{k} se_i \Big/ k$
Macro-Averaged Specificity	$sp_{Macro} = \sum_{i=1}^{k} sp_i \Big/ k$	Macro-Averaged Precision	$prec_{Macro} = \sum_{i=1}^{k} prec_i \Big/ k$
Macro-Averaged F1-Score	$F1 - Score_{Macro} = 2 \frac{(precision_{Macro} * se_{Macro})}{(precision_{Macro} + se_{Macro})}$		

References

Andersson, J. L., Skare, S., & Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage*, *20*(2), 870-888.

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, *12*(1), 26-41.

Bagby, R. M., Parker, J. D., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia Scale—I. Item selection and cross-validation of the factor structure. *Journal of psychosomatic research*, *38*(1), 23-32.

Barrett, L. F. (2017a). *How emotions are made: The secret life of the brain*. Pan Macmillan.

Barrett, L. F. (2017b). The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, *12*(1), 1-23.

Beckmann, C. F., DeLuca, M., Devlin, J. T., & Smith, S. M. (2005). Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1457), 1001-1013.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage*, *37*(1), 90-101.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *26*(2), 211-243.

Bryant, F. B., Yarnold, P. R., & Grimm, L. G. (1996). Toward a measurement model of the affect intensity measure: A three-factor structure. *Journal of Research in Personality*, *30*(2), 223-247.

Ciric, R., Thompson, W. H., Lorenz, R., Goncalves, M., MacNicol, E. E., Markiewicz, C. J., Halchenko, Y. O., Ghosh, S. S., Gorgolewski, K. J., & Poldrack, R. A. (2022). TemplateFlow: FAIR-sharing of multi-scale, multi-species brain models. *Nature Methods*, *19*(12), 1568-1571.

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, *114*(38), E7900-E7909.

Cowen, A. S., & Keltner, D. (2021). Semantic space theory: A computational approach to emotion. *Trends in Cognitive Sciences*, *25*(2), 124-136.

Cox, R. W. (2012). AFNI: what a long strange trip it's been. *Neuroimage*, *62*(2), 743-747.

Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, *10*(4-5), 171-178.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage*, *9*(2), 179-194.

Edlow, B. L., Olchanyi, M., Freeman, H. J., Li, J., Maffei, C., Snider, S. B., Zöllei, L., Iglesias, J. E., Augustinack, J., & Bodien, Y. G. (2023). Sustaining wakefulness: brainstem connectivity in human consciousness. *bioRxiv*.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169-200.

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PloS one*, *12*(9), e0184661.

Esteban, O., Blair, R., Markiewicz, C. J., Berleant, S. L., Moodie, C., Ma, F., Isik, A. I., Erramuzpe, A., Kent, M. J. D., & andGoncalves, D. E. (2018). FMRIPrep. *Softw. Pract. Exp*.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., & Snyder, M. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111-116.

Evans, A. C., Janke, A. L., Collins, D. L., & Baillet, S. (2012). Brain templates and atlases. *Neuroimage*, *62*(2), 911-922.

Faul, L., Baumann, M. G., & LaBar, K. S. (2023). The representation of emotional experience from imagined scenarios. *Emotion*, *23*(6), 1670.

Fischl, B. (2012). FreeSurfer. *Neuroimage*, *62*(2), 774-781.

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *Neuroimage*(47), S102.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in neuroinformatics*, *5*, 13.

Gorgolewski, K., Esteban, O., Markiewicz, C., Ziegler, E., Ellis, D., Notter, M., Jarecka, D., Johnson, H., Burns, C., & Manhães-Savio, A. (2018). Nipype. Software. *Zenodo*.

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, *48*(1), 63-72.

Gross, J. J., & John, O. P. (2012). Emotion regulation questionnaire. *Journal of personality and social psychology*.

Harvey, P. D., Greenberg, B. R., & Serper, M. R. (1989). The affective lability scales: development, reliability, and validity. *Journal of clinical psychology*, *45*(5), 786-793.

Horikawa, T., Cowen, A. S., Keltner, D., & Kamitani, Y. (2020). The neural representation of visually evoked emotion is high-dimensional, categorical, and distributed across transmodal brain regions. *Iscience*, *23*(5).

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, *17*(2), 825-841.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *Neuroimage*, *62*(2), 782-790.

Kassam, K. S., Markey, A. R., Cherkassky, V. L., Loewenstein, G., & Just, M. A. (2013). Identifying emotions on the basis of neural activation. *PloS one*, *8*(6), e66032.

Keltner, D., Brooks, J. A., & Cowen, A. (2023). Semantic space theory: Data-driven insights into basic emotions. *Current Directions in Psychological Science*, *32*(3), 242-249.

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., & Chaibub Neto, E. (2017). Mindboggling morphometry of human brains. *PLoS computational biology*, *13*(2), e1005350.

Kong, R., Spreng, R. N., Xue, A., Betzel, R. F., Cohen, J. R., Damoiseaux, J. S., De Brigard, F., Eickhoff, S. B., Fornito, A., & Gratton, C. (2025). A network correspondence toolbox for quantitative evaluation of novel neuroimaging results. *Nature communications*, *16*(1), 2930.

Kragel, P. A., Koban, L., Barrett, L. F., & Wager, T. D. (2018). Representation, pattern information, and brain signatures: from neurons to neuroimaging. *Neuron*, *99*(2), 257-273.

Kragel, P. A., & LaBar, K. S. (2014). Advancing emotion theory with multivariate pattern classification. *Emotion Review*, *6*(2), 160-174.
Kragel, P. A., & LaBar, K. S. (2015). Multivariate neural biomarkers of emotional states are categorically distinct. *Social cognitive and affective neuroscience*, *10*(11), 1437-1448.
Kragel, P. A., Sander, D., & LaBar, K. S. (2024). Can Brain Data Be Used to Arbitrate Among Emotion Theories? In *Emotion theory: The Routledge comprehensive guide* (pp. 511-542). Routledge.
Lanczos, C. (1964). Evaluation of noisy data. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, *1*(1), 76-85.
Lewis, M. J., Spiliopoulou, A., Goldmann, K., Pitzalis, C., McKeigue, P., & Barnes, M. R. (2023). nestedcv: an R package for fast implementation of nested cross-validation with embedded feature selection designed for transcriptomics and high-dimensional data. *Bioinformatics Advances*, *3*(1), vbad048.
Li, H.-D., Xu, Q.-S., & Liang, Y.-Z. (2018). libPLS: An integrated library for partial least squares regression and linear discriminant analysis. *Chemometrics and Intelligent Laboratory Systems*, *176*, 34-43.
Li, X., Morgan, P. S., Ashburner, J., Smith, J., & Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of neuroscience methods*, *264*, 47-56.
Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: a meta-analytic review. *Behavioral and Brain sciences*, *35*(3), 121-143.
Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. 2008 eighth ieee international conference on data mining,
Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the penn state worry questionnaire. *Behaviour research and therapy*, *28*(6), 487-495.
Nolen-Hoeksema, S., & Morrow, J. (1991). Response styles questionnaire. *Psychological Assessment*.
Panksepp, J. (1982). Toward a general psychobiological theory of emotions. *Behavioral and Brain sciences*, *5*(3), 407-422.
Patriat, R., Reynolds, R. C., & Birn, R. M. (2017). An improved model of motion-related signal changes in fMRI. *Neuroimage*, *144*, 74-82.
Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage*, *84*, 320-341.
Pruim, R. H., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J. K., & Beckmann, C. F. (2015). ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *Neuroimage*, *112*, 267-277.
Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1), 6086.
Rorden, C. (2025). MRIcroGL: voxel-based visualization for neuroimaging. *Nature Methods*, 1-2.
Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K., & Narasimhan, G. (2020). So you think you can PLS-DA? *BMC bioinformatics*, *21*, 1-10.
Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.
Saarimäki, H., Ejtehadian, L. F., Glerean, E., Jääskeläinen, I. P., Vuilleumier, P., Sams, M., & Nummenmaa, L. (2018). Distributed affective space represents multiple emotion

categories across the human brain. *Social cognitive and affective neuroscience*, *13*(5), 471-482.

Saarimäki, H., Gotsopoulos, A., Jääskeläinen, I. P., Lampinen, J., Vuilleumier, P., Hari, R., Sams, M., & Nummenmaa, L. (2016). Discrete neural signatures of basic emotions. *Cerebral cortex*, *26*(6), 2563-2573.

Satpute, A. B., Kragel, P. A., Barrett, L. F., Wager, T. D., & Bianciardi, M. (2019). Deconstructing arousal into wakeful, autonomic and affective varieties. *Neuroscience letters*, *693*, 19-28.

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., Eickhoff, S. B., Hakonarson, H., Gur, R. C., & Gur, R. E. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *Neuroimage*, *64*, 240-256.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research*, *92*(120), 57.

Spielberger, C. D., Gonzalez-Reigosa, F., Martinez-Urrutia, A., Natalicio, L. F., & Natalicio, D. S. (1971). The state-trait anxiety inventory. *Revista Interamericana de Psicologia/Interamerican journal of psychology*, *5*(3 & 4).

Thomas Yeo, B., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., & Polimeni, J. R. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, *106*(3), 1125-1165.

Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability*, *57*(1), 1-436.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: improved N3 bias correction. *IEEE transactions on medical imaging*, *29*(6), 1310-1320.

Winkielman, P., Davis, J. D., & Coulson, S. (2023). Moving thoughts: emotion concepts from the perspective of context dependent embodied simulation. *Language, Cognition and Neuroscience*, *38*(10), 1531-1553.

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage*, *14*(6), 1370-1386.

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*, *31*(3), 1116-1128.

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, *20*(1), 45-57.