Dynamic Functional Connectivity Encodes Generalizable Representations of Emotional Arousal But Not Valence

Jin Ke*[1] (jinke@uchicago.edu)
Hayoung Song[1] (hyssong@uchicago.edu)
Zihan Bai[1] (baizh@uchicago.edu)
Monica D. Rosenberg[1,2] (mdrosenberg@uchicago.edu)
Yuan Chang Leong*[1,2] (ycleong@uchicago.edu)

[1]Department of Psychology, [2]Neuroscience Institute, The University of Chicago, Chicago, IL 60637

Abstract

Human affective experience varies along the dimensions of valence (positivity or negativity) and arousal (high or low activation). It remains unclear how these dimensions are encoded in the brain and if the representations generalize across diverse situations. Here we utilized two publicly available functional MRI datasets of participants watching movies to build predictive models of moment-to-moment valence and arousal from functional correlations in brain activity. We tested the models both within and across datasets and identified a situation-general arousal representation characterized by the interaction between multiple large-scale functional networks. The arousal representation generalized to two additional datasets. Predictions based on multivariate patterns of activation underperformed connectome-based predictions and did not generalize. In contrast, we found no evidence of a situation-general valence representation. Together, our findings reveal a generalizable representation of arousal encoded in patterns of dynamic functional connectivity, revealing an underlying similarity in how arousal is encoded across affective contexts.

* Correspondence to:
Jin Ke (jinke@uchicago.edu)
Yuan Chang Leong (ycleong@uchicago.edu)

Human experience is characterized by a continual ebb and flow of emotions, shaping our attention[1,2] and memory[3,4], as well as guiding how we make decisions and interact with others[5–7]. One prominent psychological framework proposes that emotions can be structured along two core dimensions: valence, which reflects the positivity or negativity of an emotional state, and arousal, which indicates its intensity or activation level[8,9]. This framework, often referred to as the circumplex model due to its common visualization in a circular configuration with two axes, provides a unifying perspective to characterize emotions. For example, both “excitement” and “anger” are positioned towards the high-arousal axis but occupy opposite ends on the valence axis. By placing emotions along these two dimensions, the model underscores how seemingly distinct emotional experiences share underlying similarities and relate to one another in a structured space.

The circumplex model provides a foundational framework that has been influential in guiding emotion research and theory[10–12]. Yet, how valence and arousal states are represented in the brain remains an open question. One possibility is that valence and arousal representations are situation-general, that is, there are generalizable neural representations underlying valence and arousal across various contexts. In this perspective, there would be an inherent similarity in how the brain encodes arousal due to excitement from a thrilling movie scene and arousal due to anger in response to a tense interpersonal interaction. A second possibility is that neural representations of valence and arousal are situation-specific, meaning that the brain encodes different experiences with the same valence and arousal differently. An intriguing third possibility is that one of the dimensions is encoded in a situation-general manner while the other is encoded in a situation-specific manner. For example, arousal might be represented in a consistent manner across diverse situations, while the representation of valence is context-dependent. If the neural representation of arousal is situation-general and non-specific, it might explain why people sometimes mistakenly misattribute their arousal to the wrong cause[13–15]. Determining whether valence and arousal are encoded in a situation-general or situation-specific manner contributes towards a deeper understanding on the neural basis of affective experience.

To date, evidence about the neural basis of valence and arousal remains inconclusive. Multiple studies have identified brain activity linked to the valence and arousal of emotional stimuli, with some pointing to activity in specific brain regions, and others focusing on distributed patterns of activity or connectivity spanning multiple regions[16–23]. However, other findings challenge the generalizability of these neural correlates across contexts[24,25]. For example, one study found that activity patterns associated with viewing negative images did not generalize to thermal pain stimulation, even though both experiences were negative and arousing[24]. Relatedly, other studies have found that neural activity elicited by emotional stimuli are more consistent with categorically distinct emotions than a continuous space defined by valence and arousal[26,27], calling into question the claim that neural representations of affective experience are organized around the two dimensions.

The goal of this study is to test whether neural representations of valence and arousal are generalizable across different contexts. Our approach has three key features. First, while past research has primarily focused on univariate or multivariate neural activity[16–20,24–27], our study

tests the hypothesis that valence and arousal are encoded in dynamic interactions between large-scale functional brain networks. Specifically, we trained predictive models[28] to predict the valence and arousal of viewed stimuli from functional correlations in brain activity. This perspective recognizes that affective experience might arise not just from static patterns of activity, but also from transient brain states shaped by the coordination across different brain regions[29]. Recent work has identified brain states associated with a diverse range of cognitive functions, including fluctuations in attention[30,31], memory encoding[32,33], and comprehension of complex narratives[34]. However, the extent to which such brain states correspond to dynamic changes in affective experience remains unknown.

Second, rather than presenting participants with static images or short video clips designed to elicit a specific affective state, we examined continuous affective responses to full-length TV episodes. These episodes have comprehensive narrative arcs that elicited a range of affective responses[31,35,36] and allowed us to capture the moment-to-moment affective responses across a variety of situations. Third, we assessed the out-of-sample generalizability[37] of our findings by testing our models on multiple new stimuli that were not used to train the model. This step is crucial to ascertain whether our models were merely fitting to idiosyncrasies of a specific dataset or if they genuinely reflect generalizable neural representations of valence and arousal.

In the current study, we utilized two publicly available fMRI datasets of participants watching TV episodes (N = 16[35] and N = 35[36]; **Fig. 1**). We additionally collected continuous valence and arousal ratings of the two episodes from a separate group of participants (N = 120, 30 in each condition). We used dynamic connectome-based predictive modeling (CPM) developed by Song et al.[31,34] to separately predict moment-to-moment ratings of valence and arousal from whole-brain functional connectivity in each dataset. CPMs are so named because they leverage the connectome (i.e., the comprehensive map of neural connections in the brain) to make predictions about inter- or intra-individual differences in behavior or mental states. We assessed both the within-dataset accuracy (i.e., how well the CPM predicted valence or arousal in the dataset it was trained on) as well as across-dataset accuracy (i.e., how well a CPM predicted valence or arousal in the dataset it was not trained on). To further validate the robustness and generalizability of our results, we also tested the trained CPMs on two additional fMRI datasets of participants watching different movies. Additionally, we also repeated our analyses with univariate and multivariate methods to compare the predictive accuracy of the CPM against more traditional analytical approaches. Altogether, our approach seeks to identify generalizable neural representations of moment-to-moment valence and arousal based on the interaction between brain areas, and shed new light on how affective experiences are encoded in the brain.

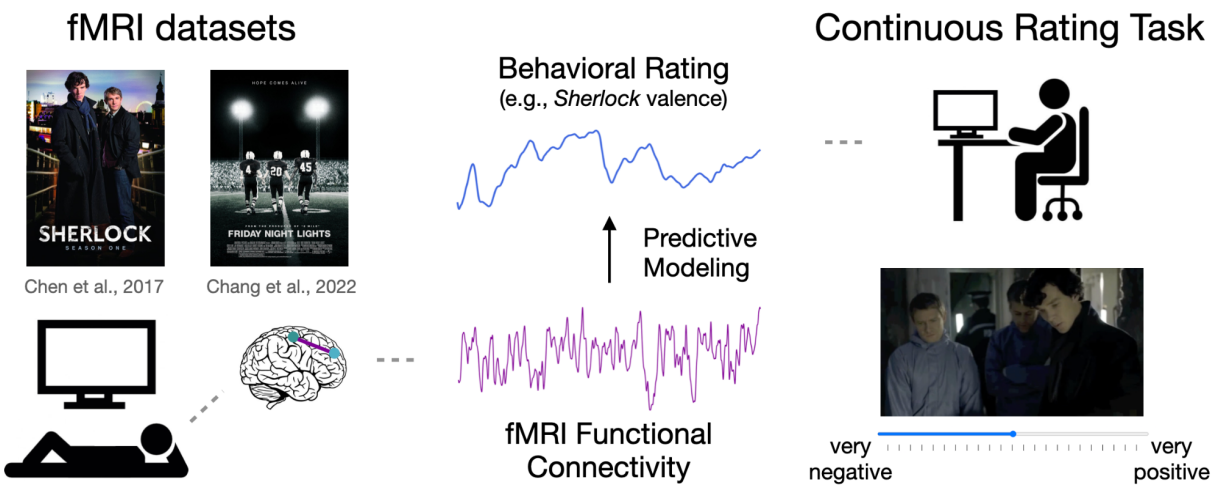

Figure. 1 Study schematic procedure. We utilized two publicly available fMRI datasets of participants watching *Sherlock* (*N* = 16) and *Friday Night Lights* (*N* = 35). Additionally, we collected behavioral ratings of the two episodes from a separate group of participants (*N* = 120). Each participant is presented with one movie clip and is instructed to continuously rate how the clip is making them feel in terms of either valence (i.e., positive to negative) or arousal (i.e., not aroused to aroused). We built within and across dataset CPMs to predict moment-to-moment affect ratings from dynamic whole-brain functional connectivity patterns.

Results

Affective experience is synchronized across individuals during movie watching

One-hundred and twenty participants performed an affect rating task while watching one of two TV episodes. One was a 48-min long episode from BBC's *Sherlock* (a British mystery crime drama series) while the other was a 45-min long episode from NBC's *Friday Night Lights* (an American sports drama series). Each participant provided continuous ratings of either valence (i.e., positive to negative) or arousal (i.e., not aroused to aroused) while watching one of the videos (**Fig. 1**). As we were primarily interested in subjective feelings during the movie, we instructed participants to indicate their affective experience (e.g., how positive or negative they were feeling), rather than the perceived emotionality of the movies (e.g., whether they thought the current scene was positive or negative). Previous work found that providing continuous affect ratings did not alter emotional or neural responses to emotion-eliciting films, indicating that this approach can capture participants' affective experiences without disrupting their natural responses to the content[38]. We will refer to the pairing of an affective dimension and a movie as an "experimental condition" (e.g., *Sherlock*-valence). Thus, we have four experimental conditions with a sample size of 30 participants each. The group-averaged time courses were treated as a proxy of the affective experience time-locked to the movie (**Fig. 2A**).

We first examined whether subjective ratings of valence and arousal were synchronized across individuals during movie watching. We computed the pairwise Pearson correlations between the ratings of every pair of participants in a condition (**Fig. S1**), and averaged the correlations across pairs (see Methods). The distribution of pairwise correlations was significantly positive in all conditions (**Fig. 2B**; *Sherlock*-valence: mean r = .487, *s.d.* = .141, p < .001; *Sherlock*-arousal: mean r = .263, *s.d.* = .105, p < .001; *Friday Night Lights*-valence: mean r = .432, *s.d.* = .112, p < .001; *Friday Night Lights*-arousal: mean r = .618, *s.d.* = .086, p < .001), with statistical significance assessed using a nonparametric permutation test[39]. These results suggest that affective experience, across both valence and arousal dimensions, was synchronized across individuals when they watched both TV episodes.

Qualitatively, the average valence and arousal time courses reflected expected fluctuations in affective experience during the episodes (**Fig. 2A**). For example, while watching *Sherlock*, participants' ratings indicated high negative valence and high arousal during the scene where John Watson has a nightmare of his time in the war, and high positive valence but low arousal during a scene where Sherlock was clueless to the lab assistant, Molly, flirting with him. Similarly, while watching *Friday Night Lights*, participants' ratings indicated high negative valence and high arousal when a star player is severely injured in a football game, and high positive valence and high arousal when the main character's team wins the game. The average valence and arousal time courses were not significantly correlated in both *Sherlock* (r = .094, p = .359) and *Friday Night Lights* (r = -.224, p = .187). The average time courses of arousal and the absolute value of valence were significantly correlated in *Friday Night Lights* (r = .706, p < .001) but not *Sherlock* (r = .101, p = .303).

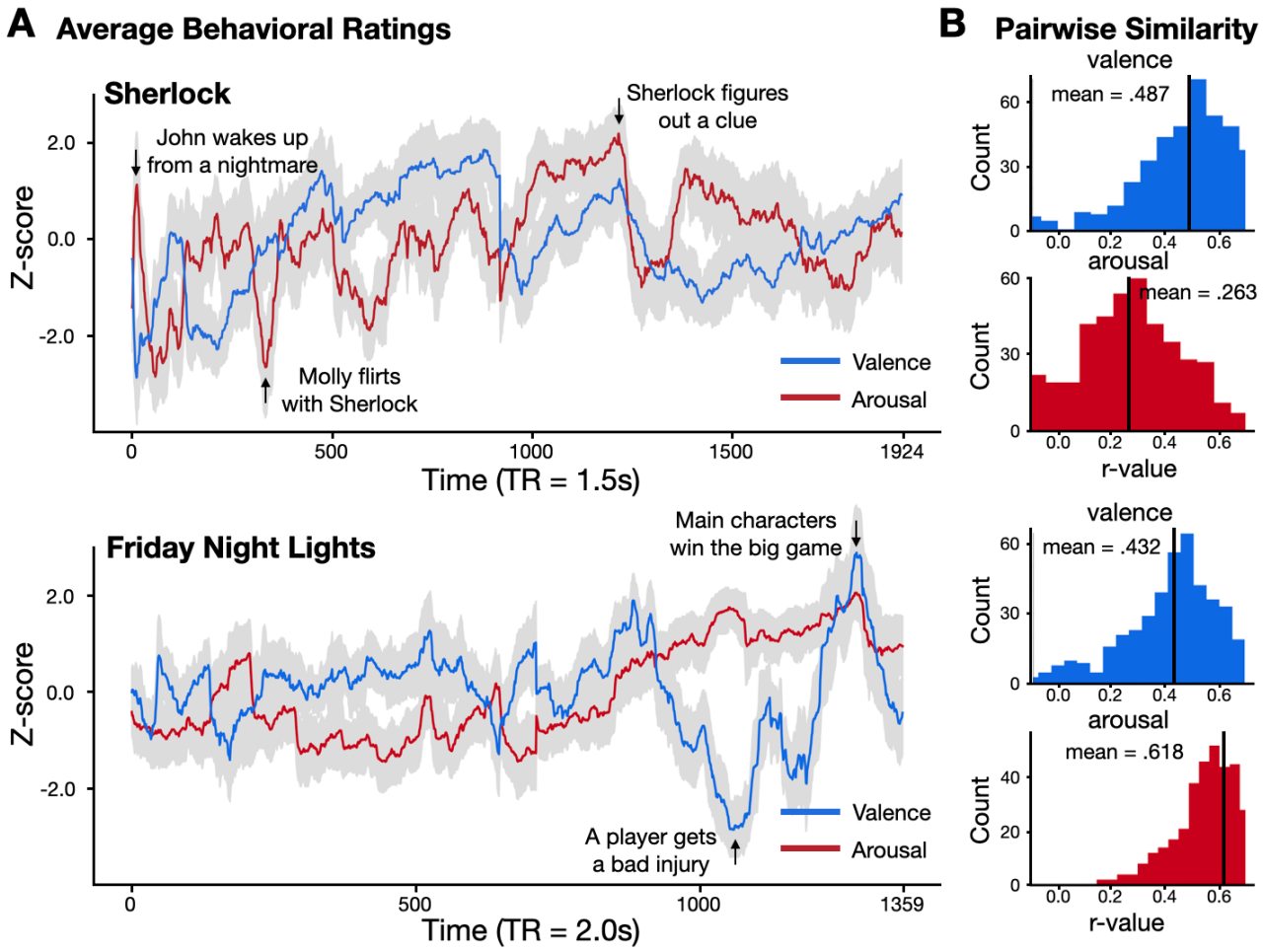

Fig. 2. Affective experience is synchronized across participants in all four experimental conditions. Each condition includes 30 participants. **A.** Participants' subjective affective experience fluctuates over time during naturalistic movie watching. The red and blue solid lines indicate average arousal and valence time courses respectively. The gray areas indicate 95% confidence interval (CI). **B.** Histograms of pairwise rating similarity between all participants in the four conditions. Higher mean *r* values indicates stronger affective synchrony.

Dynamic functional connectivity encodes arousal within and across datasets

We first asked whether dynamic functional connectivity encoded fluctuations in arousal. To that end, we trained CPMs to predict arousal rating time courses from time-resolved dynamic FC patterns (**Fig. 1**). We parcellated the whole brain into 122 ROIs, following the Yeo atlas for cortical regions[40] (114 ROIs) and the Brainnetome atlas for subcortical regions[41] (8 ROIs: bilateral amygdala, hippocampus, basal ganglia, and thalamus). The dynamic FC patterns were extracted from the 122-ROI-based BOLD time courses using a tapered sliding window approach, where the Fisher's z-transformed Pearson's correlation between the BOLD time courses of every pair of ROIs was computed within each tapered sliding window[31] (window size - *Sherlock*: 30 TRs = 45s; *Friday Night Lights*: 23 TRs = 46s; see Methods).

Separate models were trained on two openly available fMRI datasets where participants watched *Sherlock*[35] (*N* = 16) or *Friday Night Lights*[36] (*N* = 35). We assessed model performance in predicting group-mean arousal ratings within each dataset as well as between datasets. *Within-dataset performance* was computed using a leave-one-subject-out cross-validation approach, training the model on the neural data of all but one participant in a dataset, and applying the trained model to data from the held-out participant to predict the average arousal time course. In every round of cross-validation we selected FCs that significantly correlated with arousal in the training set as features (one-sample Student's *t* test, $p < .01$). Model accuracy was computed as the average correlation between the model-predicted arousal time course and the empirical arousal time course across cross-validation folds. Statistical significance was assessed by comparing model accuracy against a null distribution of 1000 permutations generated by training and testing the model on phase-randomized behavioral ratings.

Within-dataset accuracy was significantly above chance in predicting arousal ratings for both *Sherlock* (mean $r = .575$, $p = .034$) and *Friday Night Lights* (mean $r = .734$, $p = .002$; **Fig. 3A**, left), suggesting that dynamic whole-brain FC patterns encoded moment-to-moment experience of arousal during both movies. However, an alternative possibility was that the CPM was learning features specific to the particular movie it was trained on rather than arousal per se. For example, if Sherlock Holmes tended to appear in scenes that were highly arousing, the model could learn to associate the neural representation of Sherlock Holmes with higher arousal ratings. In a different movie, where Sherlock Holmes was not present, or no longer associated with arousing scenes, the model would fail to predict arousal. Thus, an important test of whether the model encodes arousal would require testing the model in a novel context.

With that goal in mind, we assessed the models' *across-dataset performance* to test whether the models generalized across datasets. For each dataset, we first identified the set of FC features that were selected in every round of the cross-validation, which we term the arousal network for that dataset. To test model generalizability, we trained a CPM in one dataset on its arousal network to predict the group-average arousal time course and applied the trained model to the other dataset. The across-dataset predictions of arousal (**Fig. 3A**, right) were significantly above chance both for a model trained on *Sherlock* and tested on *Friday Night Lights* (mean $r = .270$, $p = .010$), and a model trained on *Friday Night Lights* and tested on *Sherlock* (mean $r = .198$, $p = .001$).

To assess whether prediction results were driven by low-level audio-visual features, we reran the prediction analyses after regressing out ten low-level features (hue, saturation, pixel intensity, motion energy, presence of a face, whether the scene was indoor or outdoor, presence of written text, amplitude, pitch, presence of music) from each participant's BOLD signal time courses. Consistent with the previous analyses, we found above-chance predictions when the models were tested within each dataset (*Sherlock*: mean $r = .558$, $p = .024$; *Friday Night Lights*: mean $r = .732$, $p = .003$) and across datasets (train on *Sherlock* test on *Friday Night Lights*: mean $r = .267$, $p = .012$; train on *Friday Night Lights* test on *Sherlock*: mean $r = .220$, $p = .020$), suggesting that prediction accuracy was not driven by low-level features. Altogether, these results suggest that patterns of dynamic FC encode moment-to-moment fluctuations in arousal.

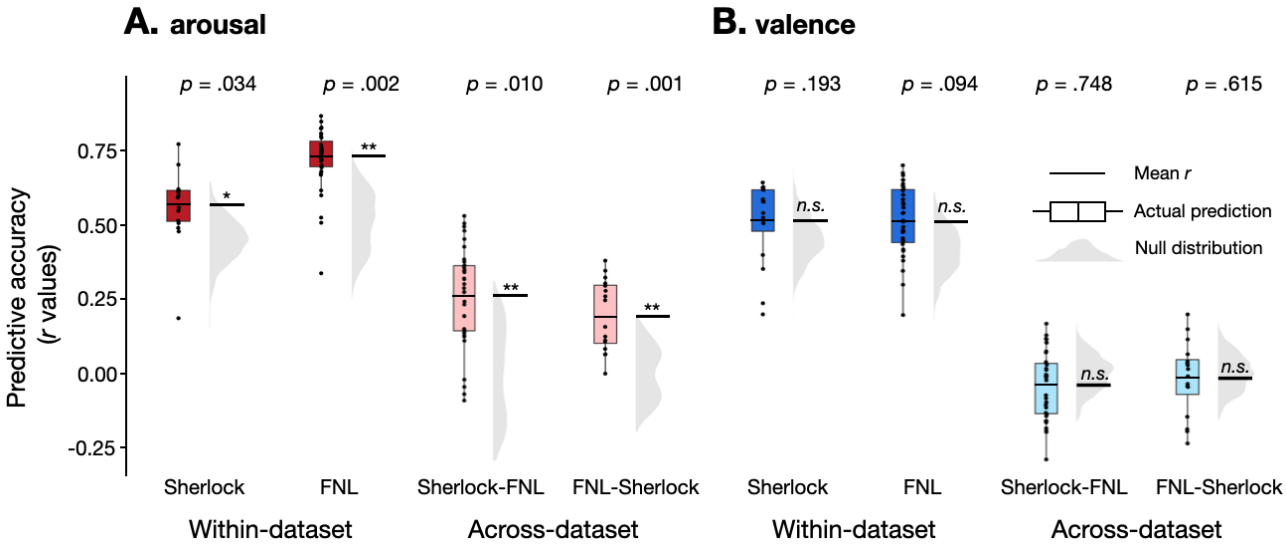

Fig. 3. Dynamic functional connectivity predicts fluctuations in arousal but not valence. CPM performance in predicting arousal (**A**) and valence (**B**) for within-dataset (the left panel) and between-dataset (the right panel). The y-axis represents the predictive accuracy, as measured by Pearson's correlation between the model predicted time course and the observed group-average time course. Each datapoint in the box plot represents the predictive accuracy of each round of cross-validation. The black horizontal lines show the Fisher-*z* transformed mean *r* value. The gray half-violin plots show the null distribution of 1000 permutations, generated by phase-randomizing the observed group-average time course before training and testing the models. *$p < 0.05$, **$p < 0.01$, n.s.: $p > 0.05$, as assessed by comparing the empirical mean predictive accuracy against the null distribution.

Arousal is encoded in FC patterns within and between large-scale functional networks

Is arousal encoded in the FCs within specific functional networks (e.g., dorsal attention network, default mode network), or is it encoded in the interactions between multiple networks? To answer this question, we examined the functional connections within the arousal network in each dataset. We term the functional connections that were positively or negatively correlated with the arousal time course as positive or negative features respectively. The *Sherlock* arousal network includes 593 FC features (439 positive and 154 negative), and the *Friday Night Lights* arousal network includes 1578 FC features (848 positive and 730 negative).

We defined the set of overlap FC features between the *Sherlock* positive arousal network and the *Friday Night Lights* positive arousal network as the high-arousal network, and the set of overlap FC features between the *Sherlock* negative arousal network and the *Friday Night Lights* negative arousal network as the low-arousal network. We then assessed the degree of overlap in the high- or low-arousal networks across datasets. Both the high-arousal network (169 overlapping FC features, $p < .001$) and low-arousal network (68 overlapping FC features, $p < .001$) have an above-chance number of overlapping FC features across the two datasets. In contrast, the *Sherlock* high-arousal network does not significantly overlap with the *Friday Night Lights* low-arousal network (1 overlapping FC feature, $p > .99$), and the *Friday Night Lights*

high-arousal network does not significantly overlap with the *Sherlock* low-arousal network (4 overlapping FC features, $p > .99$). Statistical significance of network overlap was assessed using a hypergeometric cumulative density function[42].

To further understand how connections within and between large-scale functional networks constitute the high- and low-arousal network, we followed Yeo et al.[40] to group the 122 ROIs into 8 canonical functional networks, namely, the visual (VIS), somatosensory-motor (SOM), dorsal attention (DAN), ventral attention (VAN), limbic (LIM), frontoparietal control (FPN), default mode (DMN), and subcortical (SUB) networks. We asked whether particular functional networks were represented in the arousal network more frequently than chance (**Fig. 4A**). We computed the proportion of selected FCs among all possible FCs between each pair of functional networks, and assessed the significance of the proportions by comparing them against a null distribution of 10000 permutations where functional connections were randomly distributed across networks (see Methods). Connections between the DMN and FPN, DAN and FPN, and DAN and VAN, as well as connections within the DMN were positively associated with arousal. In contrast, connections between the DAN and VIS network were negatively associated with arousal (FDR-corrected $p < .05$). These results suggest generalizable neural representations of arousal are encoded within the DMN as well as between pairs of distributed networks.

A recent study analyzing the *Sherlock* dataset revealed that self-reported attentional engagement is correlated with scene-by-scene annotations of emotional arousal during narratives. This previous study characterized a set of functional networks that predicts engagement across datasets[31]. The average engagement rating from that study was significantly correlated with the average arousal ratings that we collected (Pearson's $r = .778$, $p = .014$) but not the average valence ratings (Pearson's $r = .085$, $p = .404$). Relatedly, the arousal and engagement networks within the *Sherlock* dataset significantly overlap (361 overlapping FC features, $p < .001$; no overlap between the high arousal/low engagement and low arousal/high engagement networks). The shared FC features constitute 60.88% of FCs in the arousal network and 52.79% of FCs in the engagement network, and were distributed across different functional networks (**Fig. S2**). However, predictions of a CPM trained on engagement ratings were not significantly correlated with arousal ratings (mean $r = .498$, $p = .173$). Moreover, the correlation was significantly lower than that of a model that was trained on arousal ratings ($r = .498$ vs. $.558$; paired-t test, $t(15) = -4.368$, $p < .001$). These results suggest that attentional engagement and arousal are closely related but non-synonymous constructs, with arousal possibly comprising an aspect of attentional engagement[43,44].

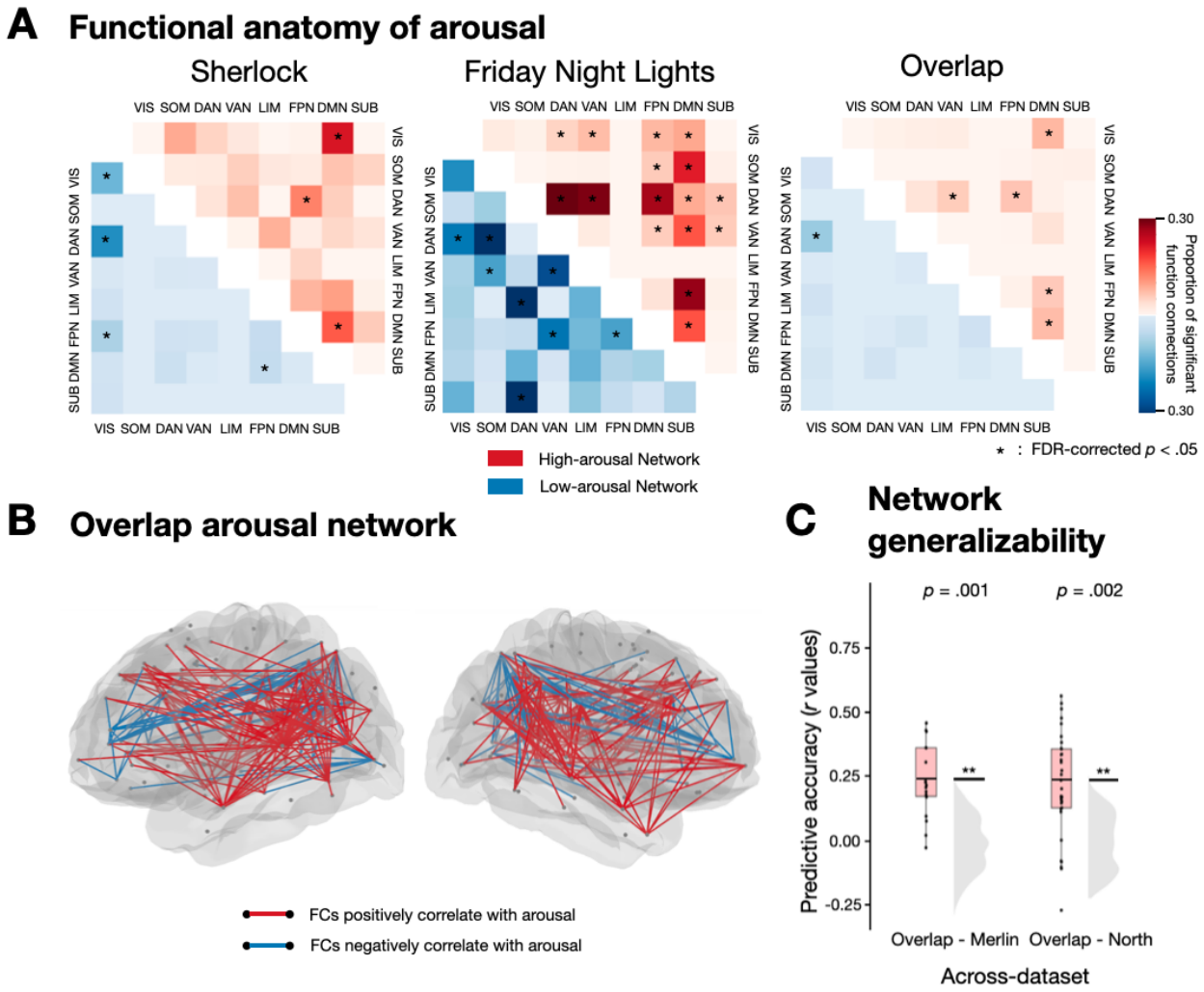

Fig. 4. Functional anatomy of arousal. A. Arousal networks for *Sherlock*, *Friday Night Lights*, and their overlap. Each cell represents the proportion of selected FCs among all possible FCs between each pair of functional networks. The cells in the upper triangle represent the high-arousal network (red), and the cells in the lower triangle represent the low-arousal network (blue). Network pairs with above-chance selected FCs are indicated with an asterisk (one-tailed *t*-test, FDR-corrected $p < .05$). **B.** Visualization of the overlap arousal network. Each dot represents a node in the brain. The lines connecting two nodes show their functional connection. Red lines represent functional connections that positively correlate with arousal. Blue lines represent functional connections that negatively correlate with arousal. (**C**) Connectome-based model trained on the overlap arousal network generalized to two more fMRI datasets, *Merlin* and *North by Northwest*. $^{**}p < .01$, n.s.: $p > .05$, as assessed by comparing the empirical mean predictive accuracy against the null distribution.

Connectome-based models of arousal generalized to two more fMRI datasets

Our findings that CPMs trained on one movie predicted arousal in another largely distinct movie provide evidence that dynamic functional connectivity encodes generalizable representations of arousal. We then proceeded to test whether our identified arousal network would generalize to additional datasets. To that end, we analyzed two additional fMRI movie datasets with distinct genre, storyline, characters, and duration. The first was a 15-min clip from the movie *North by Northwest* (a spy thriller directed by Alfred Hitchcock; N = 32) collected by our group, and the second was a 25-min episode from BBC's *Merlin* (a British fantasy adventure drama series; N = 18) from Zadbood et al[45]. To measure affective experience during the movie clips, we collected behavioral ratings of arousal and valence for both movies (N = 60, 15 in each condition).

We tested the hypothesis that dynamic FC patterns encoded an arousal neural representation that generalizes across movie stimuli. To that end, we trained a CPM on both *Sherlock* and *Friday Night Lights* datasets with the overlap arousal network (169 positive FCs and 68 negative FCs; **Fig. 4B**) as input features. We then tested the model separately on the *North by Northwest* and *Merlin* datasets. Prediction accuracy was above-chance in both *North by Northwest* (mean r = .227, p = .002) and *Merlin* (mean r = .230, p = .004). Model-predicted arousal time courses also corresponded with the plot of each movie. For example, in *North by Northwest*, the most arousing moment predicted by the model occurred when the protagonist was being chased by a plane, while the least arousing moment occurred during a long conversation between characters (**Fig. S3**, top). Similarly, in *Merlin*, the most arousing moment predicted by the model occurred when the two main protagonists had a brawl in a tavern, while the least arousing moment occurred when a nameless soldier walked towards a campfire (**Fig. S3**, bottom). These results indicate robust out-of-sample generalizability, suggesting that the models were not merely capturing the idiosyncrasies of specific datasets. Instead, the arousal networks that we identified encode neural representations of arousal that generalized across diverse situations.

Given a movie-watching fMRI dataset, how well can the model predict the moment-to-moment arousal fluctuations of the average participant? In the prior analyses, we correlated the model predictions from each participant's data with the group-average arousal ratings, and then averaged the correlations across the sample. This allowed us to assess the extent to which predictions from each individual matched the group average. If our goal was instead to maximize our ability to predict the group-average experience of arousal, we can first aggregate the model predictions across participants and predict average arousal ratings from the aggregated model predictions. To that end, we averaged the model predictions across participants watching the same movie and computed the correlation between the group-average model predictions and group-average arousal ratings. The group-average predicted arousal time course was significantly correlated with the group-average human-rated arousal time course in both *North by Northwest* (r = .649, p < .001) and *Merlin* (r = .579, p < .001; **Fig. S3**). These results provide a measure of the model's ability to predict arousal fluctuations of an average individual watching a movie that the model was not trained on, and suggest that our model could be used to generate moment-to-moment predictions of arousal during movie watching in new, independent datasets.

Dynamic functional connectivity does not predict moment-to-moment valence

Having established that dynamic functional connectivity encodes emotional arousal, we examined whether the same approach can predict valence, the other core dimension of affective states in the circumplex model. Within-dataset CPMs did not show above-chance predictive accuracy in either *Sherlock* (mean r = .518, p = .193) or *Friday Night Lights* (mean r = .514, p = .094). Across-dataset predictive accuracy was also not significantly better than chance (train on *Sherlock* test on *Friday Night Lights*: mean r = -.041, p = .748; train on *Friday Night Lights* test on *Sherlock*: mean r = -.016, p = .615; **Fig. 3B**). Equivalence tests[46–48] indicated that these accuracies were statistically indistinguishable from zero within bounds defined by a minimal effect size of interest [-.100, .100] (see **Fig. S4**). One possible explanation of these null results is that the group-average valence ratings might not be sufficiently reliable across participants, potentially due to idiosyncratic experience of subjective valence. However, the inter-subject agreement of the valence ratings was not systematically different than that of arousal in both datasets, and valence and arousal showed comparable overall inter-subject agreement across-datasets (valence: mean r = .474; arousal: mean r = .460; t-test: $t(59)$ = 1.320, p = .187). These results suggest that the model's poor performance at predicting valence fluctuations were unlikely to be due to lower reliability of the behavioral valence ratings.

The previous analyses assumed that valence is encoded as a single "bipolar" dimension, with positive and negative valence at opposing ends[8,12]. An alternative possibility is that positive and negative valence are encoded separately in the brain[49–52], in which case, a CPM trained on data that spanned the entire valence spectrum would not perform well. To test this alternative possibility, we trained separate CPMs on positive and negative moments in each movie. For positive valence, we observed above-chance within-dataset predictive accuracy in *Friday Night Lights* (mean r = .762, p = .001) but not Sherlock (mean r = .431, p = .835). However, the model failed to generalize across-dataset (from *Sherlock* to *Friday Night Lights*: mean r = -.020, p = .528; from *Friday Night Lights* to *Sherlock*: mean r = .088, p = .247; **Fig. S5B**, left). For negative valence, neither within-dataset (*Sherlock*: mean r = .553, p = .400; *Friday Night Lights*: mean r = .647, p = .192) nor across-dataset (from *Sherlock* to *Friday Night Lights*: mean r = -.093, p = .875; from *Friday Night Lights* to *Sherlock*: mean r = -.216, p = .898; **Fig. S5B**, right) models showed significant predictive accuracy. As such, we were unable to identify generalizable neural representations of valence encoded in dynamic functional connectivity, even when considering positive valence and negative valence separately.

Connectome-based models of valence failed to generalize even with more training data

One reason why the valence CPMs failed to generalize may be because they were trained on insufficiently diverse contexts. We thus sought to test if a valence CPM would generalize if trained on more data. To that end, we trained a CPM on the combined data from *Sherlock* and *Friday Night Lights*, with the 2067 FCs that significantly correlated with valence in the combined dataset as input features. This model failed to generalize to both *Merlin* (mean r = -.010, p = .611) and *North by Northwest* (mean r = -.140, p = .930), indicating that despite training the model on multiple datasets, we were still unable to identify generalizable neural representations of valence encoded in dynamic functional connectivity.

Univariate parametric mapping and multivariate-pattern based predictions of arousal and valence

Prior work has employed univariate parametric mapping and multivariate pattern analysis (MVPA) to examine the neural correlates of human affective experience[16–20]. Here, we asked if we can identify generalizable neural representations of arousal and valence using these approaches.

We first examined whether the univariate parametric maps of arousal shared overlapping brain activations across movies. We implemented a generalized linear model (GLM) to examine the association between BOLD activity and arousal ratings. We performed a conjunction analysis[53] to identify the voxels consistently significantly related to arousal across the *Sherlock* and *Friday Night Lights* datasets (TFCE-corrected $p < .01$). We observed clusters of voxels in regions including the cingulate gyrus, precuneus, and insula (**Fig. S6**, left). We then ran the same GLM to test the association between the BOLD activity and the valence ratings. In contrast to the arousal maps, few voxels exhibit significant correlations with valence across both movies (**Fig. S6**, right). These results suggest that, in our data, the same univariate activity encoded arousal, but not valence, across situations.

Having identified the voxels that were related to valence and arousal, we proceeded to investigate whether the multivariate activation patterns across voxels could predict moment-to-moment valence and arousal fluctuations. We repeated the same feature selection and predictive modeling approach that we used for training the dynamic CPM, but instead of FC patterns, we used multivariate voxel time courses as inputs to the model. Multivariate-pattern based models (MPMs) yielded significant within-dataset predictions of arousal in *Friday Night Lights* (mean $r = .498$, $p = .024$) but not Sherlock (mean $r = .409$, $p = .160$, **Fig. S7**). The MPMs on arousal failed to generalize across datasets, both when training on *Sherlock* and testing on *Friday Night Lights* (mean $r = .024$, $p = .327$), as well as when training on *Friday Night Lights* and testing on *Sherlock* (mean $r = .054$, $p = .133$, **Fig. 5A**, left).

Model predictions of valence were not significant within-dataset (*Sherlock*: mean $r = .420$, $p = .120$.; *Friday Night lights*: mean $r = .369$, $p = .344$) or across datasets (from *Sherlock* to *Friday Night lights*: mean $r = .001$, $p = .604$; from *Friday Night lights* to *Sherlock*: mean $r = -.001$, $p = .580$; **Fig. 5A**, right). Together, these results indicate that we were not able to identify generalizable representations of either valence or arousal from multivariate activation patterns.

Finally, we investigated whether CPMs outperformed MPMs in predicting arousal by comparing their Fisher's *z*-transformed predictive accuracies. CPM yielded significantly higher predictive accuracy than MPM both within-dataset (Paired *t*-test: *Sherlock*: $t(15) = 6.963$, $p < .001$; *Friday Night Lights*: $t(34) = 18.056$, $p < .001$) and across datasets (Paired *t*-test: from *Sherlock* to *Friday Night Lights*: $t(34) = 8.512$, $p < .001$; from *Friday Night Lights* to *Sherlock*: $t(15) = 3.903$, $p = .001$; **Fig. 5B**, right). These results suggest that dynamic functional connectivity patterns may encode a more reliable arousal representation than multivariate activation patterns.

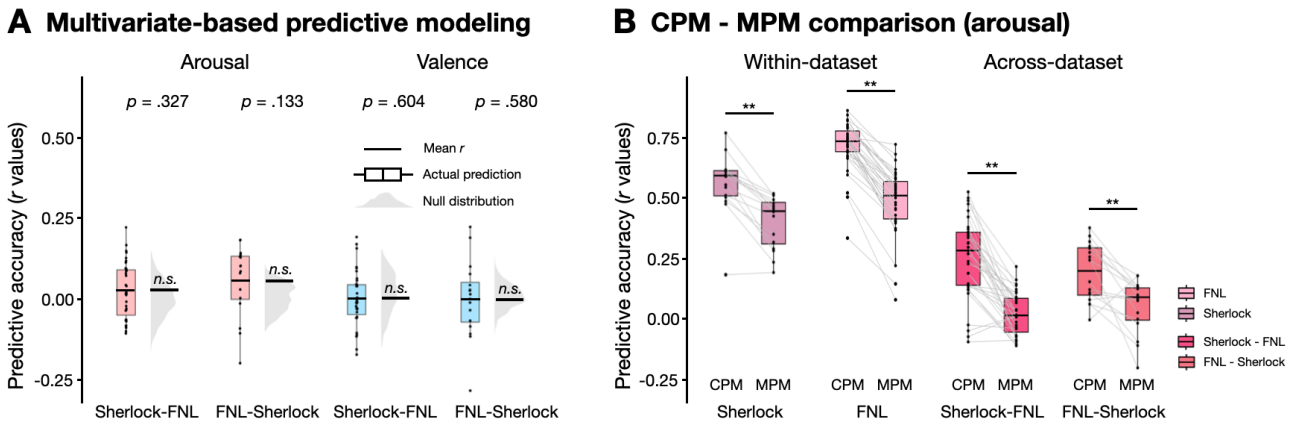

Fig. 5. A. Multivariate-pattern based model performance in predicting arousal and valence across-datasets. n.s.: $p > .05$, as assessed by comparing the empirical mean predictive accuracy against the null distribution. **B.** Comparison between the predictive accuracy of connectome-based predictive modeling (CPM) and multivariate-based predictive modeling (MPM) for both within-dataset (left panel) and across-dataset predictions (right panel). Each datapoint in the box plot represents the predictive accuracy of each round of cross-validation. The gray lines connect the same held-out participant in CPM and MPM. $^{**}p < .01$.

Discussion

This study utilized predictive modeling of affective experience to examine how valence and arousal states are represented in the human brain. We first showed that subjective valence and arousal were synchronized between individuals watching the same movie. We then identified a situation-general neural representation of emotional arousal encoded in patterns of dynamic functional connectivity. Specifically, a model trained to predict arousal from dynamic functional connectivity during one movie generalized to a second movie, and the arousal representation further generalized to two additional datasets. Moreover, predictive models trained on dynamic functional connectivity predicted arousal more accurately than predictive models trained on multivariate activation patterns. In contrast to the situation-general representation of arousal, we found no evidence of a situation-general representation of valence. These results suggest that the neural encoding of arousal might be intrinsically similar across different situations, while the neural encoding of valence might be situation-specific.

A strength of the current work is that, in addition to identifying the neural correlates of affective experience in a particular context, we explicitly tested for out-of-sample generalizability. This is important because predictive models trained and tested within the same context run the risk of fitting to the idiosyncrasies of a given stimulus. For example, a model might learn the association between high arousal and the presence of a particular character in a narrative, and will fail to predict arousal in a different narrative with different characters. A situation-general neural representation of latent psychological construct should be consistent across different contexts and situations[54–56]. Our results indicate that the neural representation of arousal encoded in dynamic functional connectivity fulfills this criteria of generalizability — the CPMs were able to predict fluctuations in arousal in movies which the models had not previously been trained on, and that also differed in low-level audio-visual features, characters, plot and genre. The robustness of the model across stimuli suggests a shared neural representation of the subjective experience of arousal across situations. That is, there seems to be an intrinsic similarity in how the brain encodes the arousal from a suspenseful moment in *Sherlock* and the arousal from an exciting moment in *Friday Night Lights.*

These findings resonate with and have implications for classic theories of emotion. In particular, the two-factor theory of emotion proposes that emotional experiences arise from the combination of physiological arousal and the cognitive interpretation of that arousal[13]. To interpret arousal, people rely on contextual cues in the immediate environment. When the source of arousal is ambiguous, individuals can attribute experienced arousal to an incorrect source, a phenomenon termed the misattribution of arousal[57–60]. For example, a person who arrives at a job interview after climbing up a flight of stairs might mistakenly attribute the arousal from the physical activity to nervousness about the interview, which can in turn affect their performance on the interview. In this view, arousal states are inherently confusable and susceptible to contextual misattribution. Our results suggest that arousal states may be confusable because there is an intrinsic similarity in how the brain encodes them across contexts. This shared representation of arousal could serve as a fundamental substrate upon which emotional experience is constructed and interpreted[61–63].

Our work extends prior studies that examined the neural correlates of arousal in univariate activity and multivariate patterns[16–18,20,64–66]. Our findings highlight the value of considering the connectivity between large-scale functional networks in addition to activity patterns within specific regions. A recent study by Young and colleagues[23] had found that connectivity within the salience network (corresponding closely to the ventral attention network), and between the salience network and the executive control network (corresponding closely to the frontoparietal control network) was associated with fluctuations in physiological arousal, as measured by heart rate. Our study complements and advances this prior work in several ways. In particular, we examined the neural representation of arousal in multiple movies with diverse affective contexts (e.g., excitement, suspenseful, comedic) and demonstrated that our model generalized across movies that the model was not trained on. In contrast to the Young et al. study, we identified connections within and between multiple networks, including the ventral attention network, the frontoparietal attention network, the dorsal attention network and the default mode network. Future work will be needed to disentangle whether these discrepancies are due to differences in the affective contexts tested (stress vs. more generalized arousal) or differences in how arousal was measured (physiological signal vs. subjective ratings).

The involvement of multiple large-scale functional networks in encoding arousal identified in our study aligns well with the growing work indicating that arousal is associated with enhanced integration across different functional networks in the brain[67,68]. Network integration refers to a state of correlated activity across distinct functional networks and is thought to facilitate rapid and coordinated behavioral responses to environmental stimuli[69,70]. For example, during heightened states of arousal, there may be increased coordination between regions involved in sensory processing, attention, and decision-making, reflecting a brain state optimized for processing salient information and preparing the body for action. As such, increased network integration induced by arousal could underlie the brain's capacity to initiate a fight-or-flight response. The generalizability of the arousal representation suggests that this arousal response may be context-independent rather than stimulus-specific.

Methodologically, our research presents a novel approach for probing moment-to-moment arousal ratings during movie watching. The collection of arousal ratings from human participants is labor-intensive and time-consuming. With the growing availability of open fMRI datasets with movie-watching data[71–73], our model can be a valuable tool for researchers who wish to obtain continuous measures of arousal without having to collect additional human ratings. A promising and important extension of the current work is to assess if the arousal representation would generalize to participants engaging in other tasks, such as viewing images or making decisions.

In contrast to arousal, we did not find evidence of a generalizable representation of valence. The correlation in valence ratings across individuals was not systematically lower for valence ratings than for arousal ratings, suggesting that the poor performance of the model was unlikely due to idiosyncratic experience of valence between participants. An alternative explanation for the poor model performance is that our models were trained across participants, while neural representations of valence may be individual specific. As we do not have data of the same participant watching multiple movies, this is not an explanation we were able to rule out.

Isolating individual-specific neural representations of valence would likely require a “precision functional mapping” approach[74], where extensive functional data is collected from the same individual experiencing diverse affective contexts.

We note, however, that our results are consistent with past studies suggesting that the neural representation of valence may be situation-specific rather than situation-general[75–78]. For example, a meta-analysis on 397 human neuroimaging studies by Lindquist and colleagues[79] found no evidence of brain regions that monotonically increase or decrease with valence. The authors suggest that valence may instead be flexibly encoded depending on the situation. Our univariate and multivariate prediction results mirror these findings, and the CPM prediction results further suggest that connectivity patterns are also unlikely to encode generalizable valence representations. Altogether, these findings are consistent with theories of emotion positing that emotional experiences are not uniform across contexts but are heterogeneous and constructed from a mix of situational cues[62], cognitive appraisals[80], and social-environmental constraints[81]. According to these theories, the “joy” from seeing an old friend after years of separation would not be the same “joy” due to winning an Olympic gold medal. The two experiences, even if similarly positively valenced, would engage different neural computations and circuits. Thus, the neural representation of valence would not be similar across situations. As our study did not exhaustively rule out other possibilities of how a situation-general representation of valence might be encoded in the brain, this interpretation remains tentative. Nevertheless, we believe that explorations into the neural mechanisms that act on situation-general representations of arousal to construct situation-specific representations of valence would be a fruitful direction for future research.

In summary, our study identified a generalizable neural representation of arousal encoded in dynamic functional connectivity, but did not find a parallel generalizable representation of valence. The findings highlight the relationship between arousal states and the dynamics of large-scale functional networks. Our work extends our understanding of how affective experience is encoded in the brain, and provides a methodological approach of probing the neural basis of affective experience using functional neuroimaging.

Methods

Subjects

One hundred and twenty individuals participated in the behavioral experiment (83 female, mean age 20.45 $\pm$ 2.15 years). Subjects were randomly assigned to four experimental conditions, namely, *Sherlock* valence, *Friday Night Lights* valence, *Sherlock* arousal, and *Friday Night Lights* arousal. Each condition includes data from 30 subjects. The rating time courses were significantly correlated across individuals in each condition (**Fig. S1**). Participants provided informed written consent before the start of the study in accordance with the experimental procedures approved by the Institutional Review Board of the University of Chicago and were compensated for participation with case or credits for class.

Stimuli and experiment design

A 48min 6s segment of the BBC television series *Sherlock*[35], and a 45min 18s segment of the NBC serial drama *Friday Night Lights* (FNL)[36] were used as the audio-visual movie stimuli in the study. The 30s audiovisual cartoon (*Let's All Go to the Lobby*) was removed from the original *Sherlock* movie stimuli. Both stimuli were divided into two segments (23min and 25min 6s for *Sherlock*, 23min 45s and 21min 33s for *FNL*). We prepended a 5s countdown video to the beginning of each segment to prepare participants to the beginning of the movie. Before rating the movie, participants performed a practice task where they rated valence or arousal of a short video from the movie *Merlin*[45] (2:03-3:10 for valence, and 5:13-6:19 for arousal) to familiarize themselves with the task.

Participants watched either *Sherlock* or *FNL*, and continuously rated either how positive or negative (valence conditions), or how aroused or not aroused (arousal conditions) the videos made them feel at each moment while watching the video. Participants who provided valence ratings were told that positive valence refers to when they were feeling pleasant, happy, excited, and negative valence refers to when they were feeling unpleasant, sad, angry. Participants who provided arousal ratings were told that high arousal refers to when they were feeling mentally or physically alert, activated, and/or energized, and low arousal refers to when they were feeling very mentally or physically slow, still, and/or de-energized. Participants pressed "left" or "right" keys on a keyboard to adjust a slider bar on a 1-25 scale. The two ends of the scale were labeled as "Very Negative" and "Very Positive" for participants in the valence conditions, and "Completely Not Aroused" to "Completely Aroused" for participants in the arousal conditions. Participants were encouraged to move the slider bar for even slight changes in how they felt. The button presses were recorded by a program coded in jsPsych (www.jspsych.org). Participants had the option to take a break after they watched the first video segment and continued to watch the second segment whenever they felt ready to do so. Participants completed a post-experiment survey after they finished the main rating experiment, where they reported the overall valence or arousal rating of the two segments, demographics and whether they had watched the video before. 15, 17, 3, and 5 participants had viewed the movie episode before the experiment in the *Sherlock*-valence, *Sherlock*-arousal, *FNL*-valence, *FNL*-arousal condition respectively. As approximately half of the participants had previously watched

Sherlock, we checked if the reliability of the ratings differed between participants who had or had not previously watched the episode. There was no significant difference in the pairwise rating similarity of the two groups both for the arousal ($t(29) = 1.157$, $p = .257$) and valence ratings ($t(29) = 1.543$, $p = .134$). In addition, the arousal models trained on the behavioral ratings from participants who had (mean r = .234, p = .019) and had not (mean r = .297, p = .008) seen the movie generalized to *Friday Night Lights*.

Movie analysis and annotations

We extracted 7 low-level visual features (hue, saturation, pixel intensity, motion energy, presence of a face, whether the scene was indoor or outdoor, presence of written text) and 3 low-level auditory features (amplitude, pitch, presence of music) from both *Sherlock* and *FNL*. We computed hue, saturation, pixel intensity ('rgb2hsv'), audio amplitude ('audioread') and pitch ('pitch') in MATLAB[82] (R2022b, The Mathworks, Natick, MA), and motion energy (pliers.extractors.FarnebackOpticalFlowExtractor) in Python[83] (version 3.9). Additionally, author J.K. coded whether each video frame contained a face (presence of face = 1, absence of face = 0), written text (presence of written text = 1, absence of written text = 0), whether the scene was indoor or outdoor (indoor = 1, outdoor = 0), and whether music played in the background (presence of music = 1, absence of music = 0).

Behavioral data analysis

Considering that rating changes can happen at any time during the movie and the time points of the changes differ across subjects, we resampled the ratings to one rating per TR (1.5s for *Sherlock* and 2s for *FNL*). Resampled time courses of the two movie segments from the same subject were concatenated, and z-scored across time. To compute the intersubject correlation of the affect ratings, we computed the Pearson correlation between the rating time courses for each pair of subjects within each condition (i.e., *Sherlock*-valence, *Sherlock*-arousal, *FNL*-valence, *FNL*-arousal). The pairwise *r*-values were then Fisher *z*-transformed and averaged across all pairs. An inverse transform was then applied to the resulting average *z*-value to obtain the average mean intersubject correlation for a given condition. To test whether group-mean ISC significantly deviated above zero, we applied a nonparametric permutation approach (10000 permutations) where 50% chance of sign-flipping was applied to every subject's rating similarity before averaging into group-mean. The normalized rating time courses of all subjects within each condition were averaged, and the group-average time courses were treated as the ground-truth affective experience time-locked to the stimuli.

Further fMRI analysis in the study involved the application of a tapered sliding window. In order to align the behavioral time courses to fMRI data time courses, we convolved the group-average time courses with the hemodynamic response function (HRF), and applied a tapered sliding window to the convolved behavioral time courses. We applied a sliding window size of 30 TR (45s) for *Sherlock* and 23 TR (46s) for *FNL*, and a step size of 1TR and a Gaussian kernel σ = 3TR. We followed Song et al.[31] in determining the weights for cropping the tail of the Gaussian kernel at the beginning and end of the time courses.

MRI acquisition and preprocessing

We acquired the raw structural and functional images of the *Sherlock* dataset from Chen et al.[35], and the *FNL* dataset from Chang et al.[36]. The *Sherlock* data were collected on a 3T full-body scanner (Siemens Skyra) with a 20-channel head coil. Functional images were acquired using a T2*-weighted echo-planar imaging (EPI) pulse sequence (TR/TE = 1500/28 ms, flip angle 64°, whole-brain coverage 27 slices of 4 mm thickness, in-plane resolution 3 × 3 mm^2, FOV 192 × 192 mm^2), with ascending interleaved acquisition. Anatomical images were acquired using a T1-weighted MPRAGE pulse sequence (0.89 mm^3 resolution)[35]. The *FNL* data were collected on a 3T Philips Achieva Intera scanner with a 32-channel head coil. Functional images were acquired in an interleaved fashion using gradient-echo echo-planar imaging with prescan normalization, fat suppression and an in-plane acceleration factor of two (i.e., GRAPPA 2), and no multiband (i.e., simultaneous multislice) acceleration (TR/TE = 2000/25 ms, flip angle 75°, resolution 3 mm^3 isotropic voxels, matrix size 80 by 80, and FOV 240 × 240 mm^2). Anatomical images were acquired using a T1-weighted MPRAGE sequence (0.90 mm^3 resolution)[36].

We applied the same preprocessing pipeline to the *Sherlock* and *FNL* dataset, which included MCFLIRT motion correction, high-pass filtering of the data with a 100-ms cut-off, and spatial smoothing using a Gaussian kernel with a full-width at half-maximum (FWHM) at 5 mm. The functional images were resampled to 3 mm^3 isotropic space, and registered to participants' anatomical images (6 d.f.), then to the Montreal Neurological Institute (MNI) space using affine transformation (12 d.f.). Preprocessing was performed using FSL/FEAT v6.00 (FMRIB software library, FMRIB, Oxford, UK).

Dynamic connectome-based predictive modeling

We applied dynamic connectome-based predictive modeling, an approach introduced in Song et al.[31] and available at https://github.com/hyssong/NarrativeEngagement, to predict arousal and valence. We first extracted BOLD signals across the whole brain using the 114-ROI cortical parcellation scheme of Yeo et al., along with the 8-ROI subcortical parcellation from the Brainnetome atlas (bilateral amygdala, hippocampus, basal ganglia, and thalamus), yielding a total of 122 ROIs. We averaged the blood-oxygen-level dependent (BOLD) time courses of all voxels in each ROI. For each dataset, the dynamic functional connectivity (FC) patterns were extracted from the ROI-based BOLD time courses using a tapered sliding window approach, where the Fisher's z-transformed Pearson's correlation between the BOLD time courses of every pair of ROIs were computed within each tapered sliding window (Window size - *Sherlock*: 30 TRs = 45s; *FNL*: 23 TRs = 46s). Hyperparameters of the sliding window for brain data were the same as those for behavioral data (Step size: 1TR, Gaussian kernel σ = 3TR).

Within-dataset FC prediction. Within each dataset, we built separate CPMs to predict valence or arousal time courses using leave-one-subject-out cross-validation. In each round of cross-validation, we conducted feature selection, where functional connections (FC) that significantly correlated with affect time course in the training set of N - 1 participants were selected (one-sample Student's t test, $p < .01$). A non-linear support vector regression model (python package svm.SVR from sklearn, https://scikit-learn.org/, kernel = "rbf") was trained in

the set of *N* - 1 participants to predict the group-average behavioral time course from the selected FC features, and tested on the held-out participant. The predictive accuracy of each round of cross-validation was calculated as the Pearson's correlation between the predicted and ground-truth arousal time course. Correlation coefficients of all rounds of cross-validation were Fisher's z-transformed, averaged across all cross-validation folds, and transformed back to an average r-value as the final measure of model performance. To assess statistical significance, we compared model performance against null distributions generated by training and testing the model on phase-randomized behavioral ratings (1000 permutations). We assumed a one-tailed significance test, with p = (1 + number of null r values $\geq$ empirical r)/(1 + number of permutations).

Across-dataset FC prediction. The set of FCs selected in every round of cross-validation in within-dataset prediction was used as the features in across-dataset prediction. All subjects' data in the training dataset (e.g., *Sherlock*) were used to train a non-linear SVM to learn the mapping between the patterns of selected FC and the group-average affect time course. The trained model was applied to the held-out dataset, to predict each subject's valence or arousal time course from their FC patterns. The predictive accuracy and statistical significance were evaluated in the same way as within-dataset FC prediction.

Arousal networks

For each dataset, we termed the set of functional connections that were selected in every round of cross-validation in predicting arousal as that dataset's arousal network. We defined FC features that positively, or negatively predict arousal as the high- or low-arousal networks, respectively. To examine the relationship between the arousal network and canonical macroscale functional networks in the brain, we grouped the 122 ROIs into 8 functional networks: the visual (VIS), somatosensory-motor (SOM), dorsal attention (DAN), ventral attention (VAN), limbic (LIM), frontoparietal control (FPN), default mode (DMN), and subcortical (SUB) networks. These networks were previously identified based on resting-state functional connectivity MRI of 1000 subjects[40]. The set of functional connections that occurred in both the *Sherlock* and *FNL* arousal networks were defined as the overlapping arousal network. Statistical significance of network overlap was assessed using a hypergeometric cumulative density function, which returns the probability of drawing up to x of K possible items in n drawings without replacement from an M-item population[42]. This was implemented in MATLAB as: p = 1 – hygecdf(x, M, K, n), where *x* refers to the number of overlapping edges, K and n refers to the number of edges in each of the two networks, and M refers to the total number of edges (7381).

We calculated the proportion of selected FCs relative to the total number of possible connections between each pair of functional networks. To statistically assess whether particular functional networks are represented in the arousal network more frequently than chance, we computed the proportion of selected FCs among all possible FCs between each pair of functional networks. To assess the significance of these proportions, we compared the proportions with a null distribution of 10000 permutations. We generated the null distribution by the following steps. First, we shuffled the positions of selected FCs (e.g., 593 for *Sherlock* high-arousal network) in the whole FC networks (i.e., 7381 FCs). Second, we divided the FC

networks into the 8 canonical macroscale functional networks. Third, we computed the proportions for every pair of functional networks following the same procedure. Finally, the significance of each proportion is computed by comparing the actual value with the null distribution of 10000 permutations using a one-tailed t-test (fdr-corrected).

Engagement behavioral ratings and networks

We acquired the engagement behavioral ratings of the *Sherlock* movie from Song et al., 2021 GitHub repository (https://github.com/hyssong/NarrativeEngagement). We averaged individual rating time courses across subjects, cropped the 52 TRs of the two cartoon segments (1-26 TR, 947-972 TR) from the group-average engagement rating time course, and z-normalized the rating time course. We computed the Pearson correlation between the average engagement and arousal rating time courses. The significance of the correlation was assessed by comparing the actual correlation with a null distribution, generated by circular shifting the average engagement rating time course 1000 times.

We trained a CPM on the engagement time course, and used the CPM to predict the arousal time course following a leave-one-subject-out cross validation procedure. The predictive accuracy in each run of cross-validation was calculated as the Pearson's correlation between model predictions and the ground-truth arousal time course. We then performed a paired *t*-test to test if model accuracy in predicting arousal was different between the CPM trained on engagement and the CPM trained on arousal.

To examine whether the neural representations of engagement and arousal shared underlying functional networks, we compared our arousal network with the engagement network acquired from Song et al., 2021. This previous work used leave-one-subject-out cross-validation to predict attentional engagement from dynamic functional connectivity in the *Sherlock* dataset. The *Sherlock* engagement FC network was the set of FCs selected in every round of cross-validation in predicting engagement. We computed the arousal-engagement overlap network by taking the overlapping FCs between the *Sherlock* arousal FC network (439 positive and 154 negative FC features) and the *Sherlock* engagement *FC* network (583 positive and 102 negative FC features), resulting in a network with 287 positive and 74 negative FC features.

Overlap arousal network predictions on two more fMRI datasets

To examine whether the overlap network between the *Sherlock* and *Friday Night Lights* arousal networks encoded a generalizable arousal representation, we trained a model on the overlap arousal network and tested on two more fMRI datasets, *North by Northwest* and *Merlin*.

The movie stimuli of the *North by Northwest* dataset was a 14:49-min long segment of the movie *North by Northwest*, an American spy thriller directed by Alfred Hitchcock. The fMRI data were collected on a 3T Philips Ingenia scanner at the MRI Research Center at the University of Chicago as part of an ongoing two-session study collecting task, movie, and annotated rest data. Only the data where participants watched the *North by Northwest* clip were analyzed here. Structural images were acquired using a high-resolution T1-weighted MPRAGE sequence (1.000 mm^3 resolution). Functional BOLD images were collected on a 3T Philips Ingenia

scanner with a 32-channel head coil. Functional scans were acquired during the movie segment in a single continuous run (951 TRs). The functional run included an extra 57 TRs after the video ends where participants were instructed to stare at the cross over a blank screen. These TRs were cropped from the end of each functional image to match the length of the movie stimuli. Data from 32 participants were included here.

The movie stimuli of the *Merlin* dataset was a 25-min long episode from BBC's *Merlin* (a British fantasy adventure drama series). The fMRI data was acquired from openneuro[45]. The fMRI preprocessing pipeline and dynamic functional connectivity analysis were the same as those of *Sherlock* and *Friday Night Lights*. Additional behavioral ratings on valence and arousal were collected on both *North by Northwest* and *Merlin* movie clips. The continuous rating paradigm was implemented in a single continuous run, separately for *North by Northwest* and *Merlin*. The same behavioral preprocessing and analysis pipeline as those of *Sherlock* and *Friday Night Lights* were performed for *North by Northwest* and *Merlin*.

The set of FCs in the overlap arousal network between the *Sherlock* arousal network and the *Friday Night Lights* arousal network was used as the input features in the current across-dataset prediction to test its generalizability. The data from both all subjects in *Sherlock* and all subjects in *Friday Night Lights* was used to train a non-linear SVM to learn the mapping between the patterns of selected FC and the group-average arousal time course. The trained model was applied separately to the *North by Northwest* and *Merlin* dataset, to predict each subject's arousal time course from their FC patterns. The predictive accuracy and statistical significance were evaluated in the same way as within-dataset FC prediction.

Examining out-of-sample generalizability of valence predictions with additional training data

To examine whether a CPM of valence could generalize with more training data, we trained a CPM on the combined data from *Sherlock* and *Friday Night Lights*. In the combined neural data, we selected the 2067 FCs that significantly correlated with valence as input features. We trained a non-linear support vector regression model to predict the valence time courses from the combined neural data of the selected FCs, and tested this trained model separately in the *North by Northwest* and *Merlin* datasets.

Defining positivity and negativity

To identify moments with positive valence (positivity), we took the raw behavioral valence ratings (range 1 - 25, 13 = neutral), averaged across participants rating the same movie, extracted the time segments where the group-averaged raw behavioral rating was above 13, and concatenated them into one single time course. We *z*-normalized and convolved the new rating time course with a HRF, and smoothed the new rating time course using a tapered sliding window (window size *Sherlock*: 30 TRs; *FNL*: 23 TRs). Moments with negative valence (negativity) were identified following the same procedure except that ratings that were below 13 were extracted from the group-average valence rating time course. The *Sherlock* movie had 1114 and 800 TRs of positivity and negativity respectively, and the *FNL* movie has 758 and 594 TRs of positivity and negativity respectively (**Fig. S5**).

To build predictive models to predict positivity and negativity from brain fMRI activity, we preprocessed the brain data to assess how they would be associated with the positivity and negativity rating time courses. We separately extracted the TRs corresponding to the behavioral positivity and negativity moments from the 122 parcels' activation magnitude time courses, and applied the same tapered sliding window approach as described above to extract the dynamic functional connectivity patterns.

Univariate generalized linear model

We ran generalized linear models (GLM) to identify the voxels whose BOLD activity were associated with valence or arousal in each movie. A Separate model was built for each of the four experimental conditions: *Sherlock*-valence, *FNL*-valence, *Sherlock*-arousal, and *FNL*-arousal. The GLMs were estimated throughout the whole brain using FSL/FEAT v.6.00.. Correction for multiple comparisons was performed using threshold-free cluster enhancement (TFCE correction) with an alpha level of 0.05, as implemented by the randomise tool in FSL[84]. To assess the overlap in statistical maps across movies, we performed a minimum statistic test compared to the global null (MS/GN) conjunction analysis[53] to identify the voxels in the brain whose activations were consistent across Sherlock and FSL ($p < .01$). The test was run separately for arousal and valence.

Multivariate-pattern based predictive modeling

To examine whether the multivariate patterns of activation encode valence and arousal, we employed the same analysis pipeline as that of dynamic CPM, except that the input features of the models were now the time courses of each voxel (Tapered sliding window size - *Sherlock*: 30 TRs = 45s; *FNL*: 23 TRs = 46s) following the approach described in Song et al.[31]. Hyperparameters of the sliding window for brain data were the same as those for behavioral data (Step size: 1TR, Gaussian kernel σ = 3TR).

Within-dataset prediction. Within each dataset, we built separate multivariate pattern-based predictive models (MPM) to separately predict valence or arousal time courses using a leave-one-subject-out cross-validation approach. In each round of cross-validation, we conducted feature selection, where voxels whose activity significantly correlated with the behavioral time course in the training set of N - 1 participants were selected (one-sample Student's t test, $p < .01$). The remaining procedures were identical to how we trained and tested the CPMs. The large number of voxels selected (24333 features) for each iteration posed a computational challenge for the permutation tests. To reduce computational cost, we ran 100 instead of 1000 permutations. We assumed a one-tailed significance test, with p = (1 + number of null r values $\geq$ empirical r)/(1 + number of permutations).

Across-dataset prediction. The set of voxels selected in every round of cross-validation in within-dataset prediction was used as the features in across-dataset prediction. All subjects' data in the training dataset (e.g., *Sherlock*) were used to train a non-linear SVM to learn the mapping between the patterns of selected voxels and the group-average affect time course. The trained model was applied to the held-out dataset, to predict each subject's affect time course

from their multivariate patterns. The predictive accuracy and statistical significance were evaluated in the same way as within-dataset predictions, except that the number of permutations to generate the null distribution was 1000.

Acknowledgements

We thank Janice Chen and colleagues for open sourcing the *Sherlock* dataset, Luke Chang and colleagues for open sourcing the *Friday Night Lights* dataset, and Asieh Zadbood and colleagues for open sourcing the *Merlin* dataset. We thank the MRI Research Center at the University of Chicago for assisting with the collection of the *North by Northwest* dataset. We thank Emily Finn for sharing code on calculating optical flow. We thank Yizhou (Louisa) Lyu for collecting part of the behavioral rating data. We thank members of the Bainbridge, Leong, Rosenberg, Bakkour (BLRB) community, especially members of the Motivation and Cognition Neuroscience Lab and Cognition, Attention, and Brain Lab at the University of Chicago for helpful feedback.

Our work was supported by National Science Foundation BCS-2043740 (M.D.R.), the University of Chicago Social Sciences Division, the University of Chicago Neuroscience Institute, and resources provided by the University of Chicago Research Computing Center.

References

1. Fredrickson, B. L. & Branigan, C. Positive emotions broaden the scope of attention and thought-action repertoires. *Cogn. Emot.* **19**, 313–332 (2005).
2. Phelps, E. A., Ling, S. & Carrasco, M. Emotion Facilitates Perception and Potentiates the Perceptual Benefits of Attention. *Psychol. Sci.* **17**, 292–299 (2006).
3. Kensinger, E. A. & Corkin, S. Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? *Mem. Cognit.* **31**, 1169–1180 (2003).
4. Kensinger, E. A. & Corkin, S. Two routes to emotional memory: Distinct neural processes for valence and arousal. *Proc. Natl. Acad. Sci.* **101**, 3310–3315 (2004).
5. Bechara, A., Damasio, H., Tranel, D. & Damasio, A. R. Deciding Advantageously Before Knowing the Advantageous Strategy. *Science* **275**, 1293–1295 (1997).
6. Heffner, J., Son, J.-Y. & FeldmanHall, O. Emotion prediction errors guide socially adaptive behaviour. *Nat. Hum. Behav.* **5**, 1391–1401 (2021).
7. Lerner, J. S., Li, Y., Valdesolo, P. & Kassam, K. S. Emotion and decision making. *Annu. Rev. Psychol.* **66**, 799–823 (2015).
8. Barrett, L. F. & Russell, J. A. The Structure of Current Affect: Controversies and Emerging Consensus. *Curr. Dir. Psychol. Sci.* **8**, 10–14 (1999).
9. Russell, J. A. A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**, 1161–1178 (1980).
10. Barrett, L. F. & Bliss-Moreau, E. Chapter 4 Affect as a Psychological Primitive. in *Advances in Experimental Social Psychology* vol. 41 167–218 (Academic Press, 2009).
11. Lindquist, K. A. Emotions Emerge from More Basic Psychological Ingredients: A Modern Psychological Constructionist Model. *Emot. Rev.* **5**, 356–368 (2013).
12. Posner, J., Russell, J. A. & Peterson, B. S. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev. Psychopathol.* **17**, (2005).
13. Schachter, S. & Singer, J. E. Cognitive, social, and physiological determinants of emotional state. *Psychol. Rev.* **69**, 379–399 (1962).
14. Dutton, D. G. & Aron, A. P. Some evidence for heightened sexual attraction under conditions of high anxiety. *J. Pers. Soc. Psychol.* **30**, 510–517 (1974).
15. Savitsky, K., Medvec, V. H., Charlton, A. E. & Gilovich, T. 'What, me worry?' Arousal, misattribution and the effect of temporal distance on confidence. *Pers. Soc. Psychol. Bull.* **24**, 529–536 (1998).
16. Anders, S., Lotze, M., Erb, M., Grodd, W. & Birbaumer, N. Brain activity underlying emotional valence and arousal: A response-related fMRI study. *Hum. Brain Mapp.* **23**, 200–209 (2004).
17. Anderson, A. K. *et al.* Dissociated neural representations of intensity and valence in human olfaction. *Nat. Neurosci.* **6**, 196–202 (2003).

18. Baucom, L. B., Wedell, D. H., Wang, J., Blitzer, D. N. & Shinkareva, S. V. Decoding the neural representation of affective states. *NeuroImage* **59**, 718–727 (2012).
19. Lewis, P., Critchley, H., Rotshtein, P. & Dolan, R. Neural Correlates of Processing Valence and Arousal in Affective Words. *Cereb. Cortex* **17**, 742–748 (2007).
20. Wilson-Mendenhall, C. D., Barrett, L. F. & Barsalou, L. W. Neural Evidence That Human Emotions Share Core Affective Properties. *Psychol. Sci.* **24**, 947–956 (2013).
21. Kim, J. *et al.* A study in affect: Predicting valence from fMRI data. *Neuropsychologia* **143**, 107473 (2020).
22. Kim, J., Shinkareva, S. V. & Wedell, D. H. Representations of modality-general valence for videos and music derived from fMRI data. *NeuroImage* **148**, 42–54 (2017).
23. Young, C. B. *et al.* Dynamic Shifts in Large-Scale Brain Network Balance As a Function of Arousal. *J. Neurosci.* **37**, 281–290 (2017).
24. Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A. & Wager, T. D. A Sensitive and Specific Neural Signature for Picture-Induced Negative Affect. *PLOS Biol.* **13**, e1002180 (2015).
25. Wang, Y., Kragel, P. A. & Satpute, A. B. *Neural predictors of subjective fear depend on the situation*. http://biorxiv.org/lookup/doi/10.1101/2022.10.20.513114 (2022) doi:10.1101/2022.10.20.513114.
26. Kragel, P. A. & LaBar, K. S. Multivariate neural biomarkers of emotional states are categorically distinct. *Soc. Cogn. Affect. Neurosci.* **10**, 1437–1448 (2015).
27. Wager, T. D. *et al.* A Bayesian Model of Category-Specific Emotional Brain Responses. *PLOS Comput. Biol.* **11**, e1004066 (2015).
28. Shen, X. *et al.* Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nat. Protoc.* **12**, 506–518 (2017).
29. Barrett, L. F. & Satpute, A. B. Historical pitfalls and new directions in the neuroscience of emotion. *Neurosci. Lett.* **693**, 9–18 (2019).
30. Rosenberg, M. D. *et al.* A neuromarker of sustained attention from whole-brain functional connectivity. *Nat. Neurosci.* **19**, 165–171 (2016).
31. Song, H., Finn, E. S. & Rosenberg, M. D. Neural signatures of attentional engagement during narratives and its consequences for event memory. *Proc. Natl. Acad. Sci.* **118**, e2021905118 (2021).
32. Avery, E. W. *et al.* Distributed Patterns of Functional Connectivity Predict Working Memory Performance in Novel Healthy and Memory-impaired Individuals. *J. Cogn. Neurosci.* **32**, 241–255 (2020).
33. Keerativittayayut, R., Aoki, R., Sarabi, M. T., Jimura, K. & Nakahara, K. Large-scale network integration in the human brain tracks temporal fluctuations in memory encoding performance. *eLife* **7**, e32696 (2018).
34. Song, H., Park, B., Park, H. & Shim, W. M. Cognitive and Neural State Dynamics of

Narrative Comprehension. *J. Neurosci.* **41**, 8972–8990 (2021).

35. Chen, J. *et al.* Shared memories reveal shared structure in neural activity across individuals. *Nat. Neurosci.* **20**, 115–125 (2017).
36. Chang, L. J. *et al.* Endogenous variation in ventromedial prefrontal cortex state dynamics during naturalistic viewing reflects affective experience. *Sci. Adv.* **7**, eabf7129 (2021).
37. Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of Best Practices for Evidence for Prediction: A Review. *JAMA Psychiatry* **77**, 534–540 (2020).
38. Hutcherson, C. A. *et al.* Attention and emotion: Does rating emotion alter neural responses to amusing and sad films? *NeuroImage* **27**, 656–668 (2005).
39. Chen, G. *et al.* Untangling the relatedness among correlations, part I: Nonparametric approaches to inter-subject correlation analysis at the group level. *NeuroImage* **142**, 248–259 (2016).
40. Thomas Yeo, B. T. *et al.* The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 1125–1165 (2011).
41. Fan, L. *et al.* The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cereb. Cortex N. Y. N 1991* **26**, 3508–3526 (2016).
42. Rosenberg, M. D. *et al.* Methylphenidate Modulates Functional Network Connectivity to Enhance Attention. *J. Neurosci.* **36**, 9547–9557 (2016).
43. Aston-Jones, G. & Cohen, J. D. An Integrative Theory Of Locus Coeruleus-Norepinephrine Function: Adaptive Gain and Optimal Performance. *Annu. Rev. Neurosci.* **28**, 403–450 (2005).
44. Hopstaken, J. F., van der Linden, D., Bakker, A. B. & Kompier, M. A. J. A multifaceted investigation of the link between mental fatigue and task disengagement. *Psychophysiology* **52**, 305–315 (2015).
45. Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A. & Hasson, U. How We Transmit Memories to Other Brains: Constructing Shared Neural Representations Via Communication. *Cereb. Cortex* **27**, 4988–5000 (2017).
46. Lakens, D. Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Soc. Psychol. Personal. Sci.* **8**, 355–362 (2017).
47. Lakens, D., Scheel, A. M. & Isager, P. M. Equivalence Testing for Psychological Research: A Tutorial. *Adv. Methods Pract. Psychol. Sci.* **1**, 259–269 (2018).
48. Goertzen, J. R. & Cribbie, R. A. Detecting a lack of association: An equivalence testing approach. *Br. J. Math. Stat. Psychol.* **63**, 527–537 (2010).
49. Diener, E. & Iran-Nejad, A. The relationship in experience between various types of affect. *J. Pers. Soc. Psychol.* **50**, 1031–1038 (1986).
50. Larsen, J. T., Peter McGraw, A., Mellers, B. A. & Cacioppo, J. T. The Agony of Victory and Thrill of Defeat: Mixed Emotional Reactions to Disappointing Wins and Relieving Losses. *Psychol. Sci.* **15**, 325–330 (2004).

51. O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J. & Andrews, C. Abstract reward and punishment representations in the human orbitofrontal cortex. *Nat. Neurosci.* **4**, 95–102 (2001).

52. Seymour, B., Daw, N., Dayan, P., Singer, T. & Dolan, R. Differential Encoding of Losses and Gains in the Human Striatum. *J. Neurosci.* **27**, 4826–4831 (2007).

53. Nichols, T., Brett, M., Andersson, J., Wager, T. & Poline, J.-B. Valid conjunction inference with the minimum statistic. *NeuroImage* **25**, 653–660 (2005).

54. Yu, H. *et al.* A Generalizable Multivariate Brain Pattern for Interpersonal Guilt. *Cereb. Cortex* **30**, 3558–3572 (2020).

55. Woo, C.-W. & Wager, T. D. Neuroimaging-based biomarker discovery and validation. *Pain* **156**, 1379–1381 (2015).

56. Kragel, P. A., Koban, L., Barrett, L. F. & Wager, T. D. Representation, Pattern Information, and Brain Signatures: From Neurons to Neuroimaging. *Neuron* **99**, 257–273 (2018).

57. Cotton, J. L. A review of research on Schachter's theory of emotion and the misattribution of Arousal. *Eur. J. Soc. Psychol.* **11**, 365–397 (1981).

58. Marin, M. M., Schober, R., Gingras, B. & Leder, H. Misattribution of musical arousal increases sexual attraction towards opposite-sex faces in females. *PLOS ONE* **12**, e0183531 (2017).

59. Vosgerau, J. How Prevalent Is Wishful Thinking? Misattribution of Arousal Causes Optimism and Pessimism in Subjective Probabilities. *J. Exp. Psychol. Gen.* **139**, 32–48 (2010).

60. Harvey, J. H., Ickes, W. J. & Kidd, R. F. *New Directions in Attribution Research: Volume 1*. (Psychology Press, 2018).

61. Lindquist, K. A., Siegel, E. H., Quigley, K. S. & Barrett, L. F. The hundred-year emotion war: Are emotions natural kinds or psychological constructions? Comment on Lench, Flores, and Bench (2011). *Psychol. Bull.* **139**, 255–263 (2013).

62. Barrett, L. F. The theory of constructed emotion: an active inference account of interoception and categorization. *Soc. Cogn. Affect. Neurosci.* nsw154 (2016) doi:10.1093/scan/nsw154.

63. Barrett, L. F. *How emotions are made: The secret life of the brain*. xv, 425 (Houghton Mifflin Harcourt, 2017).

64. Lewis, P., Critchley, H., Rotshtein, P. & Dolan, R. Neural Correlates of Processing Valence and Arousal in Affective Words. *Cereb. Cortex* **17**, 742–748 (2006).

65. Chang, C. *et al.* Tracking brain arousal fluctuations with fMRI. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 4518–4523 (2016).

66. Goodale, S. E. *et al.* fMRI-based detection of alertness predicts behavioral response variability. *eLife* **10**, e62376 (2021).

67. Shine, J. M. *et al.* The Dynamics of Functional Brain Networks: Integrated Network States during Cognitive Task Performance. *Neuron* **92**, 544–554 (2016).

68. Lee, K. *et al.* Arousal impacts distributed hubs modulating the integration of brain functional connectivity. *NeuroImage* **258**, 119364 (2022).

69. Varela, F., Lachaux, J.-P., Rodriguez, E. & Martinerie, J. The brainweb: Phase synchronization and large-scale integration. *Nat. Rev. Neurosci.* **2**, 229–239 (2001).

70. Deco, G., Tononi, G., Boly, M. & Kringelbach, M. L. Rethinking segregation and integration: contributions of whole-brain modelling. *Nat. Rev. Neurosci.* **16**, 430–439 (2015).

71. Aliko, S., Huang, J., Gheorghiu, F., Meliss, S. & Skipper, J. I. A naturalistic neuroimaging database for understanding the brain using ecological stimuli. *Sci. Data* **7**, 347 (2020).

72. Visconti di Oleggio Castello, M., Chauhan, V., Jiahui, G. & Gobbini, M. I. An fMRI dataset in response to “The Grand Budapest Hotel”, a socially-rich, naturalistic movie. *Sci. Data* **7**, 383 (2020).

73. Hanke, M. *et al.* A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. *Sci. Data* **1**, 140003 (2014).

74. Gordon, E. M. *et al.* Precision Functional Mapping of Individual Human Brains. *Neuron* **95**, 791-807.e7 (2017).

75. Barrett, L. F. & Bliss‐Moreau, E. Chapter 4 Affect as a Psychological Primitive. in *Advances in Experimental Social Psychology* vol. 41 167–218 (Academic Press, 2009).

76. Murphy, F. C., Nimmo-Smith, I. & Lawrence, A. D. Functional neuroanatomy of emotions: a meta-analysis. *Cogn. Affect. Behav. Neurosci.* **3**, 207–233 (2003).

77. Liu, X., Hairston, J., Schrier, M. & Fan, J. Common and distinct networks underlying reward valence and processing stages: a meta-analysis of functional neuroimaging studies. *Neurosci. Biobehav. Rev.* **35**, 1219–1236 (2011).

78. Lindquist, K. A. & Barrett, L. F. A functional architecture of the human brain: Emerging insights from the science of emotion. *Trends Cogn. Sci.* **16**, 533–540 (2012).

79. Lindquist, K. A., Satpute, A. B., Wager, T. D., Weber, J. & Barrett, L. F. The Brain Basis of Positive and Negative Affect: Evidence from a Meta-Analysis of the Human Neuroimaging Literature. *Cereb. Cortex* **26**, 1910–1922 (2016).

80. Moors, A., Ellsworth, P. C., Scherer, K. R. & Frijda, N. H. Appraisal Theories of Emotion: State of the Art and Future Development. *Emot. Rev.* **5**, 119–124 (2013).

81. Adolphs, R. & Andler, D. Investigating Emotions as Functional States Distinct From Feelings. *Emot. Rev.* **10**, 191–201 (2018).

82. Lee Masson, H. & Isik, L. Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. *NeuroImage* **245**, 118741 (2021).

83. Grall, C. & Finn, E. S. Leveraging the power of media to drive cognition: A media-informed approach to naturalistic neuroscience. (2021).

84. Smith, S. M. & Nichols, T. E. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage* **44**, 83–98 (2009).