nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-022-01310-0

Check for updates

Decoding naturalistic affective behaviour from spectro-spatial features in multiday human iEEG

Maryam Bijanzadeh[1], Ankit N. Khambhati[1], Maansi Desai[2], Deanna L. Wallace[3], Alia Shafi[1], Heather E. Dawes[1], Virginia E. Sturm[4] and Edward F. Chang[1]✉

The neurological basis of affective behaviours in everyday life is not well understood. We obtained continuous intracranial electroencephalography recordings from the human mesolimbic network in 11 participants with epilepsy and hand-annotated spontaneous behaviours from 116 h of multiday video recordings. In individual participants, binary random forest models decoded affective behaviours from neutral behaviours with up to 93% accuracy. Both positive and negative affective behaviours were associated with increased high-frequency and decreased low-frequency activity across the mesolimbic network. The insula, amygdala, hippocampus and anterior cingulate cortex made stronger contributions to affective behaviours than the orbitofrontal cortex, but the insula and anterior cingulate cortex were most critical for differentiating behaviours with observable affect from those without. In a subset of participants ($N = 3$), multiclass decoders distinguished amongst the positive, negative and neutral behaviours. These results suggest that spectro-spatial features of brain activity in the mesolimbic network are associated with affective behaviours of everyday life.

The outward expression of internal feeling states, affective behaviours play an integral role in everyday human life. Functional magnetic resonance imaging (fMRI) and scalp-based electroencephalography (EEG) studies have used task-based paradigms to reveal a distributed neural network that supports the generation of emotions and their accompanying affective behaviours. The insula and anterior cingulate cortex (ACC) are tightly connected structures within a mesolimbic network that are critical for producing and sensing the motor and autonomic changes that arise during emotions. Activity in the mesolimbic network increases as affective experience intensifies but decreases with engagement of emotion regulation systems anchored by orbitofrontal cortex (OFC)[1–3]. Engagement of emotion regulation systems allows individuals to control their feelings and reduces activity in emotion-generating structures such as the amygdala[1,4,5]. Whilst some previous neuroimaging studies have indicated that certain regions in the mesolimbic network play predominant roles in specific emotions (for example, insula in disgust[6–10], subgenual ACC in sadness, amygdala in fear[11,12] and ventral striatum in joy[13]), there is also evidence that the insula and ACC, together with the mesolimbic network, coactivate during a wide range of affective states[14]. Electrical stimulation of these structures offers convergent evidence that activation or deactivation of distinct mesolimbic network nodes can alter emotions, mood and behaviour.

EEG provides additional insights into mesolimbic network functioning in affective contexts. EEG studies, which quantify neural activity on faster timescales than fMRI and in different frequency bands, have demonstrated that mesolimbic structures exhibit rapid responses (<300 ms) in local field potentials to emotional faces[11,15–19] and evocative images[20]. These studies offer some evidence that different affective reactions are accompanied by distinct patterns of spatial activity in the mesolimbic network, with certain structures playing more prominent roles in some affective states than in others. Images that elicit negative emotions, for example, increase high gamma band activity in the amygdala more than in other mesolimbic regions[11]. There is mixed evidence, however, as to whether different spectral patterns across the mesolimbic network differentiate amongst affective states. Whilst some studies have found that increased high gamma band activity in mesolimbic structures characterizes both positive[11,21,22] and negative emotions[11,15,21,23,24], others have found that increased activity in lower frequency bands (for example, theta and alpha) may be a distinguishing feature of positive emotions[25].

Although prior fMRI and EEG studies have helped to elucidate the role of the mesolimbic network during emotion-relevant, task-based paradigms, methodological constraints have limited investigations in more ecologically valid contexts. Little is known, therefore, about how the brain produces the affective behaviours that arise amidst the ups and downs of everyday life. Here, we obtained multiday video recordings of participants undergoing surgery for intractable epilepsy who had intracranial EEG (iEEG) electrodes[26,27] implanted in the mesolimbic network. Participants' spontaneous affective behaviours ('naturalistic affective behaviours') were hand-annotated from the video recordings of their hospital stay and used to probe attendant neural activity patterns. We tested whether binary models could decode positive and negative affective behaviours from those lacking clear affect ('neutral behaviours') from the iEEG recordings, and we then examined which mesolimbic features influenced the decoders' performance. In a subset of participants ($N = 3$), we also tested whether multiclass decoders could distinguish amongst all three behaviour types.

Our central hypothesis was that power changes in specific frequency bands (that is, spectral features) in specific network hubs (that is, spatial features) would together create 'spectro-spatial' patterns across the mesolimbic network and distinguish amongst different types of naturalistic behaviour. Positive and negative affective

[1]Department of Neurological Surgery, University of California San Francisco, San Francisco, CA, USA. [2]Department of Communication Sciences and Disorders, Moody College of Communication, University of Texas at Austin, Austin, TX, USA. [3]Department of Mechanical Engineering, Psychology and Neurology, University of Texas at Austin, Austin, TX, USA. [4]Department of Neurology, UCSF Weill Institute for Neurosciences, University of California San Francisco, San Francisco, CA, USA. ✉e-mail: Edward.Chang@ucsf.edu

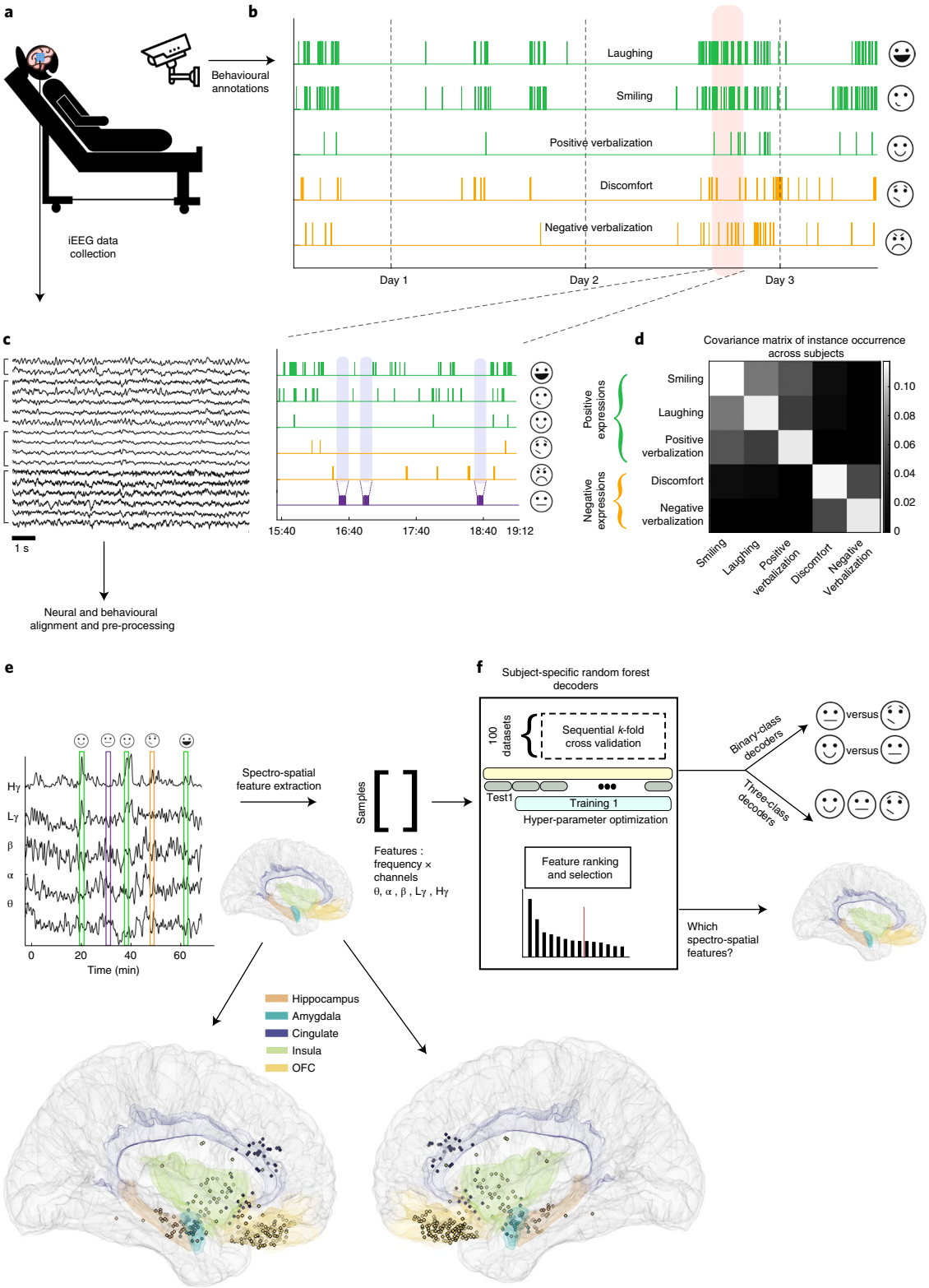


Fig. 1 | Collection and processing pipelines for the behavioural and neural data streams. a, Schematic of an example participant who underwent continuous neural and video recordings during a multiday hospital stay. **b**, Video recordings were hand-annotated to identify instances of positive affective behaviours (green), negative affective behaviours (orange) and neutral behaviours. Inset: zoom of a 3-h period (orange shading) to illustrate examples of neutral behaviours (purple shading). **c**, Example 10-s raw iEEG data traces from four regions. **d**, Covariance matrix of occurrences of affective behaviours across participants. **e**, Magnitude of Hilbert transform in five frequency bands from an insula channel across 60 min overlaid on instances of affective behaviours with right and left hemisphere views of the MNI template brain (bottom) to show the verified electrode coverage of mesolimbic structures across the sample. **f**, Pipeline for training the RF decoder models.

behaviours, by definition, differ in valence, but they can also have similar qualities such as comparable levels of arousal (that is, intensity), which may be represented by shared network changes. Thus, we expected that, though common spectral changes in mesolimbic structures might characterize both positive and negative affective behaviours, each behavioural class would also have spectro-spatial features that make it unique. As gamma band activity is thought to reflect neuronal activity in humans[28], we hypothesized that both positive and negative affective behaviours would be characterized by increased gamma activity. We further hypothesized that gamma activity in the insula and ACC, regions that facilitate emotions[13,29], would contribute more strongly to the production of affective behaviours than gamma activity in the OFC, a region that often inhibits emotions[19,30–32]. We anticipated that, whereas distributed network-level spectral changes may characterize affective behaviours in general, spatial differences may serve to differentiate between positive and negative affective behaviours given that some regions play prominent roles in certain emotions. Given that stimulation of the ventral ACC can induce positive behaviours and feelings such as laughter and mirth[10,12,33], and stimulation of the dorsal ACC and amygdala can produce feelings of fear and doom[34], we expected that the ventral ACC might contribute more to positive affective behaviours whereas the dorsal ACC and amygdala might participate more in negative affective behaviours[15,25,35].

Results

We obtained 24-h bedside audiovisual recordings and continuous iEEG data in participants with intractable epilepsy. Participants were hospitalized for clinical seizure monitoring and had temporary implanted subdural electrodes (Fig. 1a, Table 1 in Supplementary Information). To examine the neural mechanisms underlying naturalistic affective behaviours, we analysed a total of 116 h (mean 10.5 h, s.d. 5.48 h) of behavioural (Fig. 1b) and neural data (Fig. 1c) in 11 participants with electrodes placed in at least three mesolimbic structures, which here included the insula, ACC, OFC, amygdala and hippocampus (Supplementary Table 1). Although participants had extensive coverage across the mesolimbic network, electrode placement was based on each participant's clinical needs and, thus, varied somewhat across individuals.

Eleven human raters annotated participants' spontaneous behaviours in the video recordings (Fig. 1b, Supplementary Information, Tables 2 and 3 and Extended Data Fig. 1A). As there was a positive correlation amongst the total number of smiling, laughing and positive verbalization instances in each participant, these behaviours were aggregated as 'positive affective behaviours'. There was also a positive correlation between the total number of pain–discomfort and negative verbalization instances, so these behaviours were aggregated as 'negative affective behaviours' (Fig. 1d). We defined 'neutral behaviours' as periods in which there were neither positive nor negative affective behaviours for at least 10 min (Fig. 1b, bottom panel, purple shading). These periods were often characterized by other behaviours lacking clear affect such as eating, sleeping or conversing (Supplementary Table 3). Although it can be argued that no activities are truly affectively neutral[35], these behaviours offered a rigorous control condition against which to compare the affective behaviours because, unlike moments of rest, they included behaviours with varying levels of engagement and movement.

After aligning the neural and behavioural data (Fig. 1e and Extended Data Fig. 2), we extracted the spectral power in conventional EEG frequency bands (Methods) from electrodes in mesolimbic structures. We computed the average power in each frequency band (that is, the spectral features) for each electrode (that is, the spatial features; Fig. 1e, bottom), using 10-s non-overlapping windows centred on each positive, negative or neutral behaviour. Together, we refer to these as the 'spectro-spatial features'.

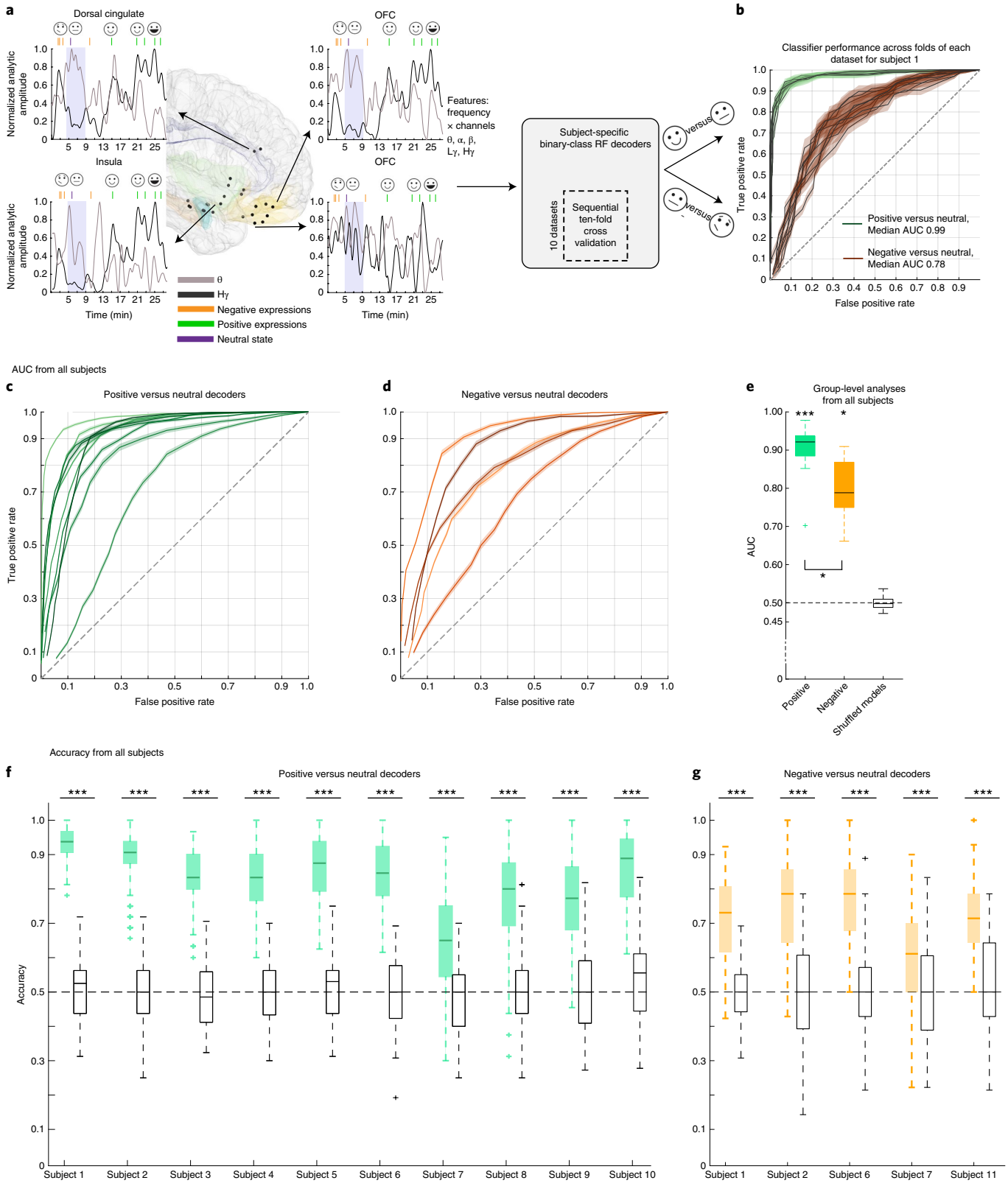
Personalized random forest (RF) models decoded affective behaviours. We first trained binary decoders (Fig. 1f) to determine whether we could distinguish affective behaviours from neutral behaviours in each participant. The goal of the positive decoder was to distinguish positive affective behaviours from neutral behaviours ($n = 10$ participants), and the goal of the negative decoder was to distinguish negative affective behaviours from neutral behaviours ($n = 5$ participants). For each participant, we constructed RF models and trained them on the spectro-spatial features for the positive, negative and neutral behaviours (Fig. 2a,b).

At the individual level, the spectro-spatial features of the mesolimbic network discriminated positive affective behaviours (10/10 participants, Fig. 2c) and negative affective behaviours (5/5 participants, Fig. 2d) from neutral behaviours significantly better than chance. The group-level results replicated the successful performance of the positive and negative decoders at the individual level (mean ± s.e.m. area under the curve (AUC) 0.90 ± 0.02, $n = 10$, Wilcoxon rank-sum test, $P < 0.001$; 0.80 ± 0.04, $n = 5$, $P = 0.0012$; Fig. 2e). Similar findings were also obtained using accuracy measures (number of true predicted samples/all samples) across all participants (Fig. 2f,g). A comparison of decoding performance revealed that the positive decoders performed significantly better than the negative decoders (Wilcoxon rank-sum test, $P = 0.04$; Fig. 2e).

As the periods of neutral behaviour included activities with varying levels of arousal, engagement and movement, we conducted two additional analyses to investigate whether these factors influenced our results. First, we removed periods in which participants engaged in miscellaneous activities (for example, sleep/eye closure, conversation, eating, drinking, etc.) and selected a subset of moments in

Fig. 2 | Within-subject RF models decoded positive and negative affective behaviours from neutral behaviours. a, Locations of the leads (black dots) and spectral features used in the decoder models for an example participant illustrated using the MNI template brain (Electrode localization section). Insets indicate 27 min of high gamma (black) and theta (grey) analytic amplitudes for four example channels, aligned with the affective behaviours in green and orange for the example participant; purple shadings show neutral periods. The analytic amplitudes were averaged using a 10-s non-overlapping window and then convolved by a Gaussian with an s.d. of 20 s. **b**, Receiver operating characteristic curve for the example participant across ten datasets (neutral behaviours were selected from different recording times to reduce selection bias) for positive decoders (green) and negative decoders (orange). The shadings represent the s.e.m. across ten folds. **c**,**d**, Area under the curve (AUC) for all ten participants and for five participants on which the positive (**c**) and negative decoders (**d**) were trained. Each solid line represents one participant. The shadings show the s.e.m. across all 100 datasets. **e**, Distribution of average AUC for positive (green, $n = 10$ participants), negative (orange, $n = 5$ participants) and permuted models (black, $n = 15$ from both positive and negative) trained in the same way using the shuffled labels across all participants (with a significance level of 0.05 since the average of 100 runs from each participant is included). The AUCs of the positive and negative decoders were significantly different from the shuffled models ($P = 0.00003$ and $P = 0.0012$, respectively). The positive decoders reached a larger AUC than the negative decoders ($P = 0.04$). **f**,**g**, Accuracies of the same models as in **c** and **d**. Accuracy of all $n = 100$ RF models were significantly different from 100 permuted models for all participants ($P < 0.0001$; Supplementary Tables 10 and 11). In the box plots **e**–**g**, central lines represent the median whilst the two edges represent 25 and 75 percentiles; whiskers show the most extreme data points, and outliers are shown individually (see MATLAB boxplot function). All statistics reported in **e**–**g** are from two-sided pairwise rank-sum test, and asterisks show significance levels of $^{***}P < 0.0001$ and $^{*}P < 0.05$.

which no behaviours were annotated (Supplementary Table 3). These periods represented a quiet yet alert resting state ('rest') that was presumably characterized by lower arousal and lower movement than the broader neutral behaviour category. Binary RF decoders, trained as above in each participant, successfully decoded positive affective behaviours ($n=9$ participants) and negative affective behaviours ($n=5$ participants) from rest using the spectro-spatial features (Extended Data Fig. 3). Despite some participant-level variability, when examined across the sample, there was no significant difference between the mean AUC from the binary decoders comparing affective behaviours with neutral behaviours and the mean AUC from the binary decoders comparing affective



behaviours with rest. Next, we limited our analyses to examine whether the decoders could distinguish between affective and neutral behaviours only during conversations. As most affective behaviours arose during periods of conversation (mean across all participants 75%, s.d. 19%, $n = 11$ participants; Supplementary Table 4), comparing affective and neutral behaviours in this context was a rigorous test of our results because participants' engagement and movement (speech, gesture, etc.) were likely comparable between the affective and neutral moments of the conversations. The binary decoders again successfully distinguished positive and negative affective behaviours from neutral behaviours during conversation (Extended Data Fig. 4). Taken together, these additional analyses suggest that our primary results withstood additional behavioural contrasts and were, thus, unlikely to be explained solely by variability in arousal, engagement or movement across the behavioural classes.

Shared spectral changes discriminated affective behaviours. To identify the features that enabled the RF decoders to discriminate affective behaviours from neutral behaviours in each participant, we used the trained decoder models to rank the spectro-spatial features that best discriminated positive and negative affective behaviours from neutral behaviours at the individual level. This unbiased feature selection approach (Feature selection section and Supplementary Information) revealed that high gamma activity and low-frequency activity across multiple mesolimbic structures contributed to the successful decoding of positive affective behaviours from neutral behaviours (Fig. 3a,b). A similar spectral pattern emerged when we examined the features that decoded negative affective behaviours from neutral behaviours (Supplementary Fig. 1). These findings confirmed that a data-driven approach could uncover personalized biomarkers that distinguished both positive and negative affective from neutral behaviours.

Next, we selected the personalized spectro-spatial features from each decoder type (Supplementary Figs. 2 and 3) and investigated the preference of these features for the positive and negative affective behaviours. Additional evidence for the robustness of the selected neural features came from the three other statistical methods (Supplementary Figs. 4–7) that we used to replicate our results. Across the sample, positive affective behaviours were again characterized by increased power in the high and low gamma bands and decreased power in the low-frequency bands (for example, theta and beta) (Fig. 3c and Supplementary Table 5). Negative affective behaviours were also characterized by increased high gamma band activity and decreased low-frequency band activity (alpha and beta, Fig. 3d). To map the features in two affective spaces (that is, positive or negative affective behaviours versus neutral behaviours), we conducted hierarchical clustering in each participant (Clustering section and Supplementary Fig. 8). These analyses also uncovered 'gamma' and 'low-frequency' clusters that distinguished positive and negative affective behaviours from neutral behaviours. We found similar results within each participant and across the sample (Fig. 3e,f and 'Clustering Analyses' in Supplementary Information).

Distinct spatial patterns characterized affective behaviours. Our results revealed a common spectral pattern—increased gamma activity and decreased low-frequency activity—across the mesolimbic network during affective behaviours (Supplementary Figs. 9 and 10) within each participant. We next asked whether, despite this distributed spectral pattern, certain brain regions within the mesolimbic network were more important than others to affective behaviours, in general, and to positive or negative affective behaviours, in particular. Consistent with our prior results that emerged when the data were analysed across the network at the individual level (Supplementary Figs. 9 and 10), changes in both the low-frequency and gamma clusters also characterized both types of affective behaviour when examined across the group (Fig. 4). These results again suggested that each region within the network participated in positive and negative affective behaviours.

We next conducted a more in-depth analysis to investigate whether certain regions made stronger contributions to the cluster-level results. Visualization of the median difference scores for the gamma and low-frequency clusters in each region indicated that spectral changes in some structures were more frequently selected than others for affective behaviours. Compared with neutral behaviours, positive affective behaviours (Fig. 4b) were more often characterized by increased gamma activity than decreased low-frequency activity in certain regions (that is, ventral ACC, hippocampus and dorsal ACC; Extended Data Fig. 5E). The spectral pattern in other regions (such as OFC) during positive affective behaviours was more complex, however, as changes in both clusters contributed to these behaviours (Fig. 4a,b and Extended Data Fig. 5E). When we compared negative affective behaviours with neutral behaviours, increased gamma activity in the amygdala was the most notable distinguishing feature (Fig. 4c,d), but low-frequency activity in the amygdala did not contribute to negative affective behaviours. The spectral changes in other regions (that is, hippocampus, insula, ventral ACC and dorsal ACC) showed a preference for the low-frequency cluster during negative affective behaviours (Extended Data Fig. 5F). These results indicated that certain regions within the mesolimbic network contributed more strongly to different types of affective behaviour when neural activity was quantified within specific frequency bands.

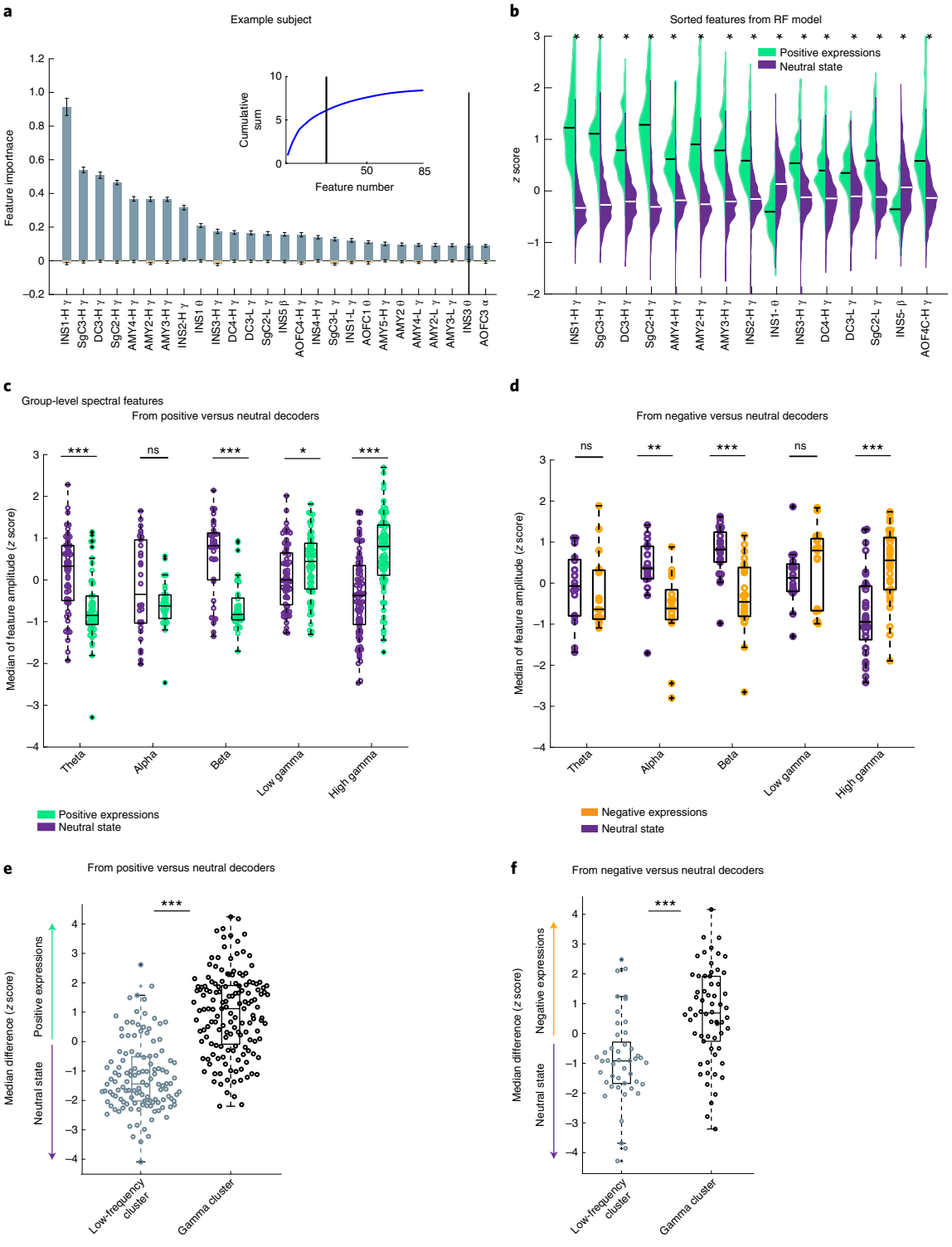
As a more rigorous test of these results, we re-trained the within-subject positive (Extended Data Fig. 6 and Supplementary Table 6) and negative (Extended Data Fig. 7 and Supplementary Table 7) decoders in each region, one at a time, leveraging all of its spectral features. Across the sample, widespread spectral changes in the insula (7/9 participants), amygdala (5/5 participants),

Fig. 3 | There was increased gamma band (low and high) activity during affective compared with neutral behaviours. a, Feature importance across $n = 100$ dataset/runs from the positive decoders for an example participant (subject 1), presented as mean ± s.e.m. The inset shows the cumulative summation curve of the average feature importance value across the runs; black vertical lines are the objective threshold used to select the top features. **b**, Sample distributions of the top 15 selected features for the positive affective behaviours (green) and neutral behaviours (purple). All sample distributions were significantly different from each other ($P < 0.0001$). **c**, Normalized median distributions of the positive affective behaviours and the neutral behaviours for selected features across the sample. The median values from the positive decoders were first normalized to the maximum absolute spectral amplitude across selected features at a within-subject level and then pooled across all participants ($n = 10$). The median values from the positive affective behaviours were significantly different from the neutral behaviours within theta ($n = 55$, $P = 9 \times 10^{-6}$), beta ($n = 37$, $P = 10^{-6}$), low gamma ($n = 65$, $P = 0.043$) and high gamma bands ($n = 86$, $P = 10^{-9}$). **d**, Normalized median distributions of negative affective behaviours and neutral behaviours for selected features (from five participants). The median values were significantly different within alpha ($n = 17$, $P = 0.0004$), beta ($n = 23$, $P = 6 \times 10^{-5}$) and high gamma ($n = 33$, $P = 10^{-5}$). **e**,**f**, Median difference score of the gamma cluster was selective to positive (**e**, $n = 149$) and negative affective behaviours (**f**, $n = 62$). The low-frequency cluster ($n = 124$ for positive, $n = 45$ for negative decoders) was significantly different from the gamma cluster for both positive (**e**, $P = 10^{-26}$) and negative decoders (**f**, $P = 3 \times 10^{-6}$). All pairwise statistical comparisons are based on non-parametric two-sided rank-sum test (**b**–**f**). In the box plots **c**–**f**, central lines represent the median and the two edges represent 25 and 75 percentiles; whiskers show the most extreme data points, and outliers are shown individually (see MATLAB boxplot function). INS, insula; SgC, subgenual cingulate; DC, dorsal cingulate; AMY, amygdala; H, high; L, low.

hippocampus (6/7 participants) and ventral ACC (4/4 participants) were more likely (>50% of participants) to discriminate positive affective behaviours from neutral behaviours than spectral changes in the OFC (4/9 participants) or dorsal ACC (4/10 participants; Extended Data Fig. 6). Spectral changes in the insula (4/4 participants), amygdala (1/1 participant), hippocampus (2/2 participants) and dorsal ACC (3/5 participants) were more likely than spectral changes in the ventral ACC (1/2 participants) or OFC (1/5 participants) to distinguish negative affective behaviours from neutral behaviours (Extended Data Fig. 7).

Insula and ACC were the most generalizable spatial features. To examine the generalizability of our within-subject results, we performed cross-subject decoding using a subset of the sample with iEEG electrodes implanted in the insula, OFC and dorsal ACC, the three most commonly sampled regions for the positive and negative decoders (Methods).

The positive decoders discriminated positive affective behaviours from neutral behaviours in 5/6 participants with a generalizability score (Methods) of 0.73 ± 0.13 where chance is 0.50 (Fig. 5a). We then used five spectral bands within each region and



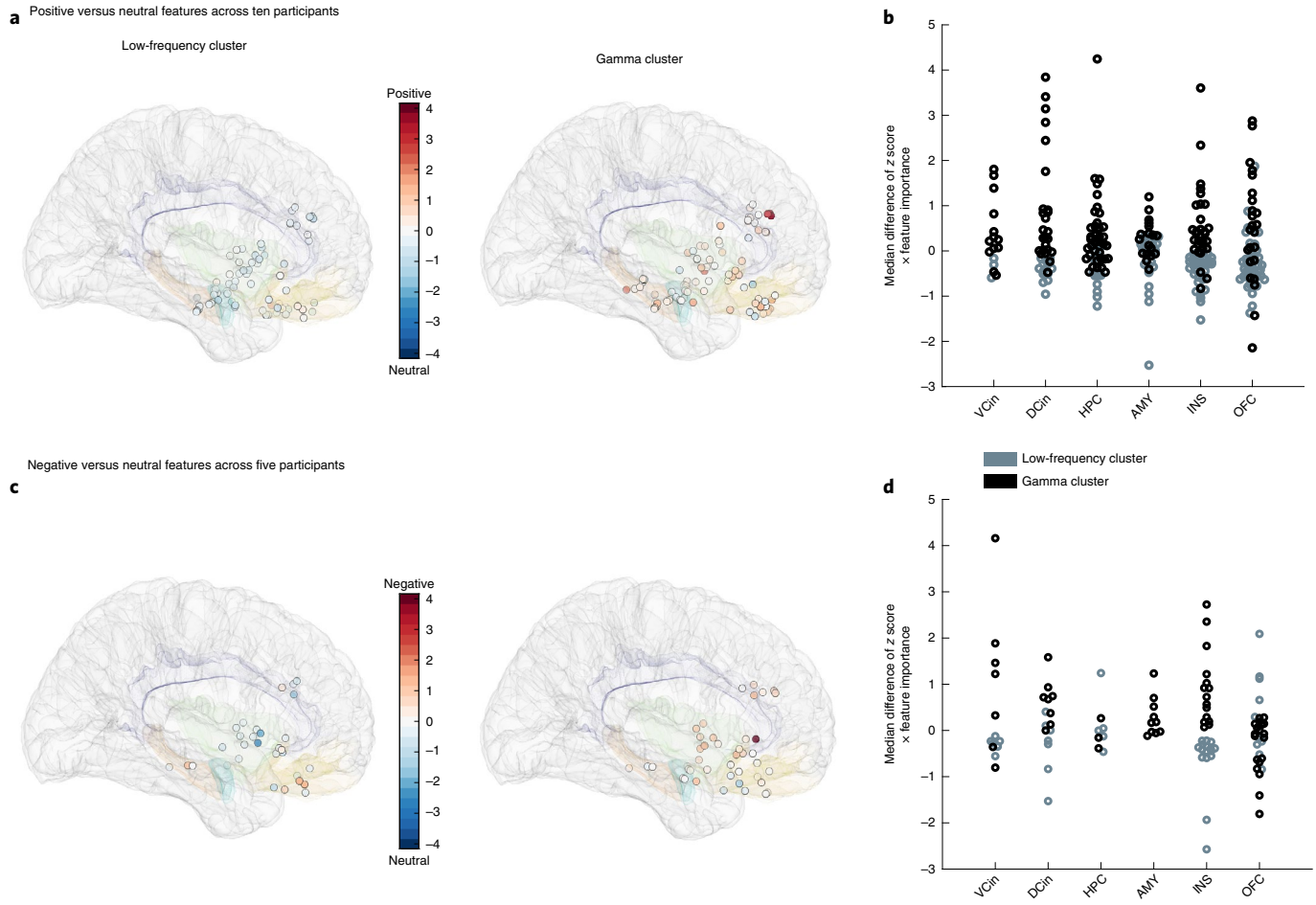


Fig. 4 | 'Gamma' and 'low-frequency' clusters belonged to a distributed network. a, Each of the electrodes on the MNI template brain (Electrode localization section) with their corresponding contribution to the 'low-frequency' (left) and 'gamma' (right) clusters from the positive decoders in all ten participants. White circles on the MNI brain indicate that the low-frequency cluster, which is scaled close to 0, was less important than the gamma cluster. **b**, Median difference scores (see Fig. 3 for details) from the gamma and low-frequency clusters from the positive decoders grouped by location. **c**,**d**, As in **a** and **b**, but pooled data from the negative decoders in five participants. INS, insula; VCin, ventral cingulate; DCin, dorsal cingulate; AMY, amygdala; HPC, hippocampus; OFC, orbitofrontal cortex.

trained the decoders in the same way for each region, one at a time. These analyses revealed that spectral features from the dorsal ACC (0.71 ± 0.12) and the insula (0.70 ± 0.12) led to greater generalizability score for the positive decoder than the spectral features from the OFC (0.60 ± 0.09, Fig. 5b). To investigate the role of the ACC further, we used another subset of participants with electrode coverage in the ventral ACC and trained the decoder in the same way, which resulted in a generalizability score of 0.76 ± 0.07 (Fig. 5c).

We next trained the negative decoders using spectral features from 4/5 participants with electrodes in the insula, OFC and dorsal ACC, resulting in a generalizability score of 0.65 ± 0.02 (Fig. 5d). Similar to the positive decoders, the dorsal ACC (0.61 ± 0.04) and insula (0.63 ± 0.06) both had a higher generalizability score than the OFC (0.55 ± 0.07, Fig. 5e). Decoders that were trained using spectral features from the ventral ACC in two participants had an average generalizability score of 0.61 ± 0.02.

Consistent with our within-subject results, the cross-subject decoding results demonstrated that the insula and ACC were important contributors to affective behaviours in general, but the role of OFC was less consistent. Although the ventral and dorsal ACC had similar generalizability scores when classifying negative from neutral behaviours, the ventral ACC was more important for discriminating positive affective behaviours (Fig. 5f). These results suggested that spectral features from the dorsal and ventral ACC made similar contributions to discriminating negative from neutral behaviours. The ventral ACC, however, appeared to be more important for distinguishing positive affective behaviours from neutral behaviours than the dorsal ACC.

Multiclass decoders classified three types of affective behaviour. We next trained within-subject multiclass RF decoders to distinguish amongst positive, negative and neutral behaviours in three participants with sufficient instances of each behavioural class ($\geq$15 samples within each fold of the dataset). Using all of the spectro-spatial features from the mesolimbic network, the multiclass decoder distinguished amongst all three types of behaviour with an average accuracy of 0.68 ± 0.016, which was significantly above chance level (33%) in each of the participants (Fig. 6a, Supplementary Fig. 12 and Table 8).

The multiclass decoder performance (F1 score) was better for positive than negative affective behaviours in each of the three participants (Supplementary Table 8 and RF classification section). This result was consistent with our prior finding with the binary decoders, which distinguished positive from neutral behaviours more effectively than negative from neutral behaviours (Fig. 2e). The decoding performance of positive versus negative affective behaviours

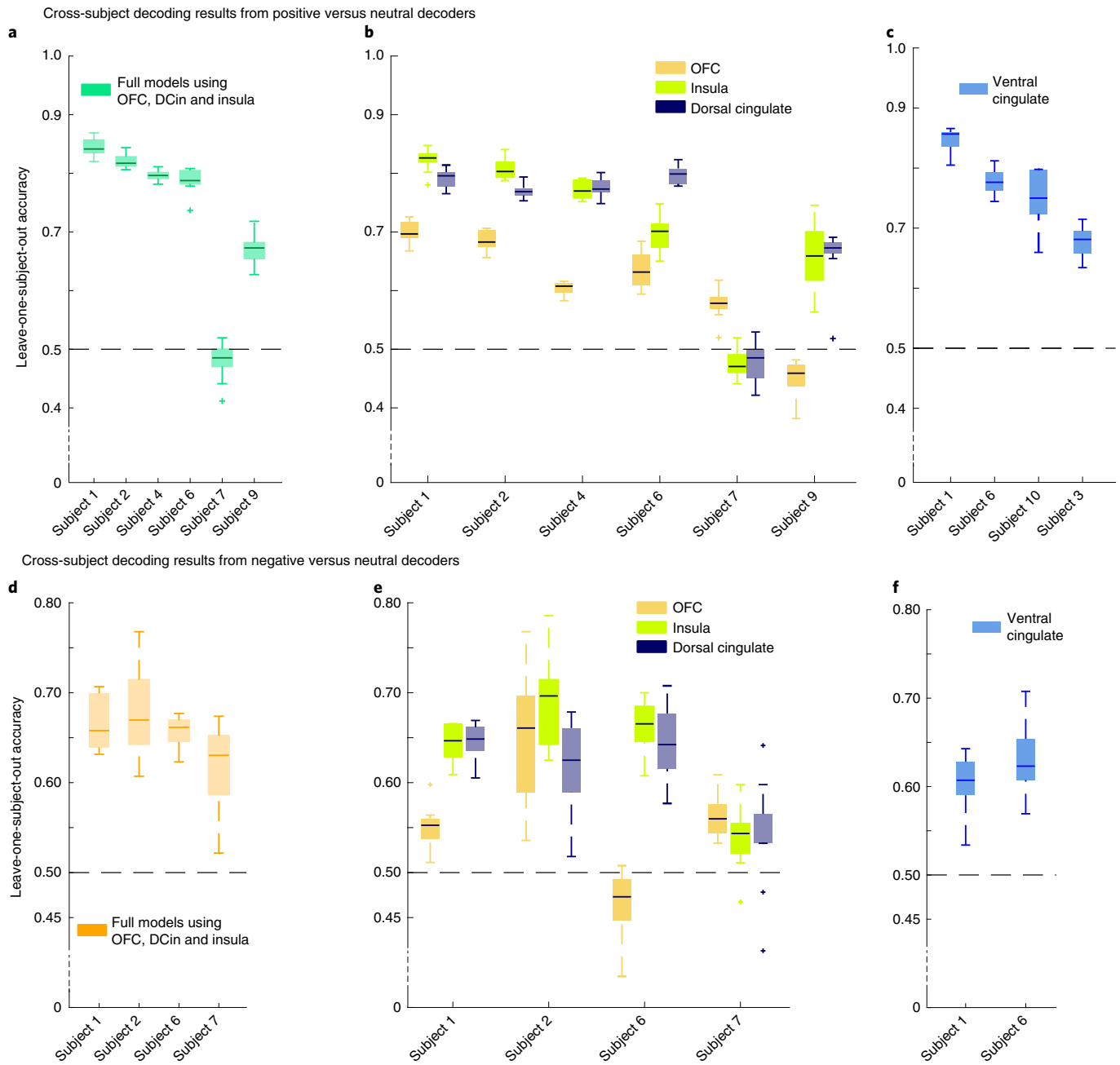


Fig. 5 | Cross-subject decoding showed that the spectral features from OFC, dorsal ACC and insula were generalizable across participants with implanted leads in these regions. For both positive and negative affective behaviours, the insula, dorsal ACC and ventral ACC were generalizable features compared with the OFC. **a**, The leave-one-subject-out accuracy for the positive decoders across $n=100$ datasets. Spectral features from all five frequency bands were averaged across contacts within each region for each participant, then each participant was omitted for training the model. The reported accuracies are test accuracies on each leave-one-out subject. All accuracy metrics are above 50% chance level except for subject 7. The average leave-one-out accuracy (referred to as the 'generalizability score') is 0.73 with s.d. 0.13 ($n=6$ participants). **b**, The leave-one-out accuracies of the decoders trained on the spectral features of each region ($n=5$ features), one at a time. Mean ± s.e.m. generalizability scores for OFC, insula and dorsal ACC are 0.60 ± 0.09, 0.70 ± 0.12 and 0.71 ± 0.12 ($n=6$ participants). **c**, We trained the cross-subject positive decoders in four participants with electrodes implanted in ventral ACC. The generalizability score for this region was 0.76 ± 0.07 ($n=4$). **d**, The leave-one-out accuracies for four participants included in training cross-subject models for negative versus neutral behaviours. The generalizability score was 0.65 ± 0.02. **e**, Similar to **b**, but for the negative decoders. The generalizability scores for OFC, insula and dorsal ACC were 0.55 ± 0.07, 0.63 ± 0.06 and 0.61 ± 0.04 ($n=4$ participants). **f**, Similar to **c**, where two participants out of five had implanted electrodes in ventral ACC. The generalizability score was 0.61 ± 0.02. VCin, ventral cingulate; DCin, dorsal cingulate. In box plots **a**–**f**, central lines represent the median and the two edges represent 25 and 75 percentiles; whiskers show the most extreme data points, and outliers are shown individually (see MATLAB boxplot function).

was consistent across participants, suggesting that these results were robust and not due to a larger number of positive than negative affective behaviours in each analysis (Supplementary Fig. 13).

We next looked across the group to examine which features were most important for the performance of the multiclass decoder across the participants. Group-level analyses of selected spectro-spatial

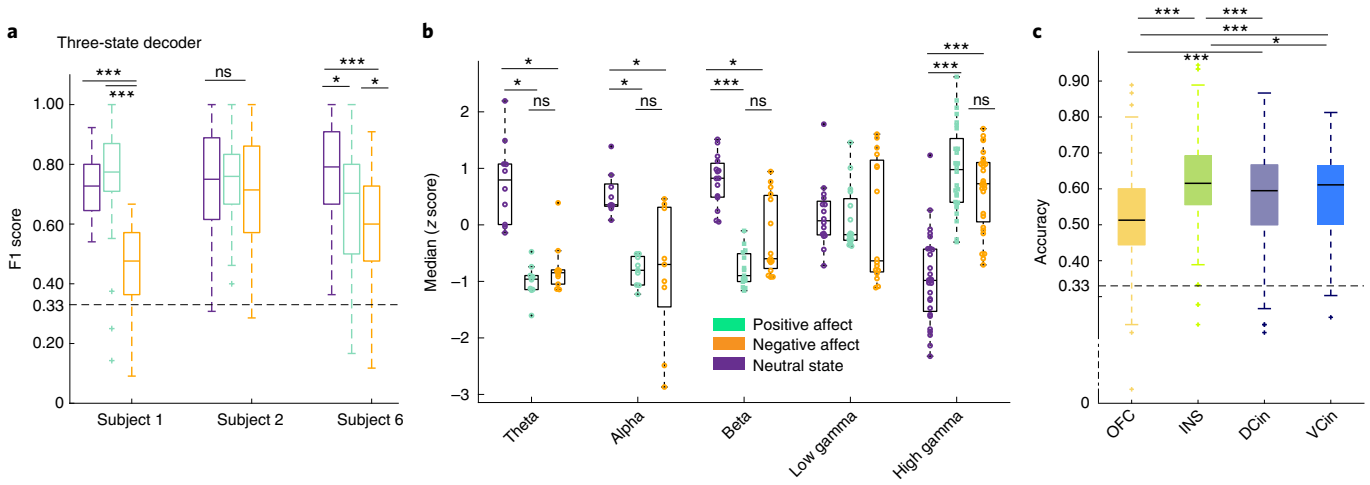


Fig. 6 | The multiclass decoder distinguished amongst positive, negative and neutral behaviours using the spectro-spatial features of the mesolimbic network. a, F1 scores for the three-class RF models from the three participants. All F1 scores were significantly above chance level (33%, dashed lines) and different from the shuffled models ($P < 0.0001$ for all participants, two-sided pairwise rank-sum test; Supplementary Table 8). Asterisks represent one-way multiple-comparison Kruskal–Wallis tests corrected with the Bonferroni method across the F1 scores of each affective behaviour within each participant. In participants 1 and 6, both positive ($P = 1.75 \times 10^{-35}$ and $P = 0.043$) and neutral ($P = 4.4 \times 10^{-24}$ and $P = 10^{-10}$) behaviours had significantly larger performance than negative behaviours. In subject 6, the positive is significantly different from the negative class ($P = 0.0001$). $^{***}P < 0.0001$, $^{**}P < 0.001$, $^{*}P < 0.05$. **b**, Median distribution of the selected features across the three participants. The Kruskal–Wallis multiple-comparison test among the three behavioural classes showed the following results: Theta ($n = 10$ for each behaviour): positive and negative affective behaviours differed from neutral behaviours ($P = 0.0001$ and $P = 0.0053$, respectively); Alpha ($n = 9$): positive and negative affective behaviours differed from neutral behaviours ($P = 0.0026$ and $P = 0.014$, respectively); Beta ($n = 15$): positive and negative affective behaviours differed from neutral behaviours ($P = 9.4 \times 10^{-7}$ and $P = 0.006$, respectively). Low gamma ($n = 16$): no significant difference was observed. High gamma ($n = 28$): positive and negative affective behaviours differed from neutral behaviours ($P = 1.36 \times 10^{-9}$ and $P = 6.4 \times 10^{-7}$, respectively). **c**, The multiclass decoder models were trained using the spectral features from each region and then pooled across the three participants. Abbreviations as in Fig. 4. OFC was from four probes implanted in three participants ($n = 400$, that is, 4×100 total datasets), insula ($n = 300$) and dorsal ACC ($n = 300$) are from three probes from three participants, and ventral ACC ($n = 200$) from two probes from two participants. Using Bonferroni-corrected Kruskal–Wallis multiple-comparisons test, the insula was significantly different from dorsal ACC ($P = 6.16 \times 10^{-6}$) and OFC ($P = 6.7 \times 10^{-29}$) and from ventral ACC ($P = 0.01$). Ventral ACC ($P = 1.7 \times 10^{-10}$) and dorsal ACC ($P = 6.7 \times 10^{-9}$) were both different from OFC. In the box plots **a**–**c**, central lines represent the median and the two edges represent 25 and 75 percentiles; whiskers show the most extreme data points, and outliers are shown individually (see MATLAB boxplot function).

features from the three participants (grouped by spectral band) demonstrated that high gamma activity was greater during positive and negative affective behaviours than during neutral behaviours and discriminated affective behaviours from neutral behaviours (Fig. 6b). Low-frequency activity in the theta, alpha and beta frequency bands, in contrast, was decreased during both positive and negative affective behaviours compared with neutral behaviours but did not differ significantly between affective behaviours of differing valence. These findings suggested that increased high gamma and decreased lower frequency activity across the mesolimbic network characterized affective behaviours in general.

To investigate whether spatially localized activity within the mesolimbic network differentiated amongst the three types of behaviour, we concatenated the decoder accuracies from each participant in regions that were sampled in at least two people (that is, the amygdala and hippocampus were not included here because they were each sampled in one participant), which included the insula (3/3 participants), dorsal ACC (3/3 participants), OFC (3/3 participants) and ventral ACC (2/3 participants). Non-parametric tests found that the accuracy of the insula was the highest, followed by the ventral and dorsal ACC, and, lastly, the OFC in distinguishing amongst behaviours with the multiclass decoder (Fig. 6c and Supplementary Table 9). Moreover, after training the multiclass decoders in each participant using the spectral features from each region, one at a time, we found that multiple regions successfully decoded the affective behaviours in each participant with accuracy significantly above chance (33%; Extended Data Fig. 8). Regions with large generalizability scores from the binary decoders, such as insula (3/3 participants) and dorsal ACC (2/3 participants), in particular, were most important for distinguishing amongst the positive, negative and neutral behaviours.

Discussion

We found evidence that direct neural recordings of the human mesolimbic network discriminated naturalistic affective behaviours from neutral behaviours with high accuracy. We trained decision tree-based models on the spectro-spatial mesolimbic features and successfully decoded positive (with up to 93% accuracy) and negative affective behaviours (with up to 78% accuracy) from neutral behaviours using binary decoders in individual participants. In general, affective behaviours were associated with coordinated changes across the mesolimbic network, including increased activity in high frequency bands (that is, gamma) and decreased activity in low-frequency bands (that is, theta, alpha and beta). By examining the contributions of different mesolimbic structures to decoding performance, certain regions emerged as playing more central and consistent roles in affective behaviours. Whilst the insula and ACC (both dorsal and ventral subregions) were the most generalizable spatial features across the sample, there was more person-specific variability in OFC. Although there were some spectro-spatial similarities between the positive and negative affective behaviours, each behavioural class had a different spatial topography within the network. In a subset of participants ($N = 3$), multiclass decoders highlighted the importance of increased high gamma activity during affective behaviours and emphasized the central role of the insula and ACC relative to other regions, such as OFC.

Our results indicate that distributed spectral changes across the mesolimbic network characterize naturalistic affective behaviours. Noninvasive EEG studies, which typically use task-based paradigms, have found consistent evidence that gamma band activity in the mesolimbic network increases in response to affective stimuli[11,21,36]. Despite numerous methodological differences between prior experimental studies and the approach we took here, we also found that, compared with neutral behaviours, positive and negative affective behaviours displayed in everyday life were also characterized by increased gamma band activity—as well as decreased low-frequency band activity—across the mesolimbic network[37]. Although many unanswered questions remain regarding the role of lower frequency bands in emotions and affect, our results suggest that affective behaviours that arise in more ecologically valid contexts may engage similar neural mechanisms—particularly when measured in high frequency bands—as those observed in more controlled settings.

Despite some common spectral patterns, an examination of the mesolimbic network activity's spatial topography revealed that certain regions contributed more strongly to affective behaviours than others. The insula and ACC, which are tightly connected structures with established roles in emotions and affect, also emerged as central regions for naturalistic affective behaviours. In the insula, simultaneous increases in gamma activity and decreases in low-frequency activity characterized both[8,17] positive and negative affective behaviours[38,39]. Prior studies have shown that stimulating the insula, an interoceptive relay station that is critical for experience[7,13], causes subjective visceral sensations and cardiovascular changes[40]. Insula engagement, therefore, may reflect its role in representing the bodily changes and feelings that accompany positive and negative affective behaviours. In the ACC, increased gamma activity characterized both positive and negative affective behaviours, but differences emerged in the degree to which the ACC subregions participated in behaviours of differing valence. Our results suggest that, whereas ventral ACC played a more generalized role during positive affective behaviours, both ventral and dorsal ACC may be important for negative affective behaviours. These findings are largely consistent with neuromodulation studies that have shown that whilst stimulation of ventral ACC can cause laughter and mirth[12,41], stimulation of dorsal ACC can cause feelings of doom and fear[42]. The dorsal and ventral ACC have different anatomical projections to autonomic and motor centres that are critical for emotions[59], and our results suggest that ACC subregions may engage these distinct pathways to produce positive and negative affective behaviours.

The amygdala and hippocampus were also important structures for decoding positive and negative affective behaviours from neutral behaviours. The amygdala, though often associated with negative emotions[11,15,18,43], activates during negative and positive states of sufficient intensity and supports affiliative behaviours as well as threat responses[15,44]. Stimulation studies have also found that brief perturbation of amygdala subnuclei can induce rapid negative[11] as well as positive affective reactions (Event related potentials at ~200–400 ms)[18]. In one participant with amygdala coverage, we found that gamma activity in the amygdala played a prominent role in negative affective behaviours, but more evidence is needed to corroborate this result. With dense reciprocal connections, the amygdala[45,46] and hippocampus are essential for emotional memories, which participants may have recalled and relived during spontaneous moments of affect. Although the role of the hippocampus in mood and emotion is still debated, recent iEEG studies have found that lower mood is associated with greater beta coherence between the amygdala and hippocampus[47], which suggests that both structures, and their interaction, may be critical for positive and negative affective behaviours.

Compared with other regions in the mesolimbic network, the OFC played a less consistent role in naturalistic affective behaviours. The OFC, especially in lateral areas, is critical for emotion regulation, cognitive control and behavioural inhibition[28,48,49]. In prior research, longitudinal measures of spontaneous OFC activity have predicted variations in mood[29]. A recent neuromodulation study also found that stimulation of lateral OFC decreased theta activity across the mesolimbic network and improved mood[30], which suggested that suppression of low-frequency activity yielded affective benefits. Our results indicated that activity in low-frequency bands across the mesolimbic network decreased during both positive and negative affective behaviours, but the OFC was not a robust correlate of either behaviour. Unlike prior studies, which relate measures of self-reported mood to neural activity over minutes[47] to hours[29], our study investigated neural changes during much briefer periods, a difference in timescale that may help to explain the heterogeneous results. Our findings suggest the OFC may be engaged in different ways depending on the emotional context, thus making its contribution to affective behaviours more variable across instances and participants.

Our results offer a comprehensive window into the neural mechanisms of the mesolimbic network. Although there is ongoing debate regarding the degree to which different affective states have unique or shared representations in the brain, the present study helps to elucidate how a distributed network is associated with different affective states via spectro-spatial patterning. Positive and negative affective behaviours differ in valence, but both can vary in arousal or intensity levels. Some of our results suggested that common changes in mesolimbic network activity are associated with affective behaviours in general, regardless of whether the behaviours were positive or negative, and it is possible that these common increases reflected heightened arousal. In particular, increased gamma activity and decreased low-frequency activity characterized both positive and negative affective behaviours. There were also regions (insula, ACC, hippocampus and amygdala) that contributed more strongly than other regions (OFC) to both types of behaviour. We speculate that the shared gamma activity in these regions during both positive and negative affective behaviours may represent arousal or intensity of emotional experience, a dimension of affect that may have been on a comparable scale during both types of behaviour. Our results also indicated, however, that different structures within the mesolimbic network made distinct contributions to positive and negative affective behaviours and may have helped to shape these distinct affective states. Whereas increased gamma activity in the ventral ACC, hippocampus and dorsal ACC contributed more to positive affective behaviours, increased gamma activity in the amygdala (in one participant) played a prominent role in negative affective behaviours. A distributed network that activates through a combination of spectral changes and spatial changes would be a flexible system that is prepared to produce a variety of affective states.

The present study has limitations to consider. First, we analysed neural activity in participants undergoing seizure monitoring for epilepsy during a multiday hospital stay. Thus, there was variability in both electrode placement, which was based on clinical needs, and in the affective behaviours demonstrated by participants. Although there was overlapping electrode coverage in multiple mesolimbic structures across participants, even electrodes in a single region may have sampled distinct subregions in different people, which may have increased variability across the sample. As the naturalistic affective behaviours were spontaneous actions exhibited throughout their hospitalization, there was also variability in the number and types of affect that participants displayed. Whereas the positive affective behaviours were fairly uniform (mostly smiling and laughing), the negative affective behaviours were more heterogeneous and included a range of expressions including pain and frustration. Additional studies are needed to determine how the mesolimbic network produces each of these specific affective behaviours.

Second, due to the unconstrained nature of our study, we did not have measures of self-reported experience, arousal, engagement or movement that aligned with the continuous neural recordings.

We conducted several follow-up analyses, however, to confirm that the associations that we found between the spectro-spatial changes and the affective behaviours were robust. We found similar neural activity patterns when, instead of contrasting positive and negative affective behaviours against neutral behaviours, we compared them with rest (that is, presumably low-arousal moments during which no behaviours were annotated). When we constrained our analyses to examine positive and negative affective behaviours during conversations (that is, where affective and neutral moments presumably had comparable levels of engagement and movement), our results also remained unchanged. These analyses offered additional evidence that the spectro-spatial patterns we found for positive and negative affective behaviours were not accounted for by variations in arousal or movement.

In summary, we used statistical and machine learning approaches[31] to decode naturalistic affective behaviours from direct recordings of the human mesolimbic network. Complex, real-time decoding models trained on neural activity in sensorimotor and language cortices have made it possible to design brain–computer interfaces for those who suffer from limb[32] or speech disability[40]. Similar advances are lacking in neuropsychiatry, however, and it remains difficult to relate neural signals to complex emotions and mood[33]. More sophisticated neuroanatomical models of affective behaviours and symptoms will help to inform personalized treatments for mental health disorders and to identify biomarkers that can be monitored in treatments such as closed-loop neurostimulation.

Methods

Participants and inclusion criteria. Participants were 11 patients (6 female, 5 male, age 20–43 years; Supplementary Table 1) with treatment-resistant epilepsy who underwent (iEEG) implantation for seizure localization. Participants were included in the study if they had electrodes in at least three mesolimbic regions and displayed a sufficient number of affective behaviours to train the RF classifiers (Supplementary Table 1 and Supplementary Table 4). No statistical methods were used to pre-determine the sample size, but our sample size is similar to those in previous iEEG publications[29,47]. All procedures were approved by the University of California San Francisco Institutional Review Board. Participants provided written informed consent to participate prior to surgery. Data collection and analysis were not performed blind to the conditions of the experiments. Further information can be found in the Nature Research Reporting Summary.

On post-operative day 2, when the behavioural annotations for our study began, all participants had a normal mental status, which was determined by assessing alertness, orientation (person, place, time and situation), interaction with clinical staff, ability to follow verbal commands and ability to participate in experimental tasks without difficulty. In all participants, anti-epileptic medications were stopped by post-operative day 2. Post-operative pain was in the mild range for all participants after post-operative day 2, except for one participant who reported borderline moderate pain. Pain and emesis were treated with standing and as-needed medications. Anti-epileptic medications that were administered included clobazam, oxcarbazepine, levetiracetam, zonisamide, topiramate, lamotrigine, lacosamide, carbamazepine and phenytoin. Pain medications included acetaminophen, hydrocodone/acetaminophen, oxycodone, hydromorphone and ondansetron.

iEEG and behavioural data acquisition. Over a multiday hospitalization, participants underwent continuous 24-h audiovisual recording and iEEG monitoring through the Natus clinical recording system as a part of routine clinical care. Electrophysiological data were collected at a sampling rate of either 512 or 1024 Hz. All mesolimbic structures were sampled by subdural grid, Ad-Tech four-contact strip and Ad-Tech four/ten-contact depth electrodes (10 or 6 mm centre-to-centre spacing). Two participants had mini-grids implanted on OFC.

Electrode localization. For electrode localization, pre-operative 3-T brain magnetic resonance imaging (MRI) and post-operative computed tomography (CT) scans were obtained for all participants. Statistical parametric mapping software SPM12[34] and FreeSurfer[50] were used to reconstruct and visualize the pial surface electrodes. Electrode locations were validated by an expert's visual examinations of the co-registered CT and MRI. Montreal Neurological Institute (MNI) template brain was used for brain visualization. The MNI coordinates of the electrodes in all participants, a spherical, sulcal-based alignment, was used to non-linearly register the surface using FreeSurfer (cvs_avg35_inMNI152 template)[51]. Participants had electrodes in at least three mesolimbic regions, which included the insula (most electrodes were in anterior or mid-insula but some were in posterior insula), ACC (dorsal and ventral subregions), OFC, amygdala and hippocampus.

Behavioural annotations. Eleven human raters, blind to the study's goals and hypotheses, manually annotated the video recordings using ELAN software[52], a linguistic ethnographic software. Spontaneous affective behaviours including smiling, laughing, positive verbalizations, pain–discomfort and negative verbalizations were annotated and coded on a millisecond basis as a tick at the behaviour onset (Fig. 1b and Supplementary Table 2). Raters marked the onset of each behavioural instance with 'ON' and the offset with 'OFF,' an annotation system that allowed the raters to code the videos with high efficiency. To minimize potential bias, the raters were assigned to randomized 10-min segments of continuous video recordings. To minimize the effects of electrode implant surgery on behaviour and mood, only annotations occurring two or more days post-surgery were included in our analyses.

A subset of videos was annotated by two raters, and overall, there was high inter-rater agreement: 82% of the total instances of affective behaviour that were logged by one rater were also logged by the other. The instances were highly overlapping in time, with the onset of each instance having a median difference of 0.87 s (mean 7.4 s, Extended Data Fig. 1C) between any two raters. In cases where there was a disagreement between coders, there was a consensus meeting with a third member who served as the 'tie-breaker'. The ratings from the third coder were used in these cases. Moreover, there was somewhat lower reliability for the annotations of the negative affective behaviours (79% agreement) compared with the positive affective behaviours (89% agreement).

In addition to affective behaviours, we also considered behaviours without an observable affective component, such as eating, drinking, sleeping, etc. Raters marked the onset of the given activity with 'ON' and the offset when the behaviour was finished with 'OFF.' These continuous behaviours were used to define 'rest' (during which no activity of interest including affective and non-affective was observed; 'Behavioural Annotations' in Supplementary Information and Supplementary Tables 2 and 3) and neutral behaviours with conversation.

iEEG pre-processing. We appended and aligned the raw iEEG recordings and the annotations of the affective behaviours (Extended Data Fig. 2). All channels were de-meaned, notch filtered (second-order Butterworth filter) at 60 Hz and its harmonics, and decimated (zero-phase 30th-order finite impulse response filter) to 512 Hz. We visualized the pre-processed signals using EEG Lab[53] to remove noisy electrodes and to mark epochs in which there were motion or interictal artefacts[53]. After excluding noisy channels, we re-referenced the recordings to the common average signal across the electrodes localized to the same depth/strip leads. Next, we appended the cleaned data to form 'chunks', which ranged from 40 min to 4 h of continuous data. All analyses were programmed in MATLAB.

Time–frequency analyses and feature extraction. We applied time–frequency decomposition to each electrode located in the grey matter of the mesolimbic regions. To extract the neural features, we applied the Hilbert transform (MATLAB Hilbert function) to band pass-filtered signals using second-order Butterworth filters specific to the following five frequency bands: 4–8 Hz (theta), 8–12 Hz (alpha), 12–30 Hz (beta), 30–55 Hz (low gamma) and 70–150 Hz (high gamma). These frequency bands are thought to correlate with cognitive functions[54]. The resulting time–frequency signals for each channel (that is, five different bands) were z-scored within each chunk of data and averaged using 10-s, non-overlapping bins centred about the occurrence of each affective behaviour. This averaging window allowed us to control for inter-rater variability underlying the true occurrence of the behaviour. We aggregated the affective behaviour data into an input matrix—with dimensions of channel numbers from various brain regions times frequency for each behavioural class—for the decoder. The 10-s averaging bins were also applied to the binary time-domain trace of the affective behaviour.

Power spectral density analyses. We computed the power spectral density of each channel within a 1-s window using the Welch method (MATLAB pwelch function) and Hanning window of length fs/5 with 50% overlap, with a 256 non-uniform fast Fourier transform or the next power of 2.

RF classification. Data preparation. We compared neural signals underlying affective and neutral behaviours using a RF classifier. First, instances of neutral behaviour were extracted from 10-min periods (or more) during which there were no annotated behaviours. Specifically, these neutral states were sampled from different periods of annotations (that is, some from the first 2 h and some from the fourth period of 2-h continuous data). As the number of neutral samples exceeded the number of affective behaviours, we next used a bootstrap procedure to construct multiple balanced datasets with equal labels such that the number of neutral labels was equal to the number of labels for each affective category (positive or negative). We applied this procedure 100/k (fold number) times for each subject to make 100 datasets (Extended Data Fig. 2).

Model training. As behavioural instances could have occurred close in time, which could artificially inflate correlations between neural features and lead to model overfitting, we used a conservative sequential k-fold cross-validation classifier to train the RF models in each participant[55]. To evaluate the decoding performance of our models, we constructed surrogate permutation models by shuffling the

behavioural class labels. To avoid any information overlap between training and test sets, we selected folds such that the training and test divisions of the observations were non-adjacent in time (that is, sequential cross validation). The number of folds, k, was chosen to be five or ten for each participant such that a minimum of ten observations were included in each fold. The number of samples varied between 28 and 164 within each class across participants (that is, positive, negative or neutral behaviours; Supplementary Table 4). The RF classifiers were trained with 300 trees and were optimized for two hyper-parameters: (1) each tree was grown such that the maximum number of samples per leaf was varied in the range of 1 and 20 and (2) the number of features at each node varied in the range of 1 to the maximum number of features minus 1.

Likewise, multiclass decoders were trained in the same way as the binary decoders, with equal number of samples for each class. To train and optimize the RF models, we used the 'TreeBagger' and 'bayesopt' MATLAB functions. We measured decoding performance for binary and multiclass decoders using the AUC and the F1 score (as below), respectively. We also measured decoding accuracy, which is the total number of true predicted samples divided by the total sample number, as a general metric of performance.

$$\text{F1 score} = \left(2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right).$$

Feature selection. Using the out-of-bag error estimate from the RF models, we were able to select the importance of each tree node (that is, feature). Briefly, each tree is built using bootstrap samples from the original samples after holding out one-third of the original samples to form a test set. Once the tree is built, the left-out samples are classified, and the average number of times that the predicted class is not equal to the true class is called the 'prediction error'. This is a standard practice in defining feature importance in RF classification[56]. We refer to the model prediction error for each feature as 'feature importance'. Subsequently, we ranked the average feature importance from all 100 runs and found the knee point of its cumulative summation curve for each participant using an algorithm called 'kneedle', which estimates the knee point based on the maximum curvature for a discrete set of points[41]. The cumulative set of features leading up to the knee point were selected as the important features for each decoder type. This method served as an objective threshold to select the neural features that were the dominant contributors to the positive or negative decoders. Lastly, we compared the distribution of the feature importance across all 100 RF models with permuted models and kept those features with significant difference between the main RF and permuted models (Supplementary Figs. 2 and 3).

Feature normalization for group-level analyses. Proceeding to feature extraction for each participant, we extracted the sample distributions of the important features for each behaviour type. Second, we extracted the median amplitude of each feature for each behaviour type. To avoid undue influence of participants with stronger neural activity, we z-scored the median values across all the selected features in each behaviour type, separately. The normalization procedure is also depicted in Supplementary Fig. 11. We next performed group-level analyses using the z-scored median of selected spectro-spatial features to examine the extent to which the spectral patterns in each participant held across individuals, and then grouped these values by their frequency bands.

Also, to account for within- and between-subject variability in the feature importance values from the RF models, we normalized these values to the maximum value within each participant because the feature importance has a positive value when the model does not overfit due to noise.

Clustering. To assess collinearity between selected features (for example, correlations between high gamma band activity in different regions) and to map the spectro-spatial features of each behavioural class, we computed the correlation matrix across samples for the positive, negative and neutral behaviours used in the binary decoders. Next, we applied hierarchical clustering to the correlation matrix (Supplementary Fig. 8) to group the features into two main groups for each participant with an objective approach. This clustering analysis identified two clusters from the positive and negative decoders that separated affective from neutral behaviours based on spectral bands rather than regions in most of the participants.

We then populated the spectro-spatial features across participants ($n = 10$) from the positive decoders and observed that 56% of features consisted of low and high gamma power, and 44% of features consisted of theta, alpha and beta band power (Extended Data Fig. 5A, pie chart). When we investigated the contribution of each frequency band to each cluster from the positive decoders, however, 80.7% of the features in cluster 1 were in theta, beta and alpha bands, and 19.3% were in low and high gamma bands. Therefore, we named cluster 1 the 'low-frequency cluster'. Similarly, 85.2% of the features in cluster 2 were in the low and high gamma bands (Extended Data Fig. 5A, histogram). We observed qualitatively similar results with the negative decoders ($n = 5$ participants, Extended Data Fig. 5B). Indeed, 98% of features in the low-frequency cluster were in theta, beta and alpha bands, and 78% of features in the gamma cluster were in low and high gamma bands.

Then the z-scored median of the selected spectro-spatial features (Feature normalization for group-level analyses section) for each cluster was extracted at the individual level and populated across the sample. Next, we computed a difference score for the selected spectro-spatial feature within each cluster by subtracting the z-scored median activity in that cluster during neutral behaviours from the z-scored median activity during the positive or negative affective behaviours (Supplementary Fig. 11)

To examine the contribution of gamma and low-frequency cluster activity in certain regions, we scaled the z-scored median differences for each cluster by their normalized feature importance from the RF decoder models (Supplementary Figs. 9 and 10).

Binary decoders from each mesolimbic region. To assess the contribution of the mesolimbic regions regardless of the spectral bands, we re-trained the within-subject positive (Extended Data Fig. 6) and negative decoders (Extended Data Fig. 7) in each region, one at a time. For each region, we included all spectral features from all electrodes implanted in that structure (and discarded the spectral information from all other regions across the network) and identified the top regions in each participant that distinguished positive or negative affective behaviours from neutral behaviours significantly better than other regions (Supplementary Tables 6 and 7) using the Kruskal–Wallis multiple-comparison test, corrected with the Bonferroni method.

Generalizability score. We averaged the z-scored value of each frequency band from all contacts on a given electrode, which resulted in five spectral features for each region and a total of 15 features per participant. To train the cross-subject positive decoders, all 15 features for the positive and neutral behaviours from six participants were stacked to form the feature matrix. We used leave-one-out subject cross validation to train the decoders and calculated a generalizability score as the mean leave-one-out accuracy across all participants (that is, larger leave-one-out accuracy implied greater generalizability of the decoder). To train the cross-subject negative decoders, all 15 features for the negative and neutral behaviours from four participants were stacked to form the feature matrix.

Support vector machine (SVM) model classification. Linear SVM. Linear SVMs were trained using all spectro-spatial features, as were the full RF models. We optimized hyper-parameters on 80% of the training set, then the model was trained by the optimized parameters using the 20% held out of the training dataset. Supplementary Figs. 5 and 6 demonstrate the performance of these models in comparison with the RF models, as well as the absolute values of the top 15 features sorted by SVM weights and RF model prediction error. The similarity index shows the percentage of the features for the RF and SVM models ('Additional tests for the feature selection' section in Supplementary information).

Non-linear SVM. To assess the robustness of the features selected by the RF models, we trained non-linear SVM classifiers using 'rbf' kernels on the same samples used in RF but using the selected features from the RF models (Supplementary Fig. 7). Ten-fold cross validation was used in the training set to optimize non-linear SVM parameters (that is, γ and C by 10-division grid search in the range of −1,000 to 1,000). Here, γ is the inverse of the s.d. of the rbf kernel, a type of Gaussian function (intuitively, it is a similarity measure between two data points and sets the decision boundary), and C is the regularization or penalizing parameter. All custom scripts are written in MATLAB. The 'fitcsvm' function in MATLAB was used to train and optimize the SVM models.

Statistical analyses. We assessed the statistical significance of all models by training surrogate RF models after shuffling the categorical labels within each fold of each dataset (to keep the balance between behavioural classes). All P values were computed using two-sided non-parametric rank-sum tests between pairs of distributions. For group-level statistical tests, one-way Kruskal–Wallis multiple-comparison tests were conducted followed by Bonferroni-adjusted tests to correct for multiple comparisons. For comparing the decoder performances and feature importance between the main and surrogate (permuted) models, in which they have similar features but with shuffled labels, the significance level is corrected to 0.0005 (0.05/100) since there were 100 trained models.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The collected neural and behavioural data are a modified version of clinical recordings for the purpose of seizure localization and clinical decisions. Thus, the minimum de-identified dataset used to generate the findings of this study will be available upon reasonable request to the corresponding author. Source data for the figures are available upon reasonable request. Contact M.B. via e-mail with enquiries.

Code availability

The code written to train the classifiers is available at: https://github.com/MBijanzadeh/DecodingAffect. The code to generate the figures will be available upon request. Contact M.B. via e-mail with any inquiries.

Received: 28 January 2021; Accepted: 18 January 2022;
Published online: 10 March 2022

References

1. Ochsner, K. & Gross, J. The cognitive control of emotion. *Trends Cogn. Sci.* **9**, 242–249 (2005).
2. Barrett, L. F., Mesquita, B., Ochsner, K. N. & Gross, J. J. The experience of emotion. *Annu. Rev. Psychol.* **58**, 373–403 (2007).
3. Ochsner, K. N., Silvers, J. A. & Buhle, J. T. Functional imaging studies of emotion regulation: a synthetic review and evolving model of the cognitive control of emotion: functional imaging studies of emotion regulation. *Ann. N. Y. Acad. Sci.* **1251**, E1–E24 (2012).
4. Lieberman, M. D. et al. Affect labeling disrupts amygdala activity in response to affective stimuli. *Psychol. Sci.* **18**, 421–428 (2007).
5. Lieberman, M. D. Social cognitive neuroscience: a review of core processes. *Annu. Rev. Psychol.* **58**, 259–289 (2007).
6. Touroutoglou, A., Hollenbeck, M., Dickerson, B. C. & Feldman Barrett, L. Dissociable large-scale networks anchored in the right anterior insula subserve affective experience and attention. *NeuroImage* **60**, 1947–1958 (2012).
7. Uddin, L. Q. Salience processing and insular cortical function and dysfunction. *Nat. Rev. Neurosci.* **16**, 55–61 (2015).
8. Zhang, Y. et al. The roles of subdivisions of human insula in emotion perception and auditory processing. *Cereb. Cortex* **29**, 517–528 (2019).
9. Seeley, W. W. et al. Dissociable intrinsic connectivity networks for salience processing and executive control. *J. Neurosci.* **27**, 2349–2356 (2007).
10. Chouchou, F. et al. How the insula speaks to the heart: cardiac responses to insular stimulation in humans. *Hum. Brain Mapp.* **40**, 2611–2622 (2019).
11. Oya, H., Kawasaki, H., Howard, M. A. & Adolphs, R. Electrophysiological responses in the human amygdala discriminate emotion categories of complex visual stimuli. *J. Neurosci.* **22**, 9502–9512 (2002).
12. Adolphs, R., Tranel, D., Damasio, H. & Damasio, A. Fear and the human amygdala. *J. Neurosci.* **15**, 5879 (1995).
13. Takahashi, H. et al. Brain activations during judgments of positive self-conscious emotion and positive basic emotion: pride and joy. *Cereb. Cortex* **18**, 898–903 (2008).
14. 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).
15. Phelps, E. A. & LeDoux, J. E. Contributions of the amygdala to emotion processing: from animal models to human behavior. *Neuron* **48**, 175–187 (2005).
16. Strange, B. A. & Dolan, R. J. Adrenergic modulation of emotional memory-evoked human amygdala and hippocampal responses. *Proc. Natl Acad. Sci. USA* **101**, 11454–11458 (2004).
17. Krolak-Salmon, P. et al. An attention modulated response to disgust in human ventral anterior insula: disgust in ventral insula. *Ann. Neurol.* **53**, 446–453 (2003).
18. Meletti, S. et al. Fear and happiness in the eyes: an intra-cerebral event-related potential study from the human amygdala. *Neuropsychologia* **50**, 44–54 (2012).
19. Omigie, D. et al. Intracranial markers of emotional valence processing and judgments in music. *Cogn. Neurosci.* **6**, 16–23 (2015).
20. Hajcak, G. & Nieuwenhuis, S. Reappraisal modulates the electrocortical response to unpleasant pictures. *Cogn. Affect. Behav. Neurosci.* **6**, 291–297 (2006).
21. Jung, J. et al. Intracerebral gamma modulations reveal interaction between emotional processing and action outcome evaluation in the human orbitofrontal cortex. *Int. J. Psychophysiol.* **79**, 64–72 (2011).
22. Wang, X.-W., Nie, D. & Lu, B.-L. Emotional state classification from EEG data using machine learning approach. *Neurocomputing* **129**, 94–106 (2014).
23. Merkl, A. et al. Modulation of beta-band activity in the subgenual anterior cingulate cortex during emotional empathy in treatment-resistant depression. *Cereb. Cortex* **26**, 2626–2638 (2016).
24. Zheng, J. et al. Multiplexing of theta and alpha rhythms in the amygdala–hippocampal circuit supports pattern separation of emotional information. *Neuron* **102**, 887–898.e5 (2019).
25. Hu, X. et al. EEG correlates of ten positive emotions. *Front. Hum. Neurosci.* **11**, 26 (2017).
26. Guillory, S. A. & Bujarski, K. A. Exploring emotions using invasive methods: review of 60 years of human intracranial electrophysiology. *Soc. Cogn. Affect. Neurosci.* **9**, 1880–1889 (2014).
27. Mukamel, R. & Fried, I. Human intracranial recordings and cognitive neuroscience. *Annu. Rev. Psychol.* **63**, 511–537 (2012).
28. Kragel, P. A. & LaBar, K. S. Multivariate pattern classification reveals autonomic and experiential representations of discrete emotions. *Emotion* **13**, 681–690 (2013).
29. Sani, O. G. et al. Mood variations decoded from multi-site intracranial human brain activity. *Nat. Biotechnol.* **36**, 954–961 (2018).
30. Rao, V. R. et al. Direct electrical stimulation of lateral orbitofrontal cortex acutely improves mood in individuals with symptoms of depression. *Curr. Biol.* **28**, 3893–3902.e4 (2018).
31. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
32. Nuyujukian, P. et al. Cortical control of a tablet computer by people with paralysis. *PLoS ONE* **13**, e0204566 (2018).
33. Kashihara, K. A brain–computer interface for potential non-verbal facial communication based on EEG signals related to specific emotions. *Front. Neurosci.* **8**, 244 (2014).
34. Ashburner, J. & Friston, K. Multimodal image coregistration and partitioning—A unified framework. *NeuroImage* **6**, 209–217 (1997).
35. Zajonc, R. B. Preferences need no inferences. *Am. Psychol.* **25**, 151–175 (1980).
36. Popov, T., Steffen, A., Weisz, N., Miller, G. A. & Rockstroh, B. Cross-frequency dynamics of neuromagnetic oscillatory activity: two mechanisms of emotion regulation: oscillatory activity during emotion regulation. *Psychophysiology* **49**, 1545–1557 (2012).
37. Ezzyat, Y. et al. Direct brain stimulation modulates encoding states and memory performance in humans. *Curr. Biol.* **27**, 1251–1258 (2017).
38. Seeley, W. W. The salience network: a neural system for perceiving and responding to homeostatic demands. *J. Neurosci.* **39**, 9878–9882 (2019).
39. Craig, A. D. How do you feel? Interoception: the sense of the physiological condition of the body. *Nat. Rev. Neurosci.* **3**, 655–666 (2002).
40. Anumanchipalli, G. K., Chartier, J. & Chang, E. F. Speech synthesis from neural decoding of spoken sentences. *Nature* **568**, 493–498 (2019).
41. Satopää, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a 'kneedle' in a haystack: detecting knee points in system behavior. *Proc. Int. Conf. Distrib. Comput. Syst.* https://doi.org/10.1109/ICDCSW.2011.20 (2011).
42. Inman, C. S. et al. Human amygdala stimulation effects on emotion physiology and emotional experience. *Neuropsychologia* **145**, 106722 (2020).
43. Phelps, E. A. Human emotion and memory: interactions of the amygdala and hippocampal complex. *Curr. Opin. Neurobiol.* **14**, 198–202 (2004).
44. Bickart, K. C., Dickerson, B. C. & Feldman Barrett, L. The amygdala as a hub in brain networks that support social life. *Neuropsychologia* **63**, 235–248 (2014).
45. Zheng, J. et al. Amygdala-hippocampal dynamics during salient information processing. *Nat. Commun.* **8**, 14413 (2017).
46. Fournier, N. M. & Duman, R. S. Illuminating hippocampal control of fear memory and anxiety. *Neuron* **77**, 803–806 (2013).
47. Kirkby, L. A. et al. An amygdala-hippocampus subnetwork that encodes variation in human mood. *Cell* **175**, 1688–1700.e14 (2018).
48. Gross, J. J. & Feldman Barrett, L. Emotion generation and emotion regulation: one or two depends on your point of view. *Emot. Rev.* **3**, 8–16 (2011).
49. Kragel, P. A., Knodt, A. R., Hariri, A. R. & LaBar, K. S. Decoding spontaneous emotional states in the human brain. *PLoS Biol.* **14**, e2000106 (2016).
50. Fischl, B. FreeSurfer. *NeuroImage* **62**, 774–781 (2012).
51. Fischl, B., Sereno, M. I., Tootell, R. B. H. & Dale, A. M. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum. Brain Mapp.* **8**, 272–284 (1999).
52. Sloetjes, H. & Wittenburg, P. *Annotation by category - ELAN and ISO DCR. 5* (European Language Resources Association (ELRA), 2008).
53. Delorme, A. & Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**, 9–21 (2004).
54. Schnitzler, A. & Gross, J. Normal and pathological oscillatory communication in the brain. *Nat. Rev. Neurosci.* **6**, 285–296 (2005).
55. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
56. James, G., Witten, D., Hastie, T. & Tibshirani R. *An Introduction to Statistical Learning: with Applications in R* (Springer, 2013).

Acknowledgements

We thank Chang laboratory members B. Speidel, D. Chandramohan, K. Sellers, L. Kirkby and P. Hullet and raters N. Goldberg-Boltz, L. Bederson, M. Solberg, C. Eun, J. Gordon, D. Tager, V. Cheng, N. Mummaneni and N. Kunwar. This research was funded by the NIMH (R01MH122431) and the Defense Advanced Research Projects Agency (DARPA) under Cooperative Agreement Number W911NF-14-2-0043. The views, opinions and/or findings contained in this material are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

M.B. performed all analysis. M.D., D.L.W. and E.F.C. designed the study. D.L.W. and M.D. assisted with subject recruitment, data collection and leading behavioural annotations. M.B. and A.N.K. conceptualized the analytical framework. M.B., M.D. and A.S. performed neural data cleaning from epileptiform activity. M.B., A.N.K. and V.E.S. wrote the manuscript with input from other authors. H.E.D. and E.F.C. supervised the experimental work.

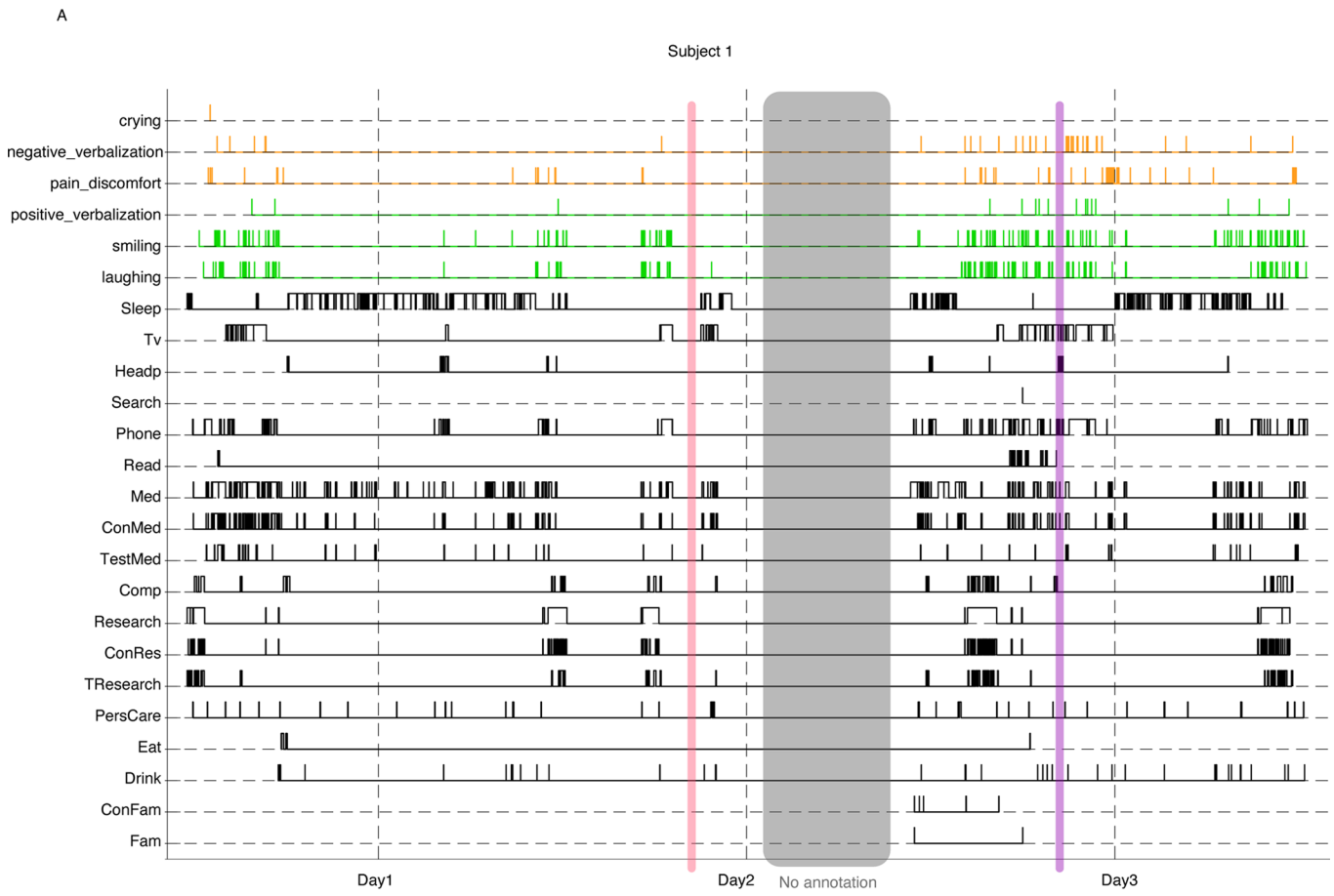
Competing interests
The authors declare no competing interests.

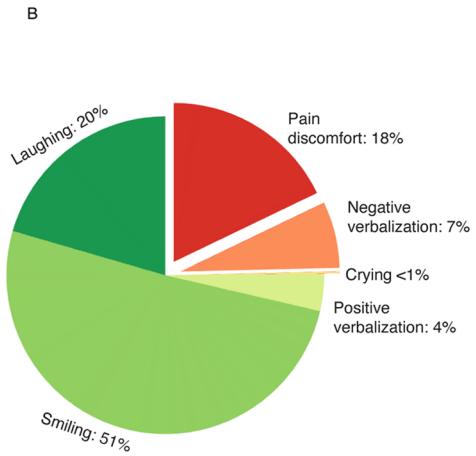
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41562-022-01310-0.

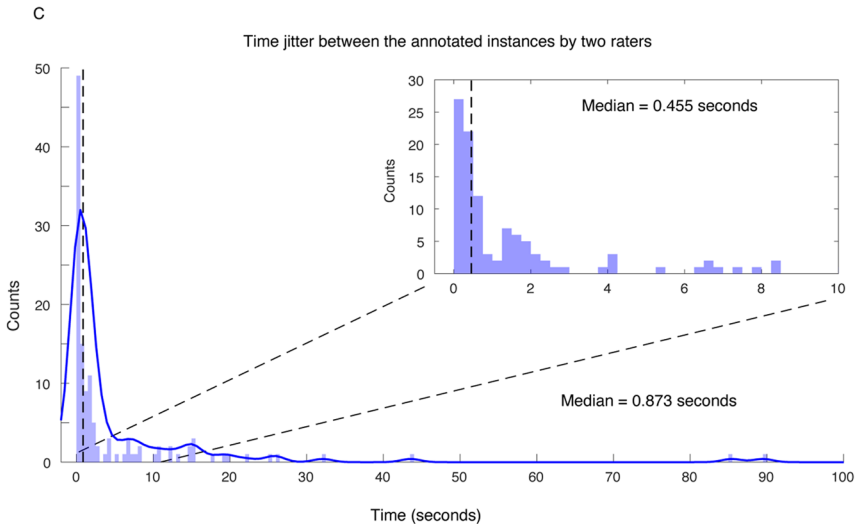
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-022-01310-0.

Correspondence and requests for materials should be addressed to Edward F. Chang.

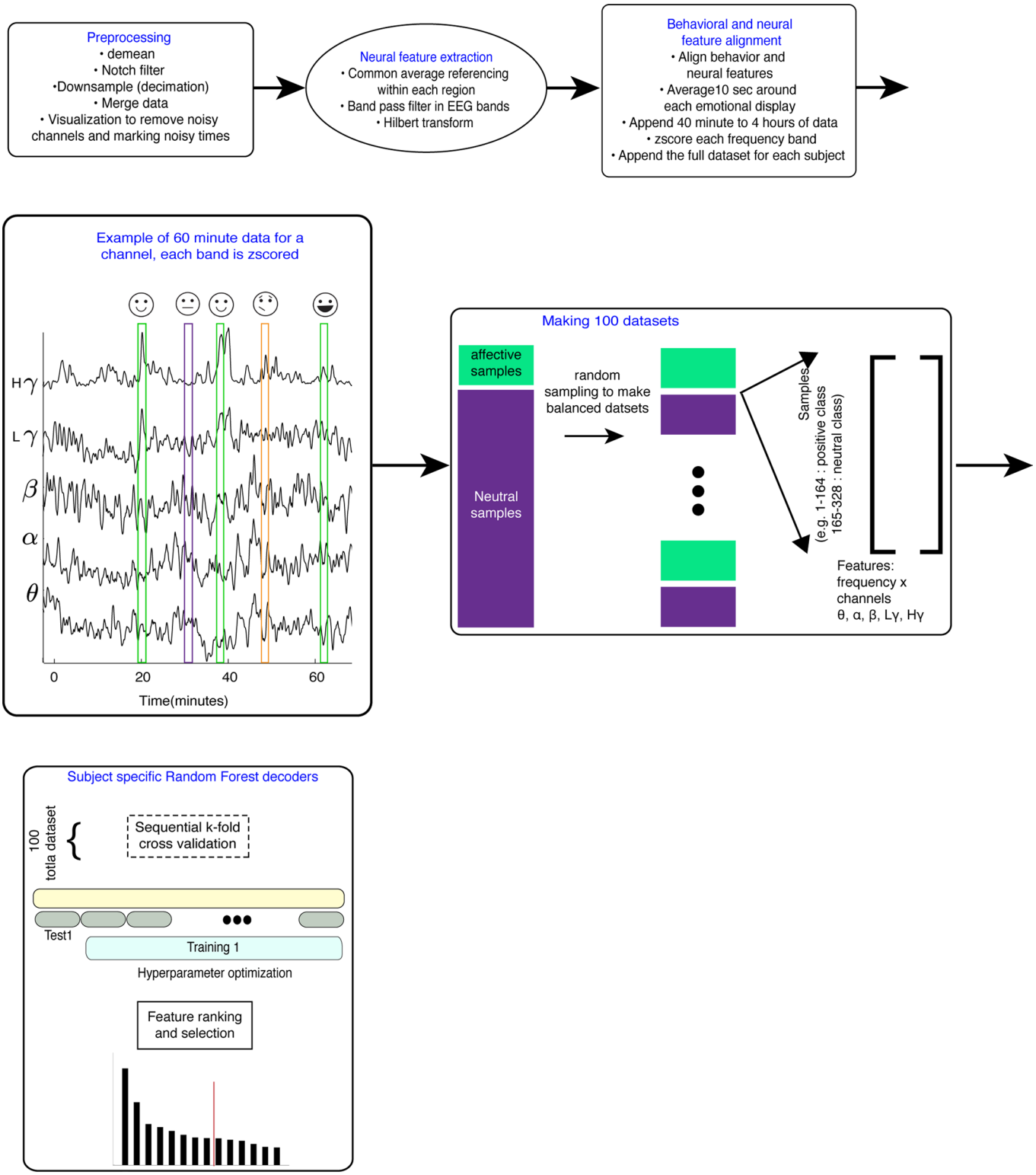
Peer review information *Nature Human Behaviour* thanks Anna Weinberg, Liang Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



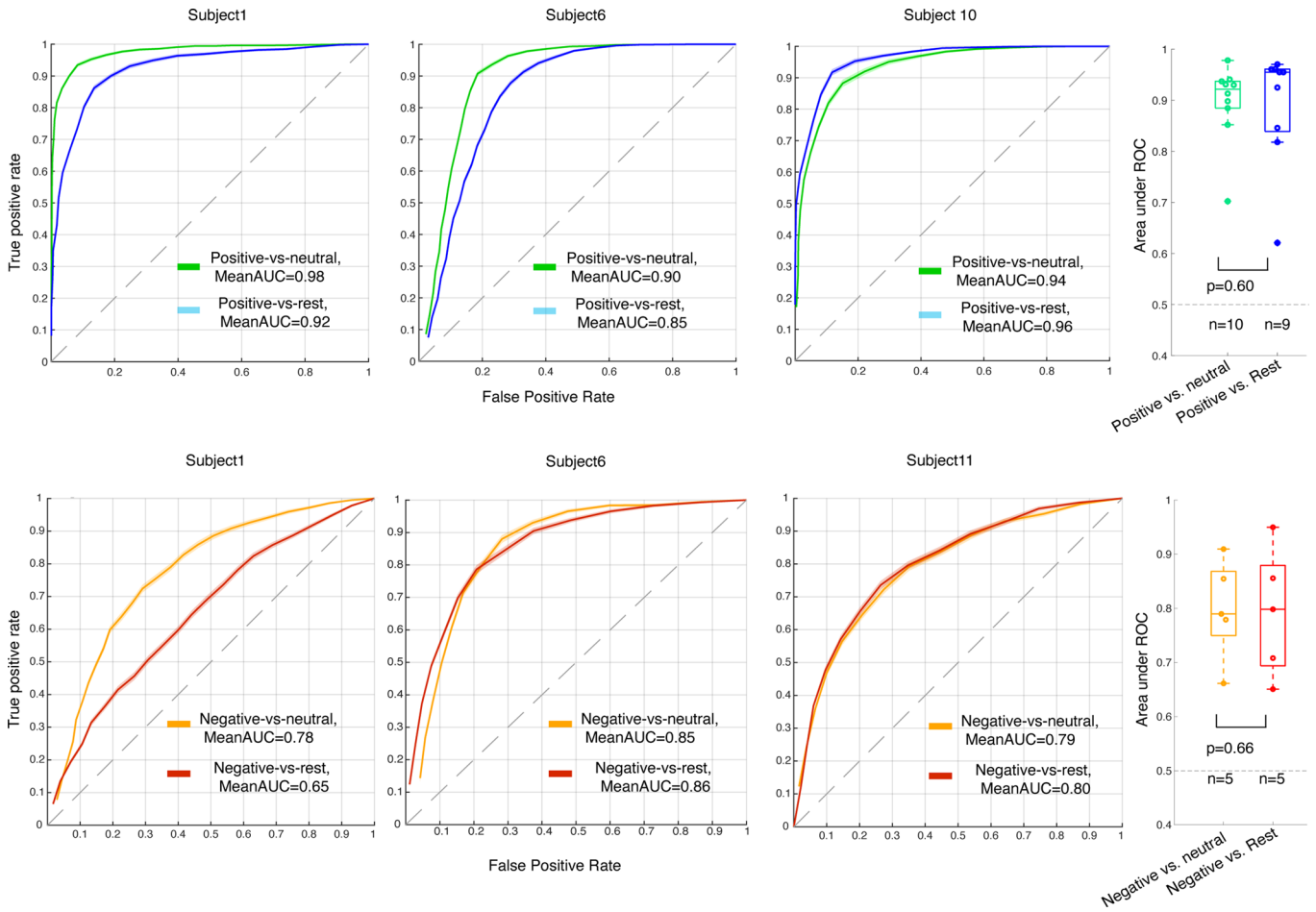




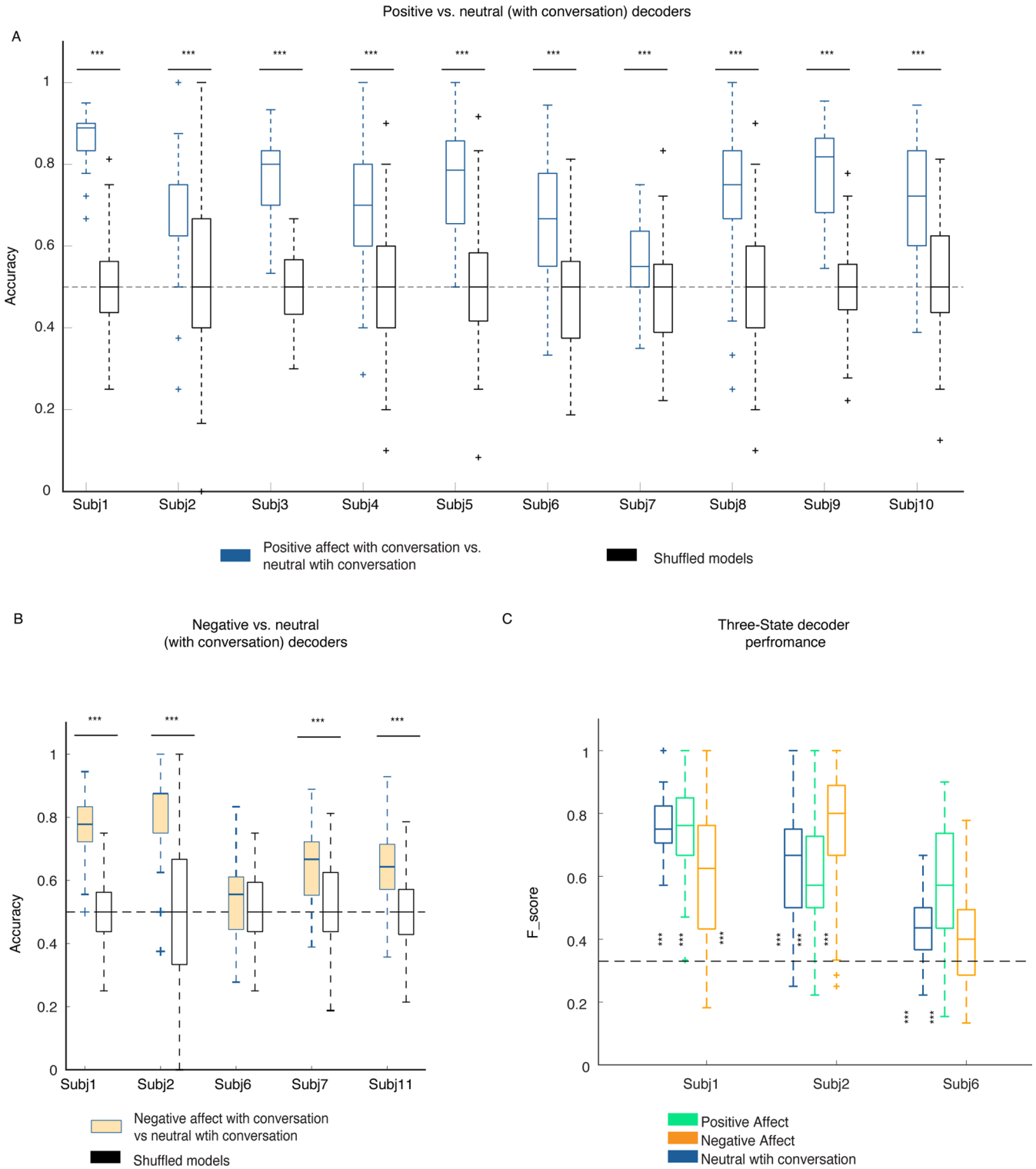
Extended Data Fig. 1 | Behavioral annotations. a) Example of annotated behaviors for an example participant during three days of their hospital stay. Behaviors in black are marked using onset and offset of the activity, while the affective behaviors are marked as instances. Purple shading represents neutral moments where there were no affective behaviors, but the participant may have been engaged in other tasks (here, using the phone). The red shading displays where there was no activity (called 'rest', per supplementary tables 1 and 2). **b**) Percentages of naturalistic affective behaviours displayed across the 11 participants in this study. **c**) distribution of the time jitter between different rater pairs for the positive and negative affective behaviours.



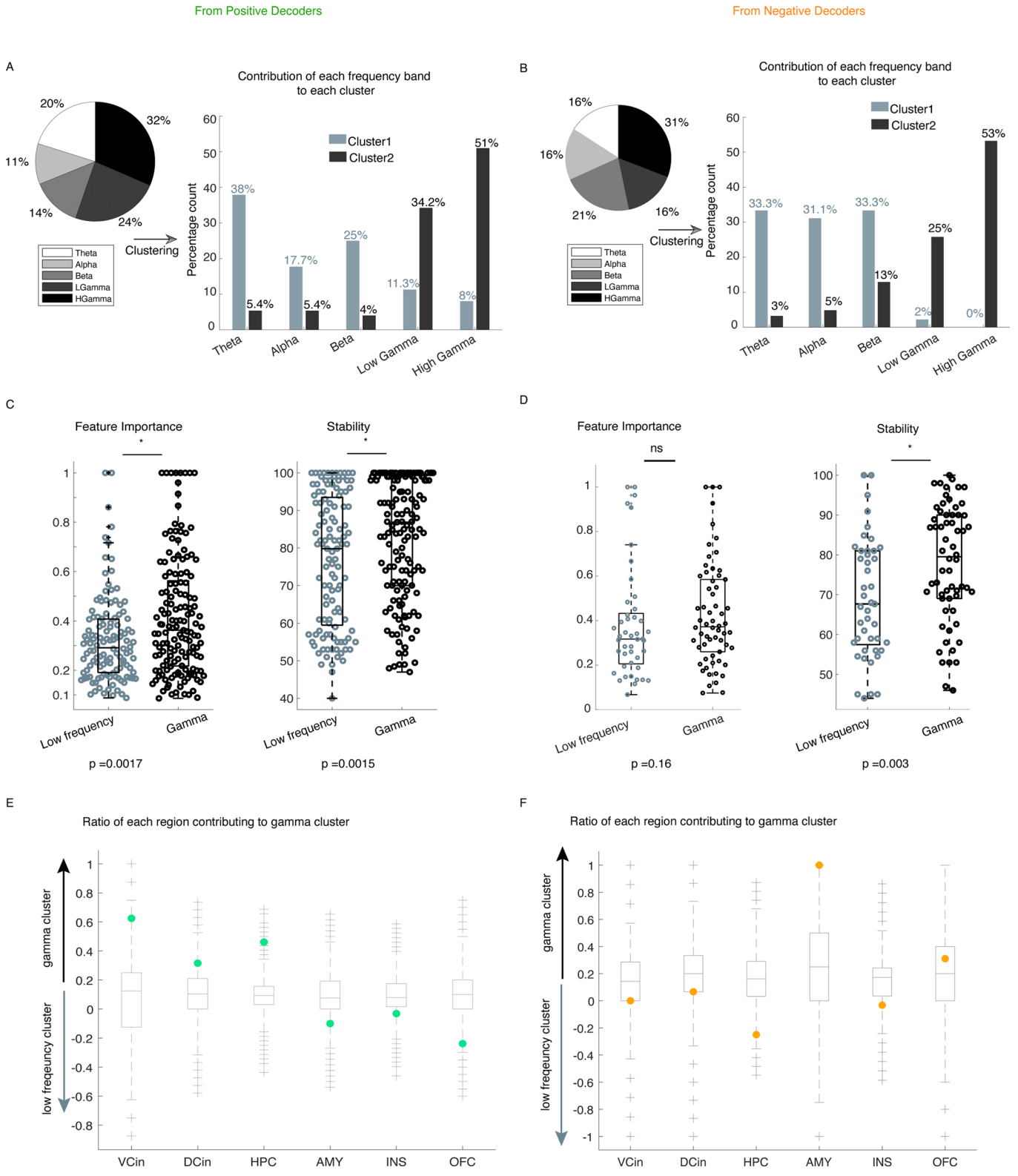
Extended Data Fig. 2 | *Preprocessing and decoding pipeline.*



Extended Data Fig. 3 | Comparison of decoder performance using rest vs. neutral moments. Decoders were trained using rest instances vs, positive (blue) and negative (red) affective behaviours. All panels are comparing these decoders with the neutral vs. affective behaviours as shown in Fig. 2. Green and orange curves show the original model AUCs for positive and negative decoders, respectively. The boxplots show the sample distribution of the average AUC for positive behaviours vs. neutral (green, n = 10 participants), and positive behaviours vs. rest (blue, n = 9 participants) in the top row and negative behaviours vs. neutral (orange, n = 5 participants) as well as negative behaviours vs. rest (red, n = 5 participants) in the bottom row. There were no significance difference between the positive ($p = 0.6$, two-sided non-parametric pairwise ranskum test) and negative ($p = 0.66$, two-sided pairwise ranksum test) decoders. In the box plots, central lines represent the median and the two edges represent 25 and 75 percentiles; whiskers show the most extreme datapoints, and outliers are shown individually (see MATLAB boxplot function).

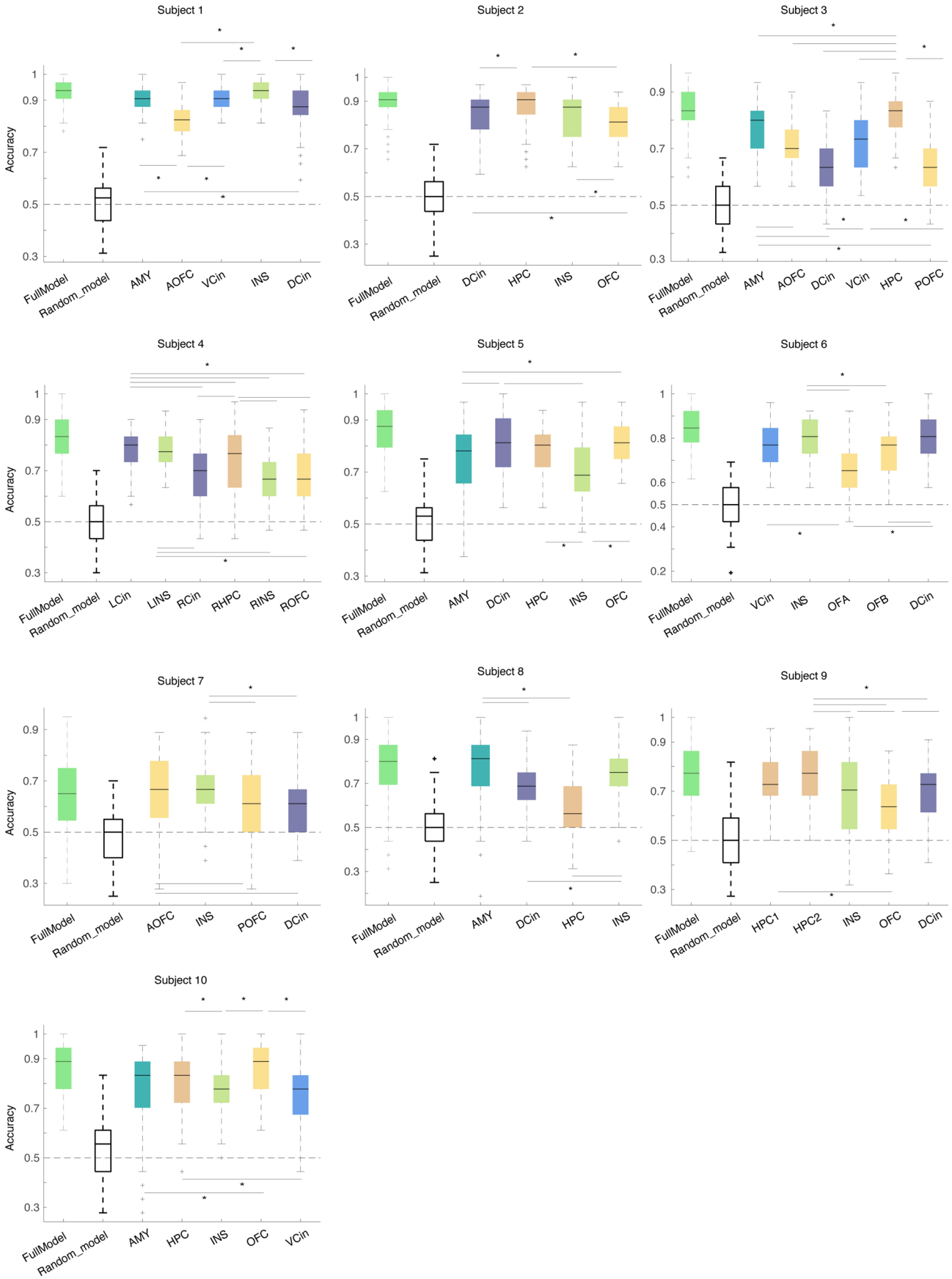


Extended Data Fig. 4 | Decoding results for neutral vs. affective behaviours that included conversational moments. a & **b**) Accuracy for all 10 and 5 participants on which the positive and negative decoders were trained, respectively. Permuted models (black) that were trained the same way using the shuffled labels across all participants. The significance level was assumed as 0.0005 to correct for n = 100 runs (refer to the Methods section 'Statistical Analyses'). P values regarding panel A are as following for all participants: $1.4 * 10^{-33}$, $5.9 * 10^{-7}$, $5.1 * 10^{-29}$, $3.3 * 10^{-16}$, $6.8*10^{-26}$, $1.6 * 10^{-13}$, $6.35 * 10^{-5}$, $2.3 * 10^{-15}$, $1 * 10^{-32}$, $2.1 * 10^{-14}$, respectively. P values regarding panel B are as following: $9.25*10^{-30}$, $9.13*10^{-27}$, 0.0031, $1.8*10^{-10}$, $2.7*10^{-11}$. **c**) F1-scores for the three-class RF models from the three participants. All F1-Scores were significantly above chance level (33%, dashed lines) and different from the shuffled models (p values are in the order of neutral, positive and negative behaviour for each participant: Subj1: $2.9*10^{-32}$, $7.1*10^{-32}$, $1.4*10^{-18}$; Subj2: $2.7*10^{-15}$: $6.9*10^{-11}$, $7.7*10^{-22}$; Subj6: $1.3*10^{-7}$, $2.1*10^{-18}$, 0.025, two-sided pairwise ranksum test). In the box plots(A-C) central lines represent the median and the two edges represent 25 and 75 percentiles, whiskers show the most extreme datapoints and outliers are shown individually (see MATLAB boxplot function). *** signifies $p < 0.0001$.



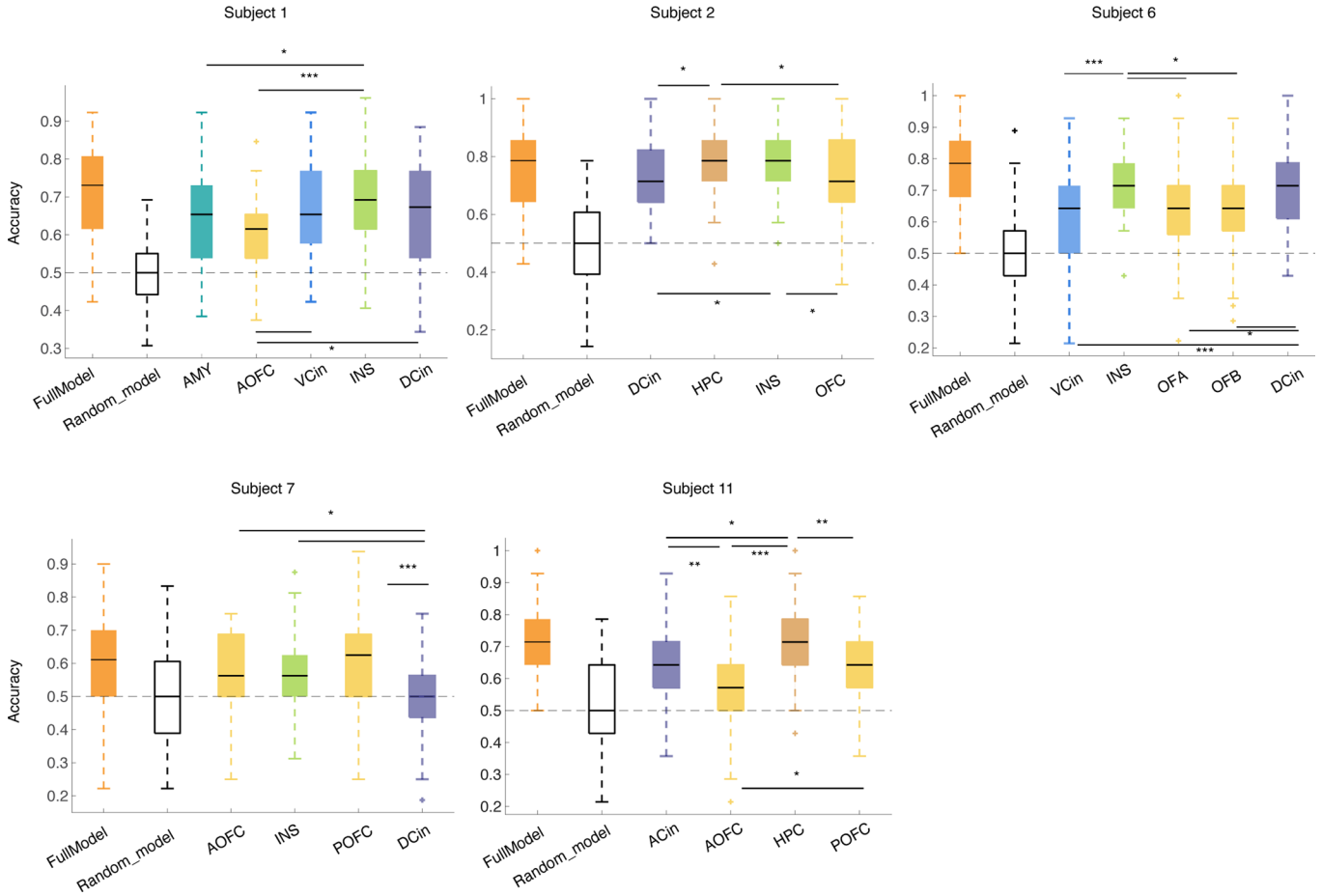
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Clustering analyses populated across all participants for the binary classifiers. a,**b**) Pie charts show the percentage of frequency bands that were selected across all participants for positive and negative decoders, respectively. The histograms show the percentage count of each frequency band within each cluster, implying that the low frequency cluster was mainly made up of the theta, alpha and beta bands. The gamma cluster was mainly made up of the high and low gamma bands for both decoder types. **c**) left and right panels show the populated normalized feature importance and the stability across all 10 participants for the positive decoders(n = 149 and n = 124 for gamma and low-frequency clusters, respectively), with p values obtained by two-sided pairwise ranksum tests at the bottom of each panel. **d**) represents similar panels as in C for negative decoders (n = 62 and n = 45 for gamma and low-frequency clusters, respectively). **e**,**f**) ratio is defined as (number of features in gamma cluster – number of feature in low frequency clyster) /total number of features contributing to both gamma and low frequency clusters (from Fig. 4b,d), positive ratio means the region have more selected features in gamma cluster and negative ratio means the region has more selected features in low-frequency cluster across subjects. INS: insula, VCin = Ventral cingulate, DCin = dorsal cingulate, AMY: amygdala, OFC = orbitofrontal cortex, HPC = hippocampus. We have generated permuted distributions (that is, null distributions) by shuffling (1000000 times) the region label of each feature and recomputing the ratio (gray boxplots). Confidence intervals are based on the t-statistics since the permuted distribution are normally distributed. All real values of the ratio shown in green(E) and orange(F) circles are outside the confidence interval of the permuted distributions. Confidence intervals in panel E are as following: VCin = [0.0908, 0.0917], DCin = [0.0914, 0.092], HPC = [0.0913, 0.0918], AMY = [0.0913, 0.0919], INS & OFC = [0.0914, 0.0919]. Confidence intervals in panel F for VCin = [0.1584, 0.1594], DCin = [0.1584, 0.1593], HPC = [0.1580, 0.1593], AMY = [0.1582, 0.1594], INS & OFC = [0.1586, 0.1592]. In the box plots(C-F) central lines represent the median and the two edges represent 25 and 75 percentiles; whiskers show the most extreme datapoints, and outliers are shown individually (see MATLAB boxplot function).

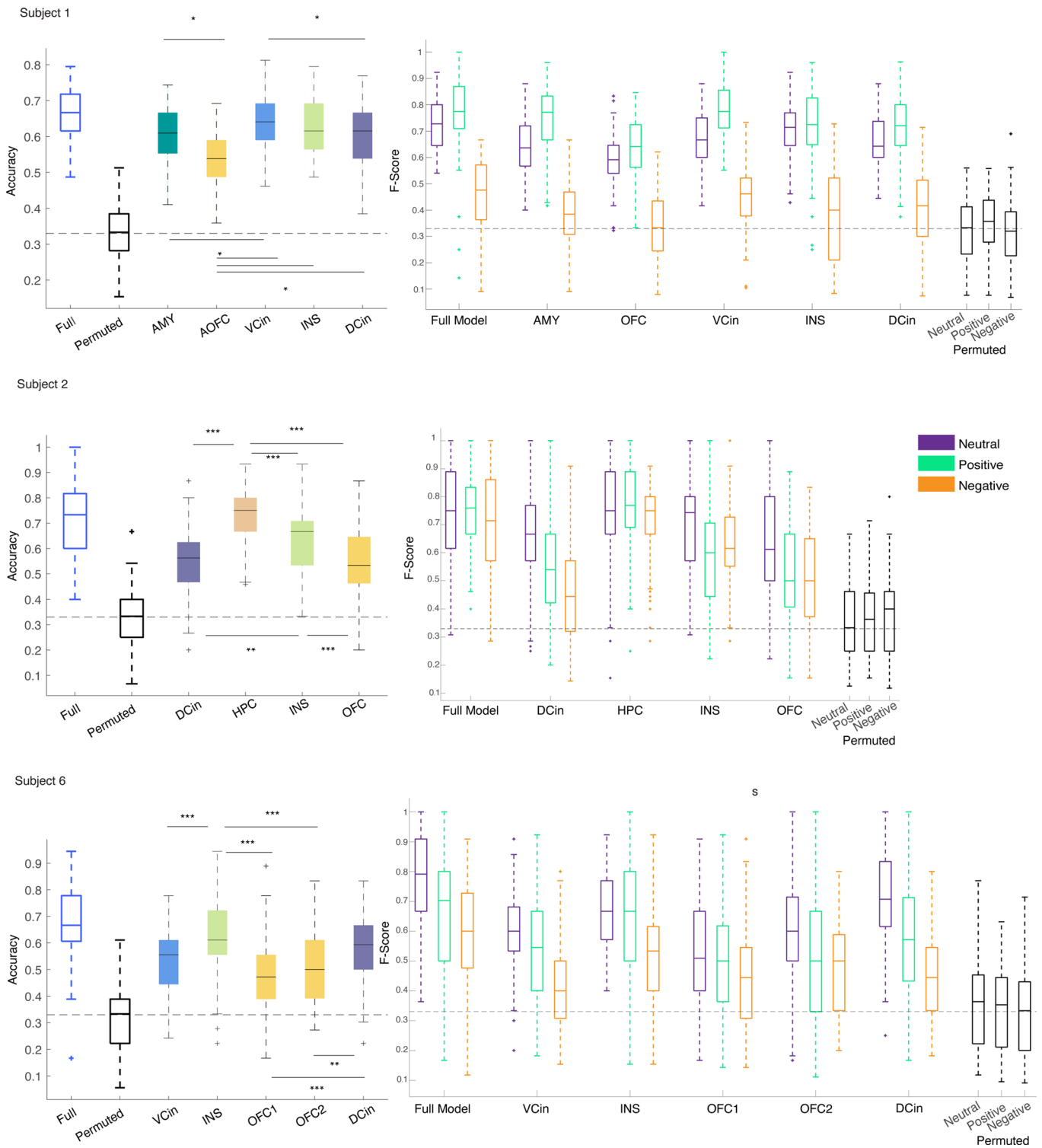


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Decoding AUC for all participants using spectral features from those contacts that are on same lead for positive vs. neutral behaviours. The green and black box plots are from the full and shuffled models across n = 100 runs as in Fig. 2-F. Other boxplots show the trained model across n = 100 datasets in which only the spectral features from each brain region were used. One-way Kruskal-wallis multi-comparison tests with Bonferroni corrections were used to examine which regions reached the highest performance (refer to supplementary table 6). OFC = orbitofrontal cortex, INS = insula, DCin = dorsal cingulate, VCin = ventral cingulate, HPC = hippocampus, AMY = amygdala. POFC = posterior OFC and AOFC = anterior OFC. In the box plots central lines represent the median and the two edges represent 25 and 75 percentiles; whiskers show the most extreme datapoints, and outliers are shown individually (see MATLAB boxplot function). *** signifies $p < 0.0001$, ** signifies $p < 0.01$ and * signifies $p < 0.05$.



Extended Data Fig. 7 | Decoding AUC for all participants using spectral features from those contacts that were on the same lead for negative vs. neutral behaviours. The orange and black box plots are from the full and shuffled models across n = 100 runs as in Fig. 2-G. Other boxplots show trained model across n = 100 datasets in which only spectral features from each brain region were used. One-way Kruskal-wallis multi-comparison tests with Bonferroni corrections were used to examine which regions reached the highest performance (refer to supplementary table 7). OFC = orbitofrontal cortex, INS = insula, DCin = dorsal cingulate, VCin = ventral cingulate, HPC = hippocampus, AMY = amygdala. In the box plots, central lines represent the median and the two edges represent 25 and 75 percentiles; whiskers show the most extreme datapoints, and outliers are shown individually (see MATLAB boxplot function). *** signifies $p < 0.0001$, ** signifies $p < 0.01$ and * signifies $p < 0.05$.



Extended Data Fig. 8 | Decoder performance of multiclass RF models run using features from each lead within a given region. Explanation of the trained models is similar as in Extended Data Fig. 7. Accuracy = number of true predicted samples / all samples. F-Score = 2*(precision*recall)/(precision + recall)). In the box plot, central lines represent the median and the two edges represent 25 and 75 percentiles; whiskers show the most extreme datapoints, and outliers are shown individually (see MATLAB boxplot function). *** signifies $p < 0.0001$, ** signifies $p < 0.01$ and * signifies $p < 0.05$.

nature research

Corresponding author(s): Edward F. Chang

Last updated by author(s): Dec 22, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The collected neural and behavioral data is a modified version of clinical recordings for purpose of seizure localization and clinical decisions. They will be made available to the lead contact
Data analysis	https://github.com/MBijanzadeh/DecodingAffect

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data used to generate the findings of this study will be available upon request to the Lead Contact.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We obtained continuous intracranial electroencephalography (iEEG) recordings from the mesolimbic network in 11 patients with epilepsy during multi-day hospitalizations. We have annotated behavioral affective moments using 24/7 audio-video recordings from subjects' hospital rooms that time-locked with the neural data. Using data driven approaches we examined the underlying neural signatures of naturalistic affective behaviors.
Research sample	Subjects included 11 patients (6 females, 5 males, age: 20-43, Table S2), who has been diagnosed with treatment-resistant epilepsy and were undergoing (iEEG) implantation for seizure localization.
Sampling strategy	Subjects that accepted to provide informed consent, or at least had implanted electrodes in 3/5 regions, or the number of behavioral emotional expressions was sufficient enough to train the RF classifiers were included in the study. We assessed the statistical significance of all models by training surrogate random forest models after shuffling the categorical labels within each fold of each dataset (to keep the balance between affective classes). All p-values were computed using non-parametric ranksum test between pairs of distribution for all pairwise statistical test mentioned in the main text or Kruksal-wallis multiple comparison test followed by Bonferroni correction for multiple groups. Please refer to the "methods" section.
Data collection	Over multiple days of monitoring, subjects underwent continuous 24-hour audio, video recording and iEEG monitoring through the Natus clinical recording system as a part of routine clinical care. Electrophysiological data were collected at sampling rates at either 512 Hz or 1024 Hz. All mesolimbic structures were sampled by subdural grid, Ad-Tech 4-contact strip and Ad-Tech 4/10-contact depth electrodes (10mm or 6mm center to center spacing). Trained human raters (11 total) manually annotated these recordings during instances of behavioral and emotional expression. In general, the annotations started two days post-electrode implantation – typically after patients recovered from the implant surgery. As a part of their review, human raters imported the video recordings into ELAN software, a linguistic ethnographic software, and used a custom template to mark individual activities and emotional states the patient engaged in throughout these continuous recordings . For more information please refer to the methods section of the paper.
Timing	The data has been collected over the course of 4 years from epileptic patients. The decoders were trained within each subject. The timing gaps is not a factor in this study.
Data exclusions	Subjects that did not provide informed consent, or at least had implanted electrodes in 3/5 regions, or the number of behavioral emotional expressions were not sufficient enough to train the RF classifiers were excluded in the study.
Non-participation	The nature of the study and patient inclusion was defined after subjects were dispatched from the hospital. Please refer to sampling strategy and Data exclusion. Thus subject drop out was not applicable in this study.
Randomization	We have included 10 and 5 subjects for training within subject positive and negative decoders, because the available number of annotated behaviors were more frequent for positive affective behaviors (e.g. smiling) than negative affective behaviors . Please refer to the method and results section of the study. We have shuffled the labels regarding affective behaviors to generate surrogate models for statistical test. Please refer to the methods section.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	"See above"
Recruitment	"See above" . Please refer to the sampling Strategy
Ethics oversight	All procedures were approved by the University of California, San Francisco Institutional Review Board. All subjects gave written informed consent to participate in the study prior to surgery.

Note that full information on the approval of the study protocol must also be provided in the manuscript.