Science Advances
AAAS

Supplementary Materials for

When self comes to a wandering mind: Brain representations and dynamics of self-generated concepts in spontaneous thought

Byeol Kim Lux *et al.*

Corresponding author: Choong-Wan Woo, waniwoo@g.skku.edu

Sci. Adv. **8**, eabn8616 (2022)
DOI: 10.1126/sciadv.abn8616

This PDF file includes:

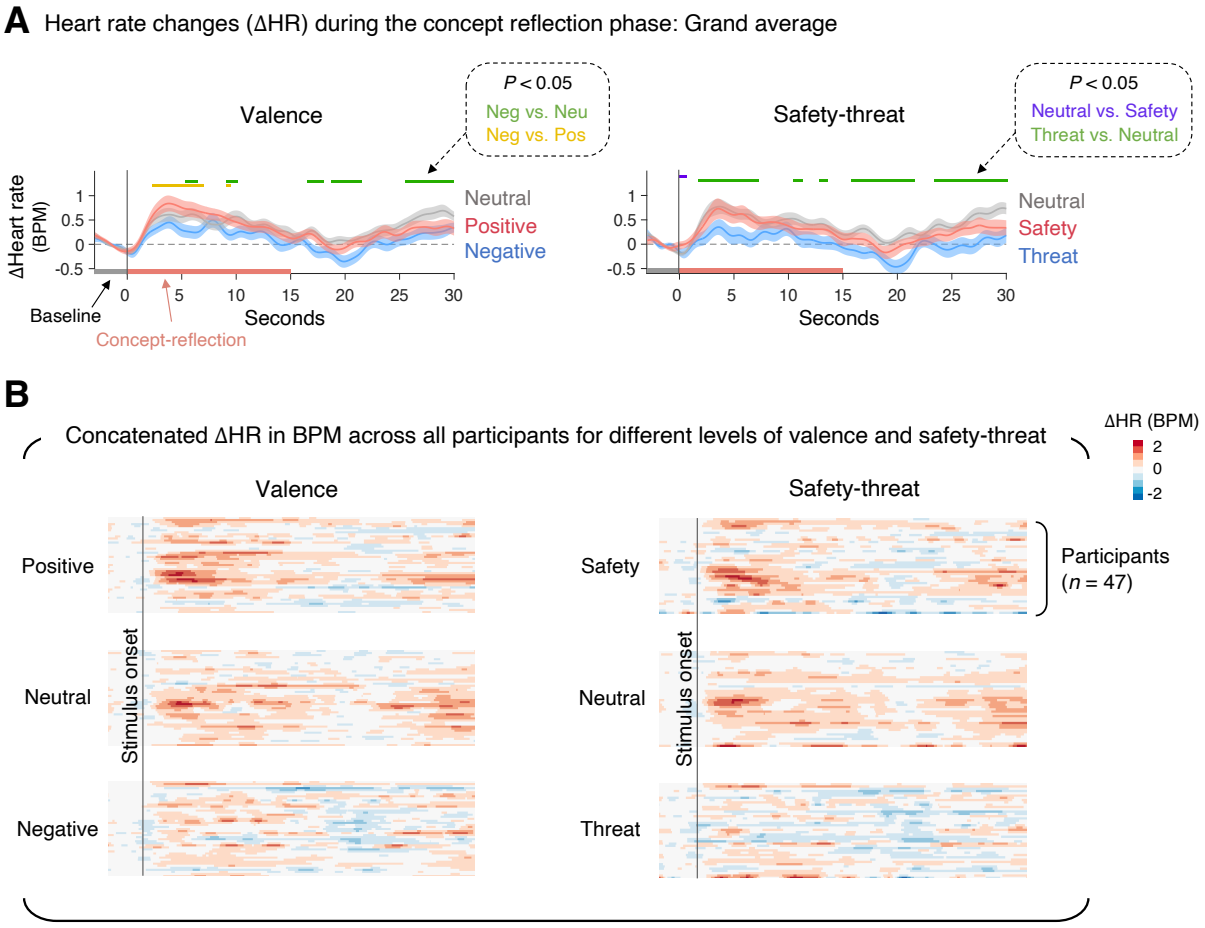


Fig. S1. Valence-induced heart rate changes during the concept reflection phase.

(A) The plots show the time-course of grand-average heart rate changes (ΔHR; in beats per minute [BPM]) during the concept reflection phase. Shading represents the standard error of the mean (s.e.m.) across participants. **(B)** The heat maps show the concatenated grand-average ΔHR across all participants for different levels of valence and safety-threat.

Main finding: Multiple time-points showed significant decreases in heart rate for the negative vs. neutral (or positive) and threat vs. neutral contrasts.

Methods. Electrocardiogram (ECG) activity was recorded using MR-compatible electrodes (Biopac Systems, Goleta, CA) placed on the right and left clavicles and lower left abdomen area. We analyzed ECG data from the first session of the fMRI experiment (n = 62). We discarded 15 participants' data: 5 participants due to recording-related issues and 10 participants due to abnormal BPM ranges caused by MR-related noise. Thus, we analyzed data from 47 participants. The ECG data were sampled at 2000Hz during the scan. We removed MR-induced noise from the ECG data with a band-pass filter (0.6 ~ 10 Hz) and a comb filter with a multiple of a reciprocal number of TR (1/0.46 Hz). Next, we used the PhysIO Toolbox (*73*) (https://www.nitrc.org/projects/physio/) to find peaks and calculate the inter-beat interval (IBI). The IBI data were down-sampled (25 Hz) and low-pass filtered (0.5 Hz). We then calculated

beats per minute (BPM) by dividing 60 seconds by the IBI (60/IBI). To examine the heart rate changes induced by the levels of valence and safety-threat scores, we divided the data into three groups—positive, neutral, and negative for valence, and safety, neutral, and threat for safety-threat—using 0.33 and -0.33 as the boundaries for defining discrete states. We grand-averaged the HR data using 3 seconds before the onset as the baseline and 30 seconds after the onset as an epoch. We conducted paired *t*-tests for each time point, and the green, yellow, and purple dots (or lines) in the plots in **(A)** show the time points that yielded significant *t*-test results, uncorrected $P < 0.05$, two-tailed, paired *t*-test, $n = 47$.

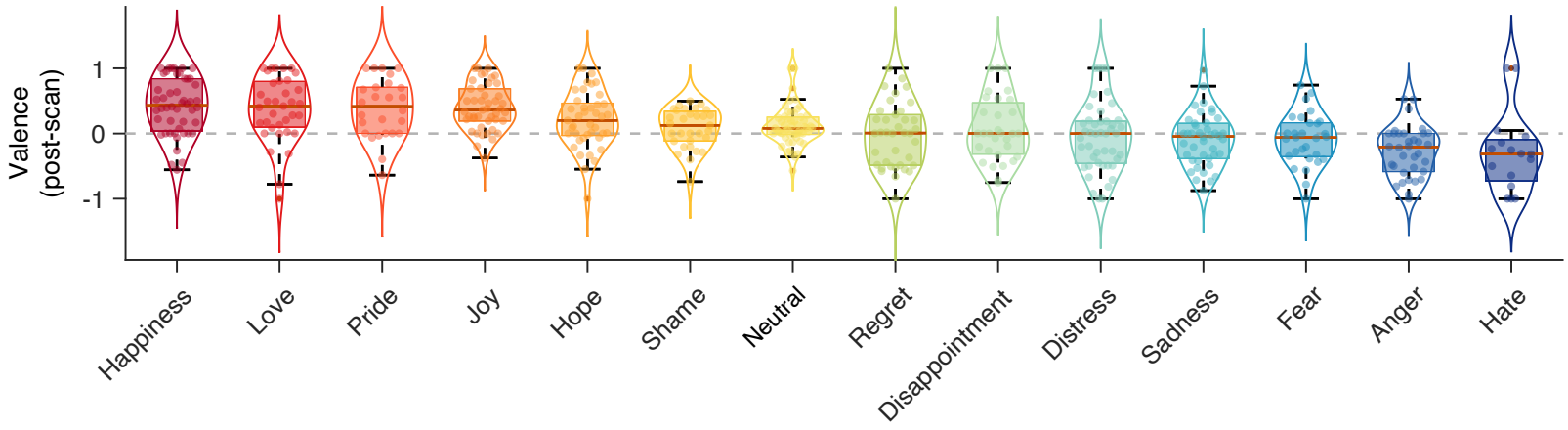


Fig. S2. The relationship between the in-scanner emotion ratings and post-scan valence ratings.
The plot shows the distribution of valence ratings (from the post-scan survey) for different emotion categories of self-generated concepts (n = 62). The emotion category data were collected from intermittent emotion ratings during the concept reflection task inside the scanner. For emotion ratings, we intermittently displayed 14 emotion words on the screen after 15 seconds of concept presentation, and asked participants to select one emotion descriptor closest to their current emotion. We obtained these emotion ratings five times per run. The box was bounded by the first and third quartiles, and the whiskers stretched to the highest and lowest values within median ± 1.5 interquartile range.

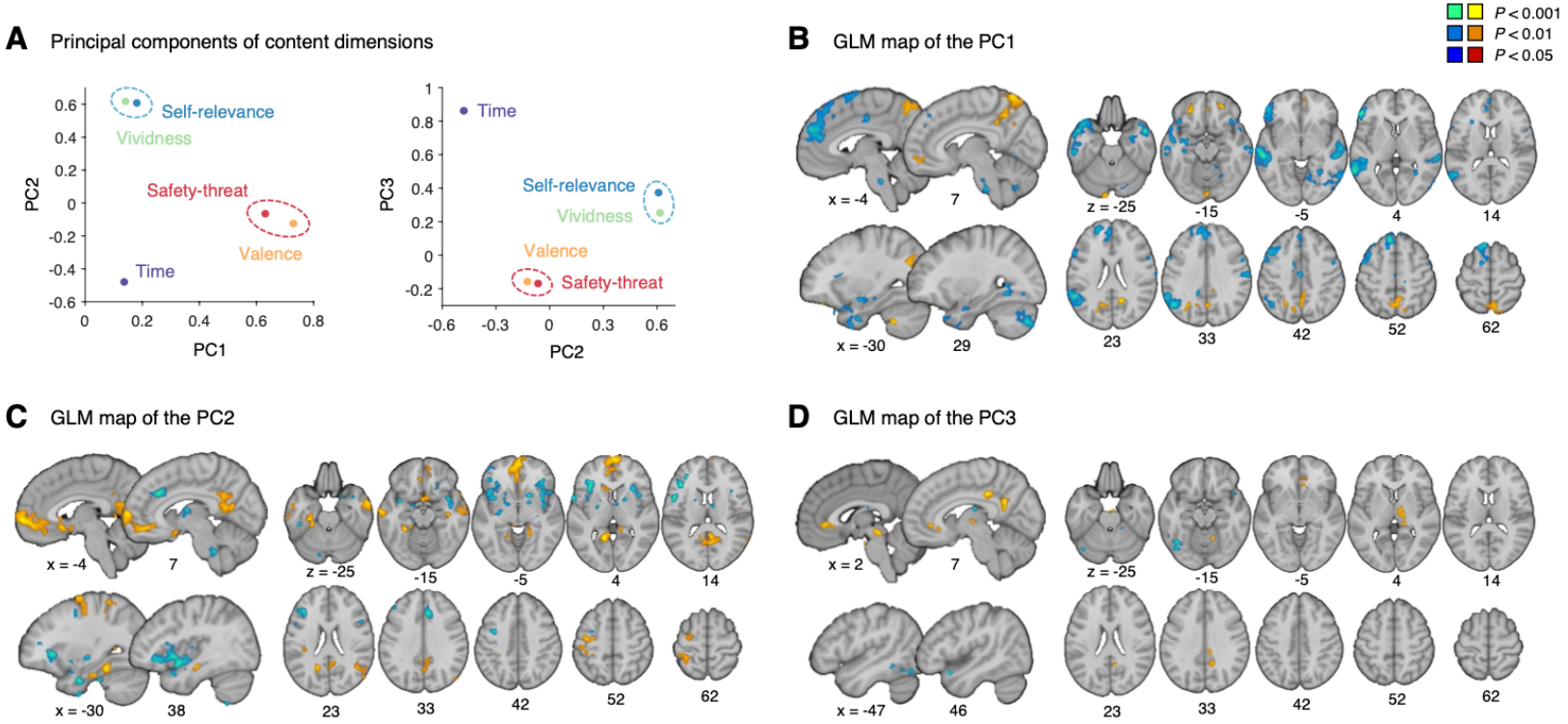


Fig. S3. Principal components of the content dimensions and their univariate general linear model maps.
(A) We conducted the principal component analysis on the five content dimensions (n = 61) and plotted three principal components (PCs), which explained 89.4% of the total variance. The valence and safety-threat dimensions showed high loadings on the first PC, while the self-relevance and vividness dimensions showed high loadings on the second PC. The time dimension showed a high loading on the third PC. **(B)-(D)**, The general linear model (GLM) results for the three principal components. To identify brain regions correlated with the principal component scores, we regressed the single-trial images on the three principal component scores for each participant. We then performed a one-sample t-test on the 61 beta images (one beta map per participant) and thresholded the map with uncorrected $P < 0.001$, two-tailed, and pruned the results using two additional, more liberal thresholds, $P < 0.01$ and $P < 0.05$, two-tailed.

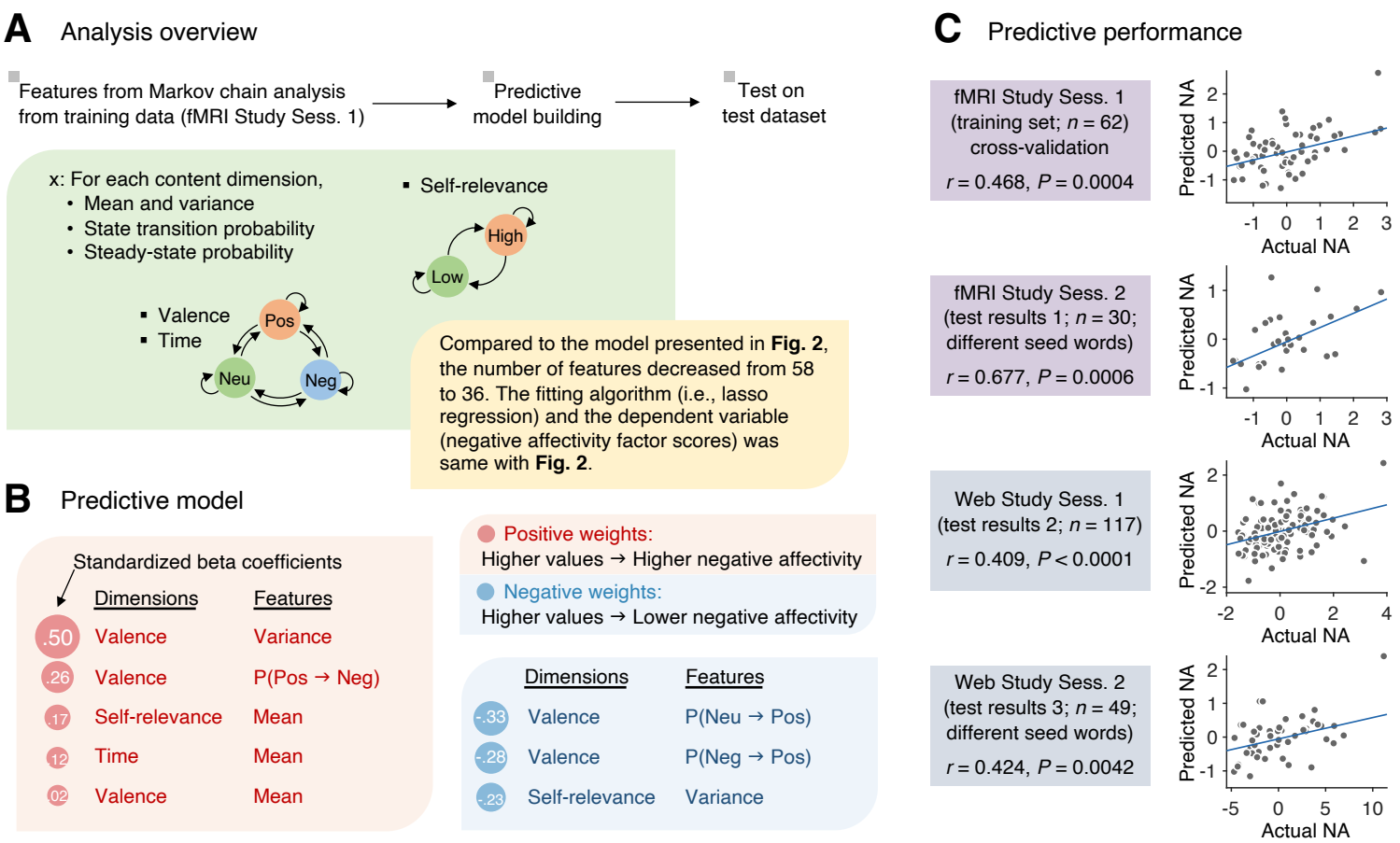


Fig. S4. Markov chain-based predictive model of negative affectivity based on valence, self-relevance, and time (VST model). **(A)** We trained an additional Markov chain-based predictive model of general negative affectivity with the valence, self-relevance, and time dimensions only to test whether these three dimensions were enough to predict the level of general negative affectivity. Other analysis procedures were identical to the main Markov chain-based predictive model with all the five content dimensions (**Fig. 2**). The total number of input features decreased from 58 to 36. **(B)** A total of 8 features were selected. All features except for "valence-mean" overlapped with the selected features in the original model. **(C)** VST model performance. From top to bottom, the plots show 1) the leave-one-participant-out cross-validated prediction results within the training dataset (n = 62, first session of the fMRI study), and

three independent test results on 2) the second session re-test data of the fMRI study with different seed words (n = 30), 3) the first session data of the FAST-web study (n = 117), and 4) the second session re-test data of the FAST-web study with different seed words (n = 49). The actual versus predicted negative affectivity factor scores are shown in the plots. Each dot represents each participant. We evaluated the model performance with robust correlation between the actual and predicted levels of general negative affectivity.

Main findings: The VST model also showed significant predictions across four datasets, and seven out of eight final features of the VST model overlapped with the original full model (**Fig. 2C**).

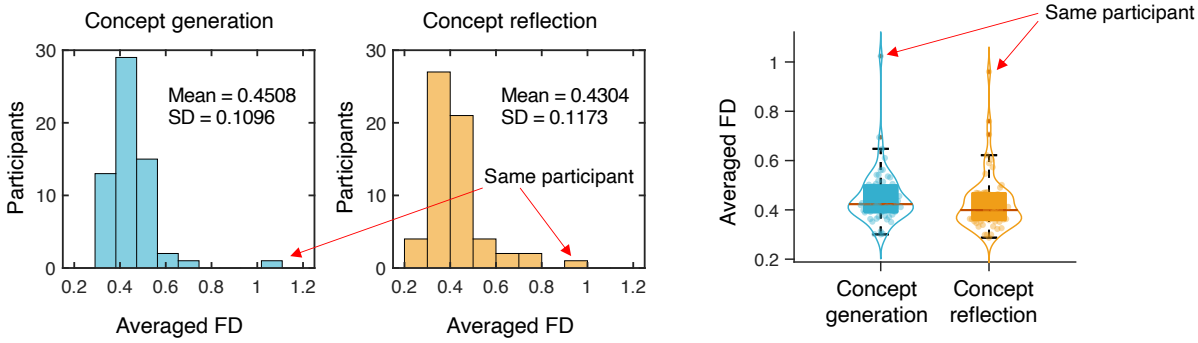


Fig. S5. Framewise displacement during the concept generation and reflection phases.
We investigated head motion during the concept generation and reflection phases with framewise displacement (FD). We averaged the TR-level framewise displacement values separately for concept generation and concept reflection runs. Both the histogram and violin plots show the averaged FD values. The histogram plots show the distribution of the averaged FD values, whereas the violin plots show the median and outliers more clearly. In the violin plot, each dot represents each individual's averaged FD value.

Though we originally predicted that the concept generation phase would show a higher level of head motion due to speaking, the result did not show a significant FD difference between the phases, $t_{60} = 0.996$, $P = 0.322$, two-tailed, paired t-test. This smaller-than-expected difference in motion suggests that it would be possible to analyze the fMRI data from the concept generation phase, which is an exciting avenue for future studies. However, while FD values were not different between the phases, speaking-induced systematic motion effects in the fMRI data may still exist, which should be carefully examined before analysis of the concept generation phase data.

In addition, we found one outlier participant with high FDs in both phases. To examine the effects of the outlier participant on our main findings, we conducted fMRI pattern-based predictive modeling of valence and self-relevance again. After removing the outlier participant, we obtained results similar to our original findings, suggesting that our main findings are robust to the motion outlier. The prediction performance after versus before removing the outlier participant is as follows.

After removing the outlier participant, predicting with leave-one-subject-out cross-validation:
valence with high self-relevant trials: mean $r = 0.009$, $z = 0.120$, $P = 0.9046$
valence with low self-relevant trials: mean $r = 0.343$, $z = 4.406$, $P < 0.0001$
self-relevance: mean $r = 0.281$, $z = 4.033$, $P < 0.0001$

Before removing the outlier participant, predicting with leave-one-subject-out cross-validation (as reported in the original manuscript):
valence with high self-relevant trials: mean $r = 0.031$, $z = 0.403$, $P = 0.6872$
valence with low self-relevant trials: mean $r = 0.307$, $z = 3.808$, $P < 0.0001$
self-relevance: mean $r = 0.304$, $z = 4.400$, $P < 0.0001$

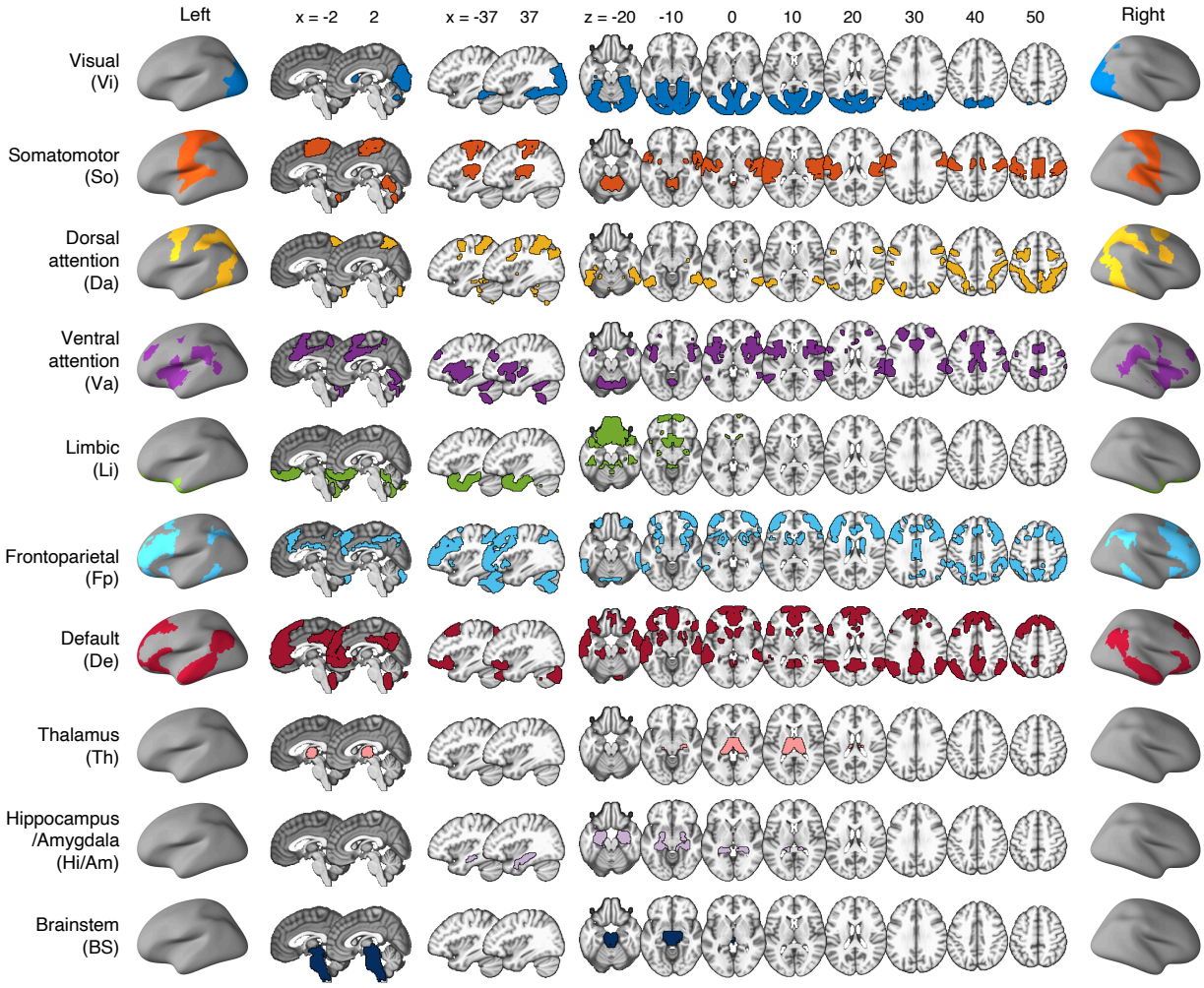


Fig. S6. Large-scale functional networks and regions for the radial network plots.

To make the radial plots in the main figures (**Figs. 3, 4,** and **6),** we used Buckner's group parcellations to define large-scale functional brain networks, including 7 networks within the cerebral cortex (*69*), cerebellum (*70*), and basal ganglia (*71*). We also added the thalamus, hippocampus, and amygdala from the SPM anatomy toolbox, as well as the brainstem region.

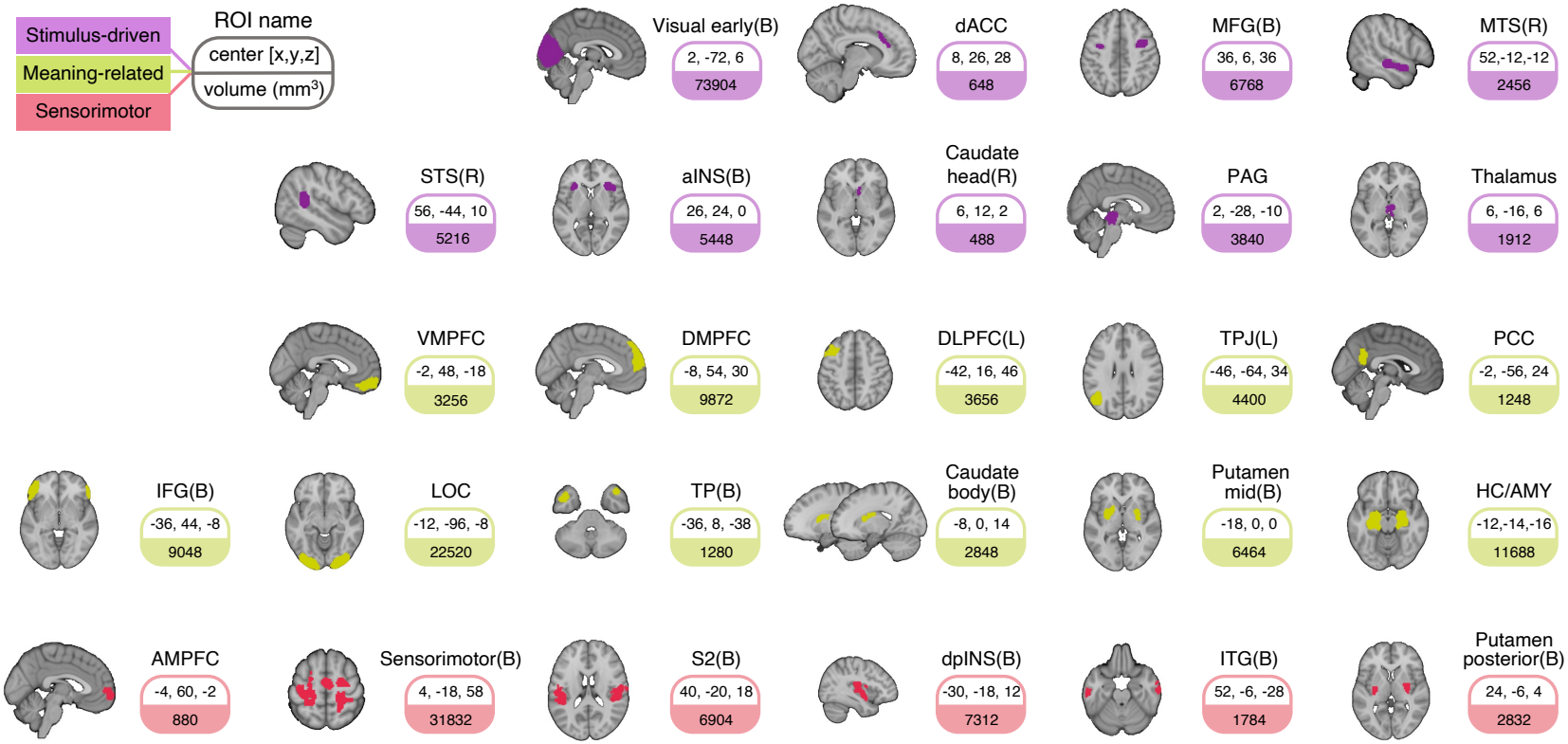


Fig. S7. 26 main regions-of-interest (ROIs) from the basic contrast map of concept reflection.

dACC, dorsal anterior cingulate cortex; MFG(B), bilateral middle frontal gyrus; MTS(R), right middle temporal sulcus; STS(R), right superior temporal sulcus; aINS(B), bilateral anterior insula; PAG, periaqueductal gray; VMPFC, ventromedial prefrontal cortex; DMPFC, dorsomedial prefrontal cortex; DLPFC(L), left dorsolateral prefrontal cortex; TPJ(L), left temporal parietal junction; PCC, posterior cingulate cortex; IFG(B), bilateral inferior frontal gyrus; LOC, lateral occipital cortex; TP(B), bilateral temporal pole; HC/AMY, hippocampus/amygdala; AMPFC, anterior medial prefrontal cortex; S2(B), bilateral secondary somatosensory cortex; dpINS(B), bilateral dorsal posterior insula; ITG(B), bilateral inferior temporal gyrus.

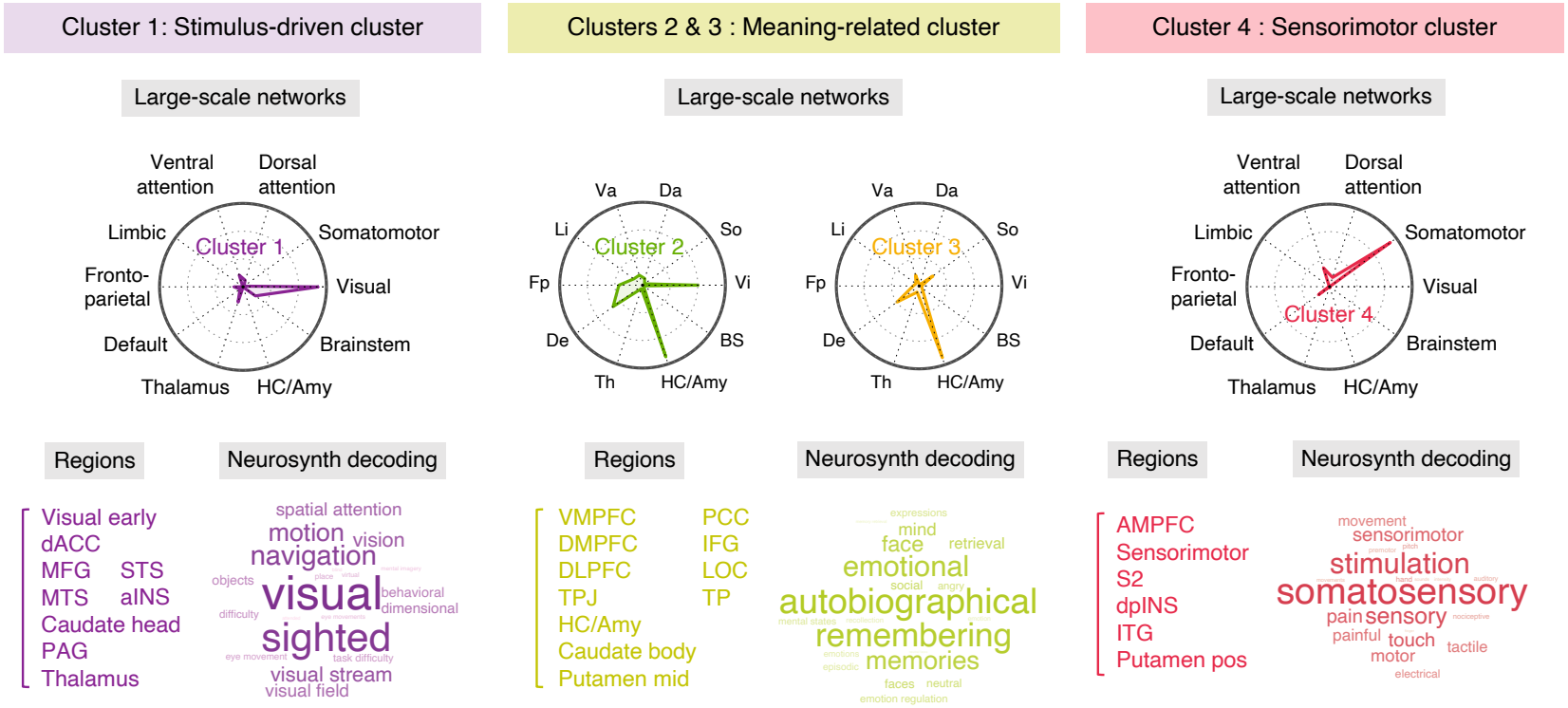


Fig. S8. Understanding clusters with large-scale networks and Neurosynth decoding.

To interpret the functional meaning of the clusters, we examined the clusters with large-scale functional networks and a term-based decoding analysis based on an automated large-scale meta-analysis, Neurosynth (*74*). The first clusters showed the largest correlations with the meta-analytic maps with functional terms including "visual," "sighted," and "navigation." The second and third cluster showed the largest correlations with the terms "autobiographical," "remembering," and "emotional." The fourth cluster was correlated with the terms "somatosensory," "stimulation," and "sensory." Based on the decoding results, we combined and named the second and third clusters to be "meaning-related" because many of the brain regions within these clusters have been shown to be involved in semantic processing and autobiographical memory. In addition, we named the first cluster "stimulus-driven," and the fourth cluster as "sensorimotor," respectively. For the radial plot, Va, ventral attention; Da, dorsal attention, So, somatomotor; Vi, visual; BS,

brainstem; HC/Amy, hippocampus/amygdala; Th, thalamus; De, default; Fp, frontoparietal; Li, limbic. For the brain regions, aINS, anterior insula; dACC, dorsal anterior cingulate cortex; MFG, middle frontal gyrus; MTS, middle temporal sulcus; STS, superior temporal sulcus; PAG, periaqueductal gray; VMPFC, ventral medial prefrontal cortex; ; DMPFC, dorsal medial prefrontal cortex; DLPFC, dorsolateral prefrontal cortex; PCC, posterior cingulate cortex; IFG, inferior frontal gyrus; LOC, TPJ, temporal parietal junction; lateral occipital cortex; S1/S2, primary and secondary somatosensory cortex; dpINS, dorsal posterior insula; AMPFC, anterior medial prefrontal cortex; ITG, inferior temporal gyrus. For the locations of each region, please see **fig. S7.**

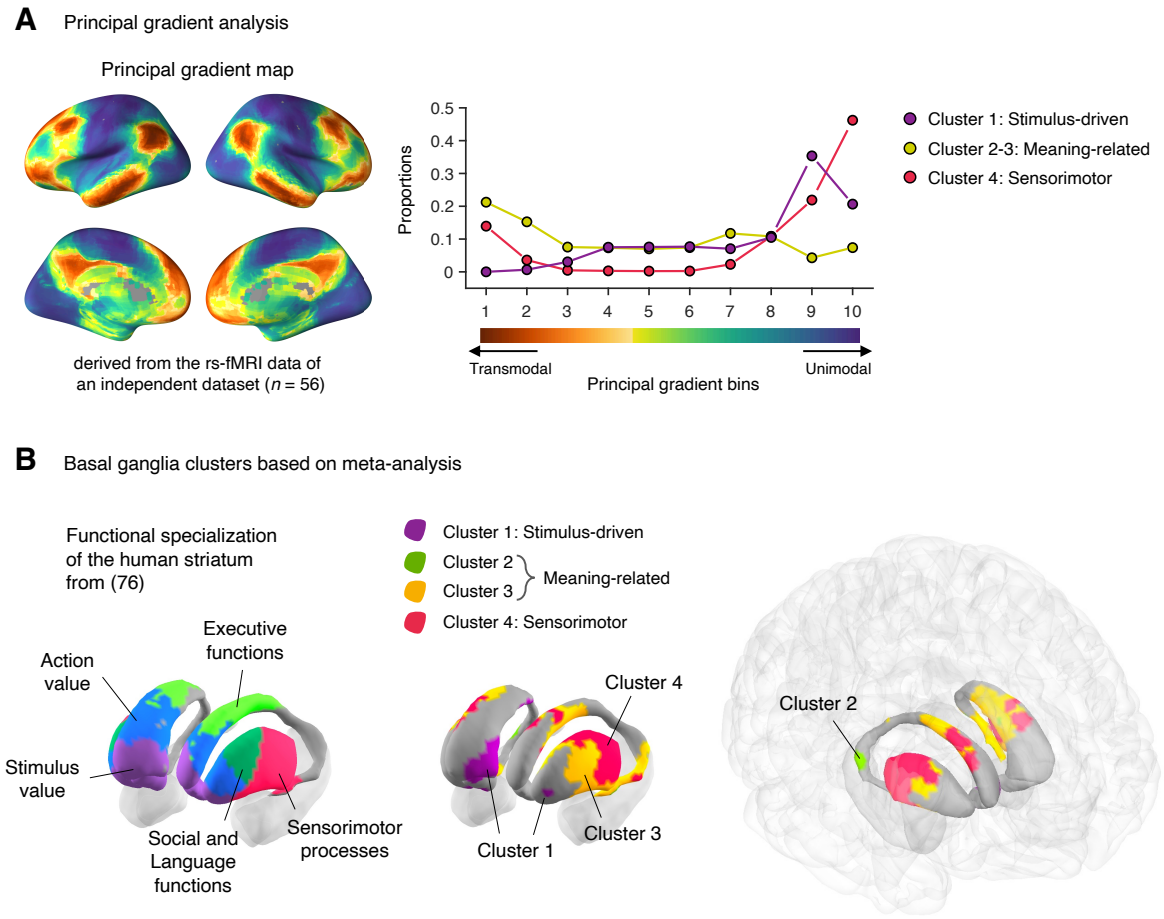


Fig. S9. Neurobiological assessment of clustering results.

We assessed whether our clustering results and their naming (shown in **fig. S8**) were neurobiologically and functionally meaningful rather than arbitrary. In addition to the term-based decoding analysis reported in **fig. S8**, here, we interpreted our clustering results with the principal gradient (*75*) and the meta-analytic basal ganglia parcellations (*76*). (**A)** The brain map on the left shows the principal gradient from unimodal to transmodal brain regions across the whole brain (*75*). We re-calculated the principal gradient map using our resting-state dataset (n = 56; 7-min resting scan) to create a volumetric principal gradient image that includes the subcortical regions. The plot on the right shows the proportions of overlapping voxels between the 10-bin maps of the principal gradient and our three clusters. (**B)** The basal ganglia map on the left shows the functional parcellations based on meta-analysis (*76*). The figures in the middle and right panels show our clustering results mapped onto the basal ganglia.

Main findings: Our region clustering and subsequent naming were largely consistent with the principal gradient in the cortex and the meta-analysis findings in the basal ganglia. For example, the "meaning-related" cluster largely overlapped with the transmodal end in the principal gradient of cortical hierarchy and the parts of the basal ganglia related to social, language, and executive functions. The "stimulus-driven" and "sensorimotor" clusters overlapped with the unimodal end of the cortical principal gradient and the basal ganglia parcellations for stimulus

value and sensorimotor processes, respectively. These findings support that our clustering analysis resulted in neurobiologically meaningful clusters, providing a basis for further analyses of our data and functional interpretations of our findings.

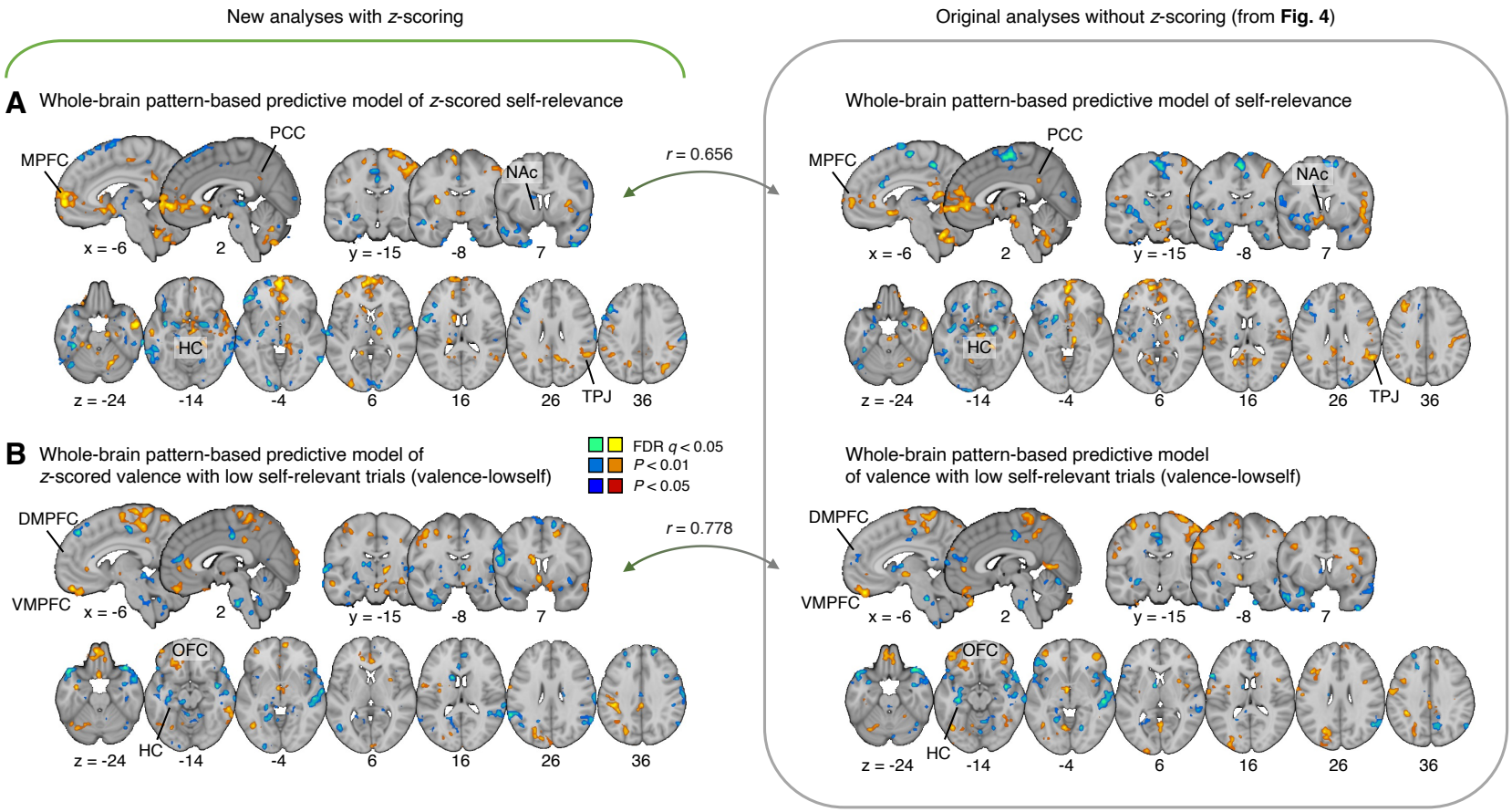


Fig. S10. Comparing predictive models of *z*-scored outcome variables with the original results from Fig. 4.

To examine the potential influences of individual differences in the scale use on the predictive modeling results, we additionally conducted the same fMRI-based predictive modeling analyses (as in **Fig. 4**) with the *z*-scored content dimension scores. To this end, we *z*-scored the outcome variables (i.e., self-relevance and low-self valence scores) before grouping the data into quartiles, and then we trained the same models in **Fig. 4** (i.e., predictive models of self-relevance and valence in low-self trials). The prediction results indicated that the two models (i.e., without vs. with *z*-scoring) showed similar prediction performances—for the self-relevance model,

the correlation between actual and predicted ratings with leave-one-subject-out cross-validation (LOSO-CV) was mean $r = 0.286$, $z = 4.440$, $P < 0.0001$, two-tailed bootstrap test, mse = 0.669 (cf. the previous model performance was mean $r = 0.304$, $z = 4.400$, $P < 0.0001$, mse = 0.155; please note that the mean squared error [mse] is different as the mse is scale-dependent), and for the valence model trained on the low self-relevance trials (named 'valence-lowself' model), mean $r = 0.316$, $z = 4.5695$, $P = 0.0001$, two-tailed bootstrap test, mse = 0.630 (cf. the previous model performance was mean $r = 0.307$, $z = 3.808$, $P < 0.0001$, mse = 0.362). In addition, the weight maps between the two models showed highly similar patterns. The bootstrap tests of the predictive weights identified similar regions as significant (at FDR $q < 0.05$), and the spatial correlations between two maps (i.e., without vs. with z-scoring) were high—for the self-relevance model, the spatial correlation between the two weight maps was 0.656, and for the valence-lowself model, the spatial correlation was 0.778. These results suggest that the individual differences in the scale use were not an important factor for our main findings.

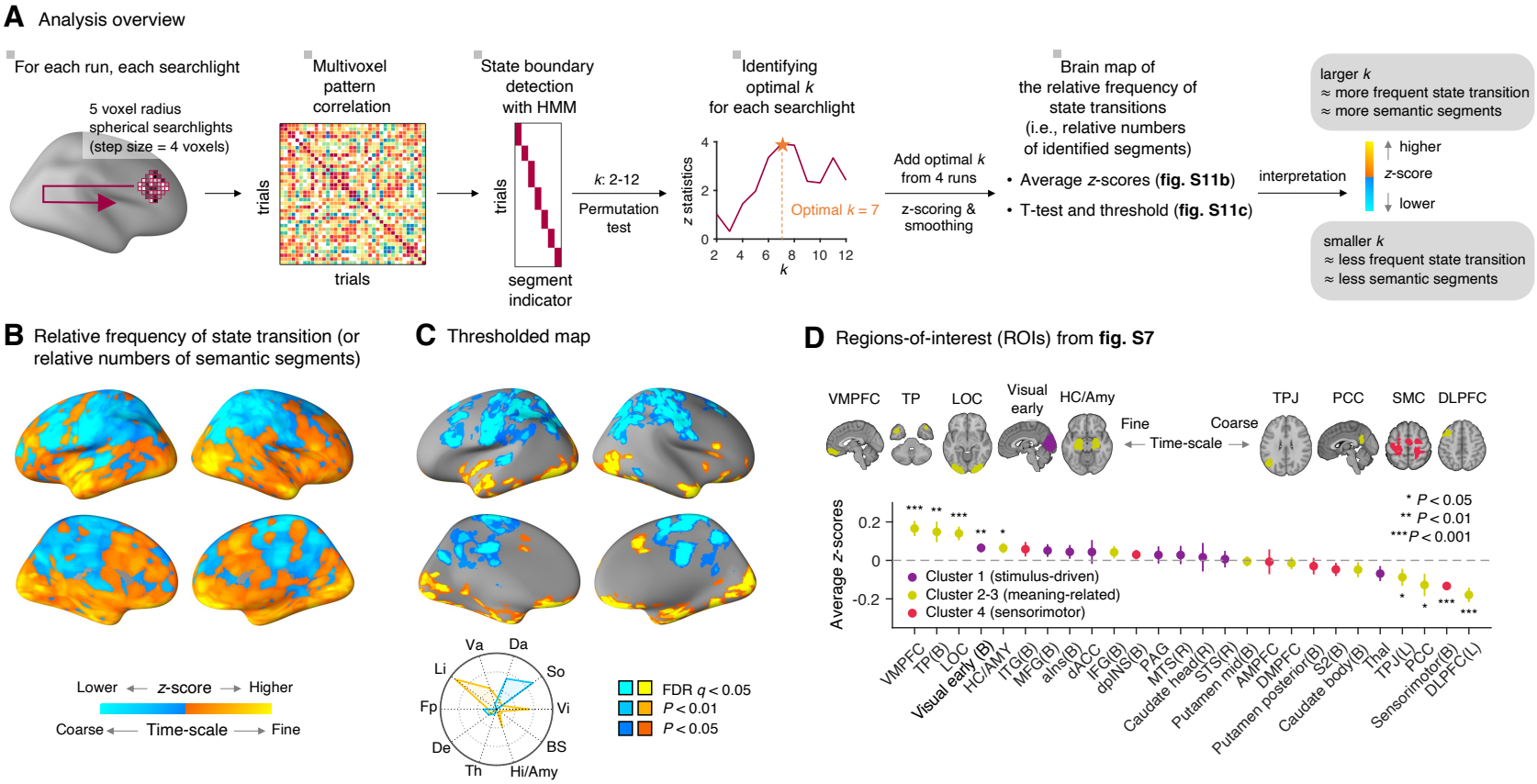


Fig. S11. Relative frequency of state transition (or relative numbers of semantic segments) across the brain.

(A) Analysis overview. To examine the relative frequency of state transition (or the relative numbers of semantic segments) of different regions based on their multivariate pattern information, we used a data-driven approach to detecting state boundaries with the Hidden Markov Model (HMM) (*77*). For details about the analysis steps, please see below. **(B)** Unthresholded group-level average of the state transition frequency. This map shows the group-averaged *z*-scores based on the total number (*k*) of state transitions across four runs. The cool color indicates a smaller number of state transitions, whereas the warm color indicates a larger number of state transitions. **(C)** Thresholded map of the relative frequency of state transition with FDR corrected $q < 0.05$, one-sample *t*-test, two-

tailed. To better show the extent of the significant areas, we pruned the results using two additional, more liberal thresholds, uncorrected voxel-wise $P < 0.01$ and $P < 0.05$, two-tailed. The radial plot shows the relative proportions of overlapping voxels between the thresholded map and large-scale networks. Va, ventral attention; Da, dorsal attention, So, somatomotor; Vi, visual; BS, brainstem; Hi/Am, hippocampus/amygdala; Th, thalamus; De, default; Fp, frontoparietal; Li, limbic. (**D)** We conducted bootstrap tests for the 26 regions-of-interest (ROIs) obtained from the basic contrast map of the concept reflection-related brain activity. The plot shows the group-average *z*-scores with the standard error of the mean (s.e.m.). $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, two-tailed, bootstrap test (for the details of results, see **table S6**). For full region names, please see **fig. S7**.

Methods: To detect the state boundaries of concept representation for each run, we used a version of the Hidden Markov Model (HMM) implemented by previous work (*77*). This HMM version is optimized for detecting boundaries based on multi-voxel pattern information of brain regions. For the current analysis, we applied the HMM to each participant's single-trial data. As shown in **(A)**, we scanned a spherical searchlight with a radius of 5 voxels (= 10 mm) across the whole brain with a step size of 4 voxels, resulting in 3,297 spherical searchlights in total. For each searchlight, we detected segment boundaries using the HMM with segment numbers ranging from $k = 2$ to 12. We used $k = 12$ as the upper limit because any number of state transitions (*k)* larger than 12 produced too many one-trial only comprised segments, and with $k \leq 12$, we were able to keep the number of one-trial segments to less than 1% of the number of total segments. The average difference in pattern similarity (i.e., spatial correlation) was calculated between intra- vs. inter-segment trials for each searchlight, each segmentation solution, and each *k*. We used this average difference as an indicator of how well the segmentation solutions captured the neural state transitions. We selected the *k* that produced the largest difference in pattern similarity as the optimal *k* for a certain searchlight, run, and participant. We then added the optimal *k*s from four runs for each participant, entered the value in the $4 \times 4 \times 4$ voxel cube for each searchlight, z-scored the sum of *k* values across the whole-brain, and applied smoothing with a 3-mm FWHM Gaussian kernel. For the final step, we conducted one-sample *t*-tests on the normalized *k* maps and thresholded the results with FDR $q < 0.05$. We also performed an ROI analysis with bootstrap tests (10,000 samples) to determine whether the normalized optimal *k* differed from zero.

Main findings: Multiple brain regions within the limbic system, including the ventromedial prefrontal cortex (VMPFC), orbitofrontal cortex, medial temporal lobe, and temporal pole (TP), consistently showed more frequent state transitions (i.e., finer-grained semantic segmentation structure) than other brain regions. Bootstrap tests on the 26 regions-of-interest (ROIs) selected from the previous analyses (see **fig. S7**) provided a similar result—the VMPFC and TP showed the largest number of segments among all ROIs (**Fig. 4D** and **table S6**), and the hippocampus and amygdala also showed a larger number of segments than other regions. In addition to the limbic areas, brain regions within the visual and ventral attention networks, including the early visual area, lateral occipital cortex (LOC), anterior insula, and mid-cingulate cortex, also showed larger numbers of segments compared to other brain regions. On the other end of the scale, multiple brain regions within the somatomotor, dorsal attention, and default mode networks showed small numbers of segments, including sensorimotor cortex, dorsolateral prefrontal cortex (DLPFC), posterior cingulate cortex (PCC), and temporal parietal junction (TPJ).

When we compared these findings with the results from the previous study (*77*) that used exogenous movie stimuli, there were consistent as well as inconsistent patterns of results. The consistent results between (*77*) and the current study include more segments in the visual cortices and less segments in the PCC and the TPJ (which was near the angular gyrus), which can be interpreted as integrating semantic information along the cortical hierarchy, from the unimodal sensory to higher-order transmodal brain regions. However, unlike the previous study (*77*), we found that some high-order transmodal brain regions within the limbic system showed more segments than other regions, including the VMPFC and TP.

We then examined whether these larger numbers of segments (i.e., more frequent state transitions) in the limbic cortical and subcortical regions (the VMPFC, TP, hippocampus/amygdala) and in the visual cortices were due to recurrent activations of similar semantic representations over time using the ratio of intra- to inter-segment pattern similarity (**fig. S12**). The results suggest that the large number of segments within the visual cortical regions could be due to the recurrent activations of similar representations. This was not the case for the limbic regions, which showed a low intra-to-inter segment pattern similarity ratio. Lastly, there is also a possibility that these results are simply due to the low signal-to-noise ratio of these limbic regions, and our supplementary analysis results (**fig. S13**) suggest that it may not be the case in our study—the correlation between the ranks of the optimal k and tSNR was not significant (Spearman's ρ = -0.207, P = 0.3092).

Interestingly, the VMPFC and TP were not even included in the analysis in the previous study (*77*) because of these regions' low inter-subject synchrony, which has been also reported in other studies (*44,78*). Together, these findings may suggest these regions' fundamental roles in endogenous cognitive and affective processes, such as storing and retrieving autobiographical memories of personal experiences and spontaneous thought. Therefore, these regions are likely to serve as a major source of the idiosyncrasy across individuals, which will be crucial for advancing personalized neuroscience and personalized treatment for psychiatric disorders.

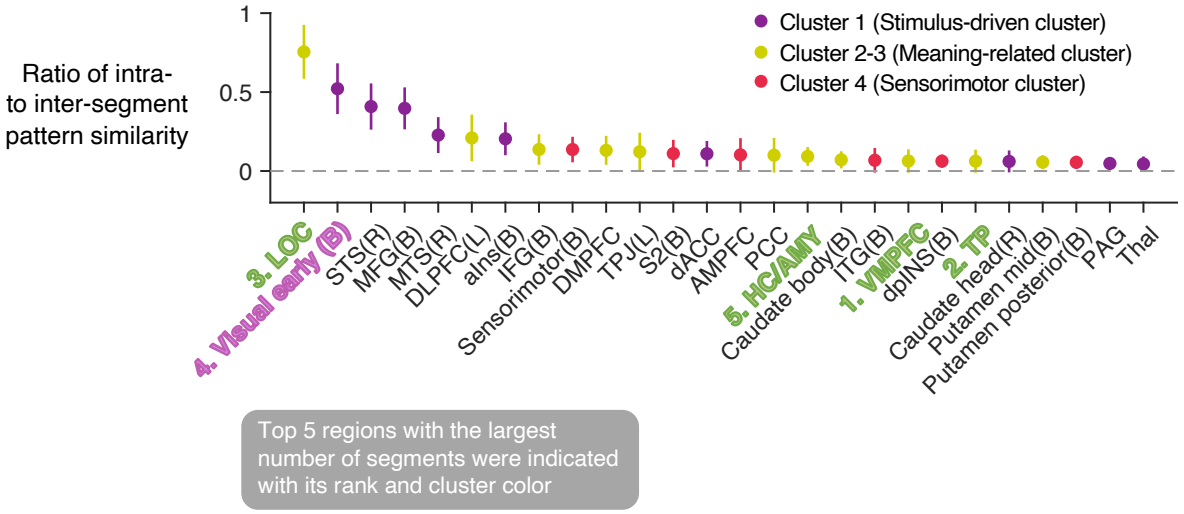


Fig. S12. Intra- versus inter-segment pattern similarity of 26 ROIs.

To examine whether a large number of segments (i.e., more frequent state transitions) in some brain regions was due to the recurrent activations of similar semantic representations, we calculated the ratio of intra- to inter-segment pattern similarity. The plot shows the average ratio of intra- vs. inter-segment pattern similarity across participants for the 26 ROIs. The error bars represent the standard error of the mean (s.e.m.).

Main findings: Among the top 5 regions with the largest number of segments, the LOC and the early visual cortex showed a high intra-to-inter segment pattern similarity ratio, whereas the VMPFC, TP, and the hippocampus/amygdala showed a low intra- to inter-segment pattern similarity ratio. This suggests that the large number of segments within the visual cortical areas could be due to the recurrent activations of similar representations, but this was not the case in limbic regions.

Methods: Based on the segmentation results with the optimal *k* of each region, we calculated the intra-segment pattern similarity using the trials within each segment. We also calculated the inter-segment pattern similarity only considering the trials within the adjacent segments. Then, we divided the intra-state pattern similarity by the inter-state pattern similarity across ROIs and runs and averaged them across four runs for each participant.

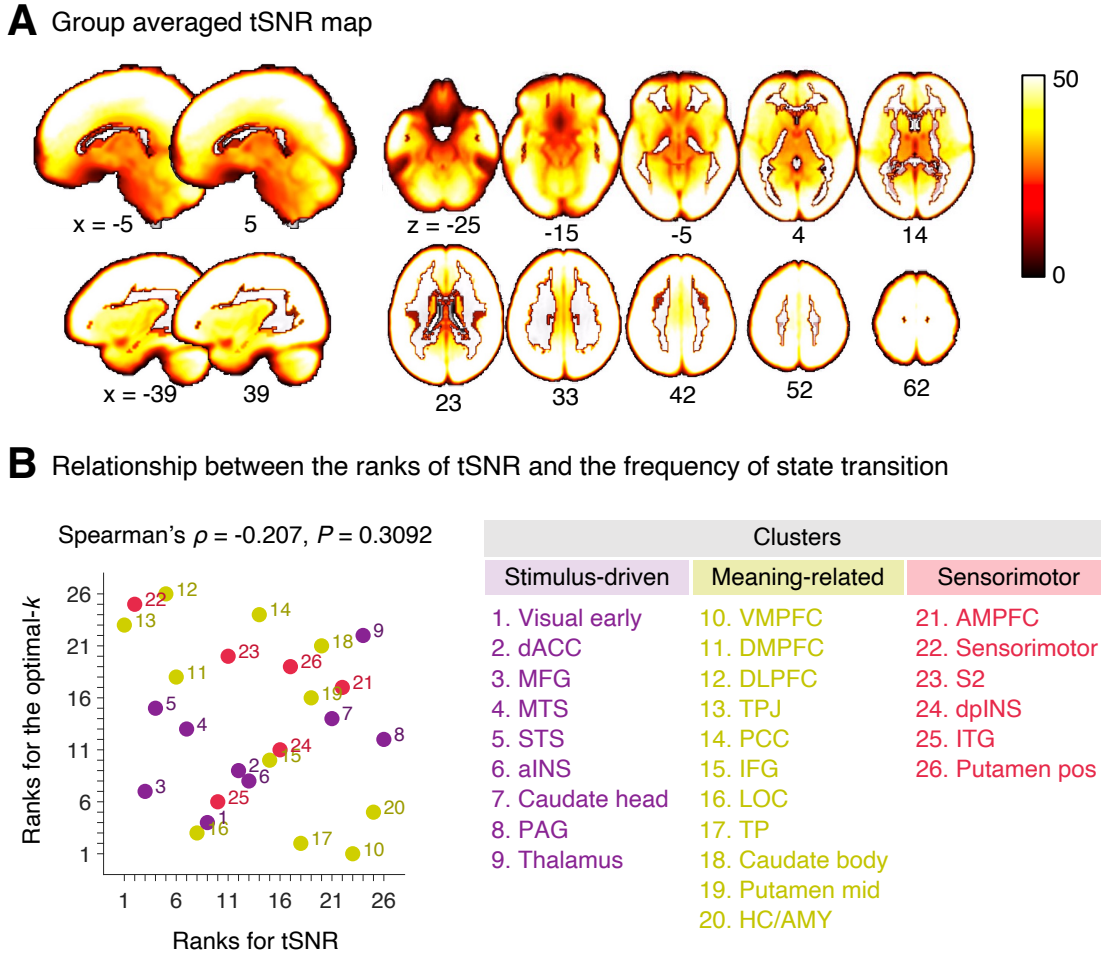


Fig. S13. The relationship between the temporal Signal-to-Noise Ratio (tSNR) and the relative frequency of state transition.

We examined whether the results of the relative frequency of state transition were confounded with the levels of signal-to-noise ratio of the BOLD signal. **(A)** We calculated the temporal signal-to-noise ratio (tSNR) using the TR images of the concept reflection runs and then averaged the tSNR values across runs and participants (n = 61). The map shows the group-average of the tSNR. **(B)** We calculated the Spearman's correlation between the ranks of two variables—the optimal number (k) of segments and the tSNR. The correlation between the two ranks was not significant (Spearman's ρ = -0.207, P = 0.3092). For the full region names, please see **fig. S7**.

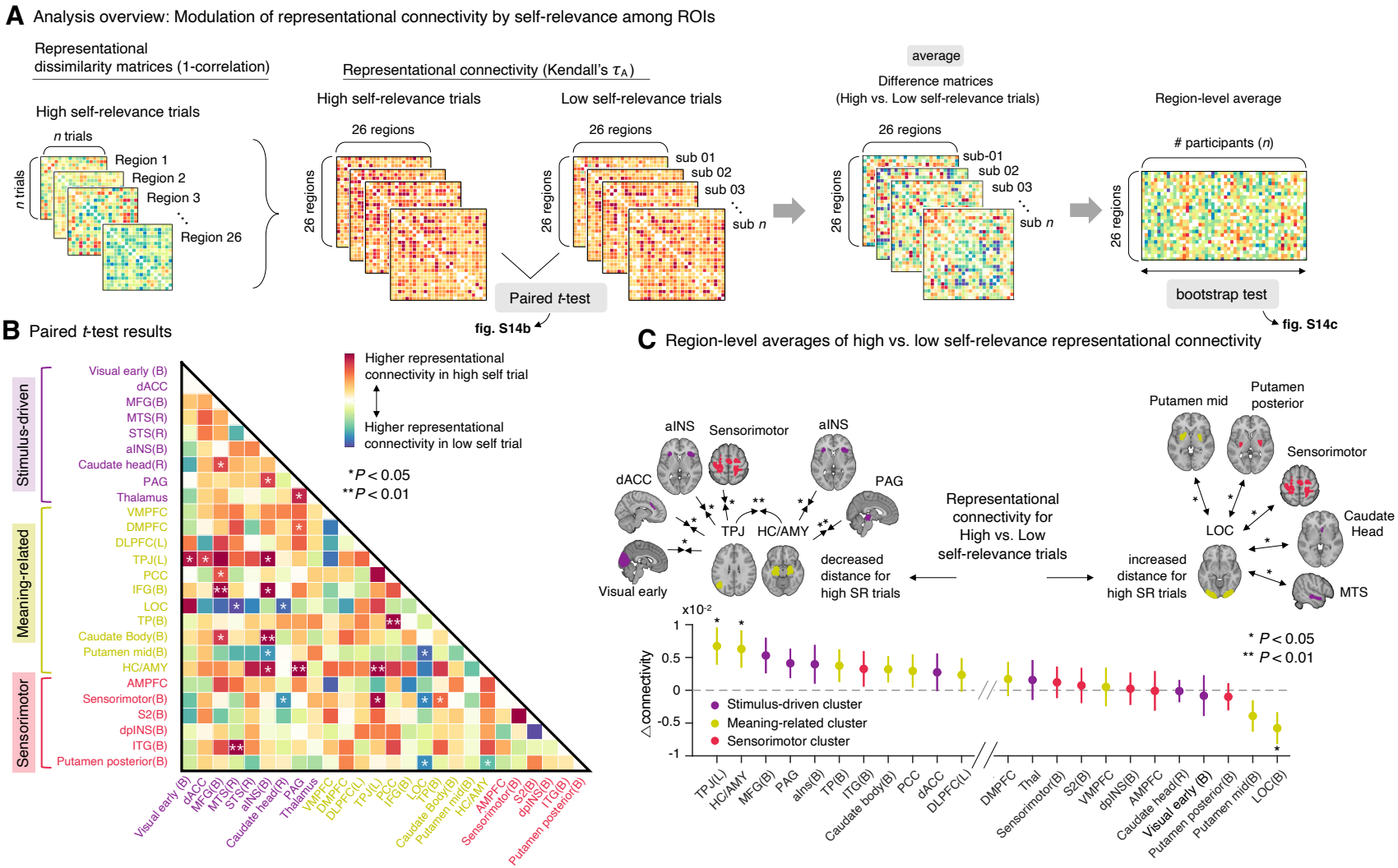

A Analysis overview: Modulation of representational connectivity by self-relevance among ROIs
Representational dissimilarity matrices (1-correlation)
High self-relevance trials
n trials
Region 1
Region 2
Region 3
Region 26
Representational connectivity (Kendall's τA)
High self-relevance trials
Low self-relevance trials
26 regions
sub 01
sub 02
sub 03
sub n
Paired t-test
fig. S14b
average
Difference matrices
(High vs. Low self-relevance trials)
sub-01
Region-level average
participants (n)
bootstrap test
fig. S14c
B Paired t-test results
Higher representational connectivity in high self trial
Higher representational connectivity in low self trial
*P < 0.05
**P < 0.01
Stimulus-driven
Meaning-related
Sensorimotor
Visual early (B)
dACC
MFG(B)
MTS(R)
STS(R)
aINS(B)
Caudate head(R)
PAG
Thalamus
VMPFC
DMPFC
DLPFC(L)
TPJ(L)
PCC
IFG(B)
LOC
TP(B)
Caudate Body(B)
Putamen mid(B)
HC/AMY
AMPFC
Sensorimotor(B)
S2(B)
dpINS(B)
ITG(B)
Putamen posterior(B)
C Region-level averages of high vs. low self-relevance representational connectivity
aINS
Sensorimotor
dACC
PAG
TPJ
HC/AMY
Visual early
decreased distance for high SR trials
Representational connectivity for High vs. Low self-relevance trials
increased distance for high SR trials
Putamen mid
Putamen posterior
LOC
Caudate Head
MTS
x10-2
△connectivity
Stimulus-driven cluster
Meaning-related cluster
Sensorimotor cluster
* P < 0.05
** P < 0.01
TPJ(L)
HC/AMY
MFG(B)
PAG
aIns(B)
TP(B)
ITG(B)
Caudate body(B)
PCC
dACC
DLPFC(L)
DMPFC
Thal
Sensorimotor(B)
S2(B)
VMPFC
dpINS(B)
AMPFC
Caudate head(R)
Visual early (B)
Putamen posterior(B)
Putamen mid(B)
LOC(B)

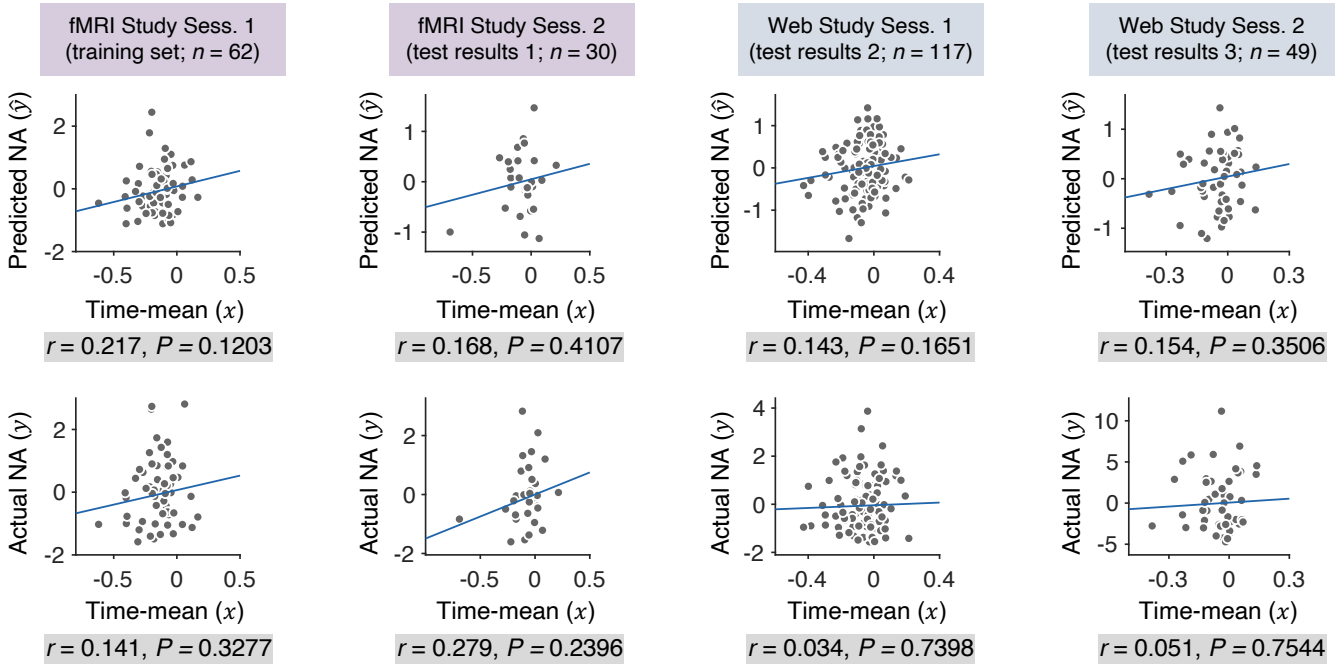
Fig. S14. Representational connectivity analysis for the modulatory effects of self-relevance on representational similarity among ROIs.

(A) Analysis overview. To identify which brain regions showed the representational changes modulated by the level of self-relevance, we first divided the trials into high vs. low self-relevance groups. Then we calculated representational dissimilarity matrices (RDMs) among high self-relevance or low self-relevance trials for each ROI using one minus correlations. With these RDMs, we calculated the representational connectivity among ROIs using Kendall's τ_A, resulting in two representational connectivity matrices per participant—one for the high self-relevance condition and the other for the low self-relevance condition. Given that we used 26 ROIs, the size of each representational connectivity matrix was 26×26. Using these representational connectivity matrices, we conducted the paired *t*-tests between the high vs. low self-relevance conditions. With the difference matrices, we calculated the region-level averages and conducted bootstrap tests (with 10,000 iterations) to identify the ROIs that showed the significant changes in the representational connectivity with other regions by the level of self-relevance. **(B)** The matrix shows the paired *t*-test results with the group average of the difference representational connectivity matrices. The ROI pairs with warm (vs. cold) colors indicate that they showed higher (vs. lower) levels of representational connectivity during high self-relevance than low self-relevance trials. $^{*}P < 0.05$, $^{**}P < 0.01$, uncorrected, two-tailed, bootstrap tests. **(C)** The bottom plot shows the region-level averages of representational connectivity for the high vs. low self-relevance conditions. A dot represents each region, and the y-axis represents the mean difference in representational connectivity for the high vs. low self-relevance comparisons. The error bars represent the standard error of the mean (s.e.m.) across individuals. The asterisk indicates the result of bootstrap tests for the paired comparisons, two-tailed.

Main findings: The left TPJ, HC/AMY, and bilateral LOC showed significant changes in the representational connectivity with other regions modulated by self-relevance. The left TPJ and HC/AMY showed the overall increases in the representational connectivity with other regions; for the left TPJ, $z = 2.385$, $P = 0.0171$, for the HC/AMY, $z = 2.178$, $P = 0.0294$, whereas the bilateral LOC showed decreased representational connectivity with other regions, $z = -2.446$, $P = 0.0144$ (for the results of all the ROIs, see **table S6**). Note that all these regions were a part of the meaning-related cluster.

We then identified the brain regions that showed significant modulations in representational connectivity with these three ROIs. The TPJ and HC/AMY showed decreased representational connectivity with the sensorimotor and salience network brain regions for highly self-relevant trials, including the S1/M1, visual cortex, dACC, and aINS. For the bilateral LOC region, we observed increased representational connectivity with some basal ganglia regions, including putamen and caudate head, S1/M1, and middle temporal sulcus, during highly self-relevant trials. Overall, these results suggest that the left TPJ and HC/AMY regions played a role as the hub attractor regions for the high self-relevance trials.

A Correlations between the time-mean variable and the predicted and actual outcomes



NA = Negative Affectivity

B Distribution of the time-mean variable

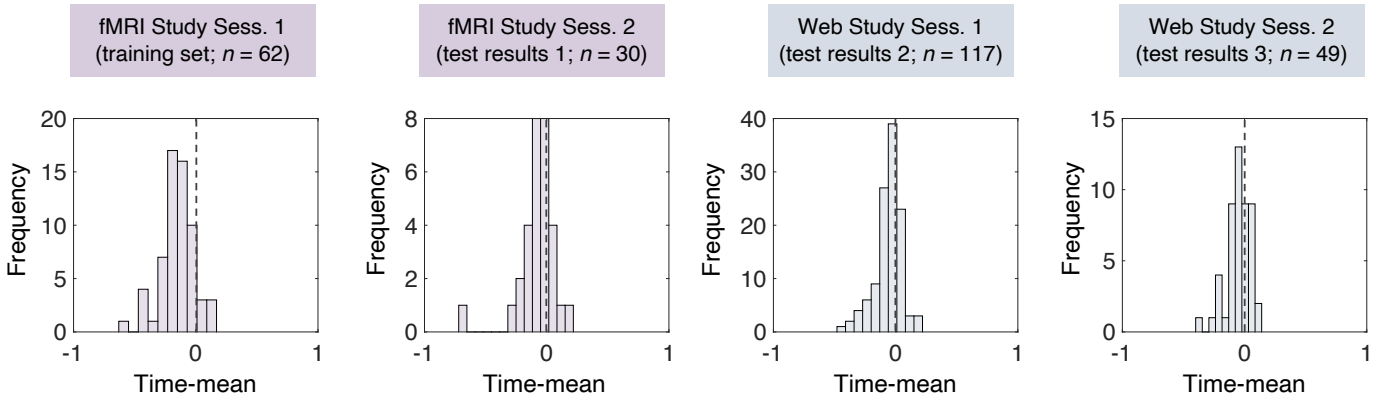


Fig. S15. Further analysis of the time-mean variable

(A) Structural coefficient analysis. To further investigate the characteristics of the time-mean variable in our Markov chain-based predictive model, we first conducted a structural coefficient analysis (*79*). In the analysis, we correlated the time-mean variable with the predicted negative affectivity ($\hat{y}$, model prediction; top row) and the actual negative affectivity (y; bottom row). The results showed that the time-mean variable was not significantly correlated with the predicted negative affectivity nor with the actual negative affectivity across all four datasets from the FAST-fMRI study and the FAST-web study. **(B)** Histograms of the time-mean variable. The distribution of the time-mean variable showed that the group averages of the time-mean variable were negative (i.e., past-oriented) across all four datasets, mean ± SD for the first session of the FAST-fMRI study = -0.145 ± 0.141 with [min, max] = [-0.620, 0.166], for the second session of FAST-fMRI study, -0.080 ± 0.153, [-0.691, 0.214], for the first session of the FAST-web study, -0.057 ± 0.114, [-0.428, 0.213], and for the second session of the FAST-web study, -0.050 ± 0.102, [-0.384, 0.138].

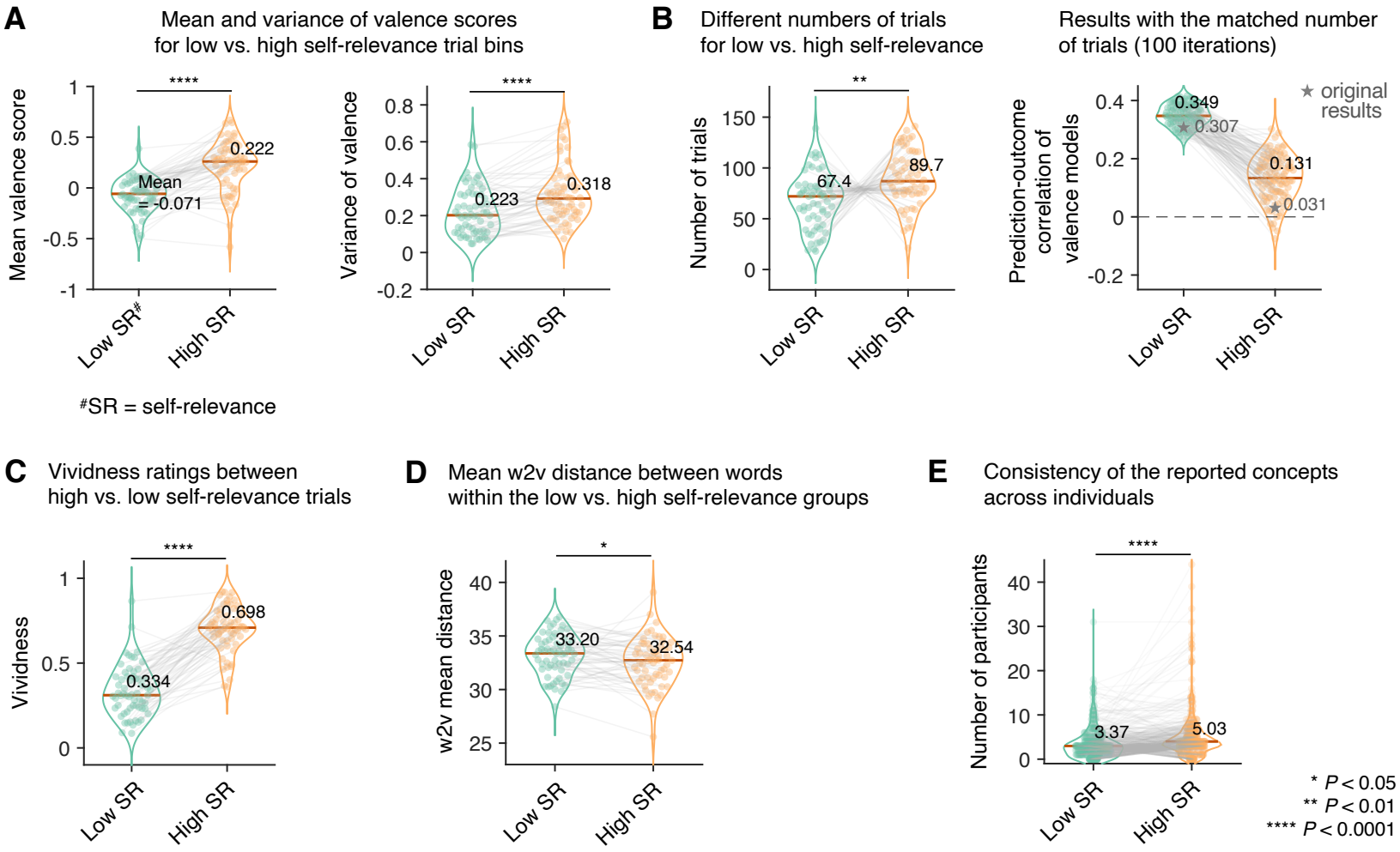

A
Mean and variance of valence scores
for low vs. high self-relevance trial bins
Mean valence score
****
0.222
Mean
= -0.071
Low SR#
High SR
Variance of valence
****
0.223
0.318
Low SR
High SR
#SR = self-relevance
B
Different numbers of trials
for low vs. high self-relevance
Number of trials
**
67.4
89.7
Low SR
High SR
Results with the matched number
of trials (100 iterations)
Prediction-outcome
correlation of
valence models
★ original
results
0.349
★ 0.307
0.131
★ 0.031
Low SR
High SR
C
Vividness ratings between
high vs. low self-relevance trials
Vividness
****
0.334
0.698
Low SR
High SR
D
Mean w2v distance between words
within the low vs. high self-relevance groups
w2v mean distance
*
33.20
32.54
Low SR
High SR
E
Consistency of the reported concepts
across individuals
Number of participants
****
3.37
5.03
Low SR
High SR
* P < 0.05
** P < 0.01
**** P < 0.0001

Fig. S16. Multiple differences between low versus high self-relevance trial bins.

To characterize the low and high self-relevance trial bins, we compared multiple variables between the low vs. high self-relevance trial bins. In the violin plots in **(A)**, **(B)** left, **(C)**, and **(D)**, dots represent individuals. In the violin plot in **(B)** right, dots represent iterations. In the violin plot in **(E)**, dots represent concepts.

(A) Mean and variance of valence scores for low vs. high self-relevance trial bins. The high self-relevance trial bin had a significantly higher mean valence compared to the low self-relevance trials, $t_{60} = 10.297$, $P < 0.0001$. In addition, the high self-relevance trials showed a significantly higher variance in valence, $t_{60} = 6.871$, $P < 0.0001$.

(B) (Left) The number of high self-relevance trials was significantly higher than that of low self-relevance trials, $t_{60} = 3.237$, $P = 0.0020$. (Right) To test whether these different numbers of trials influenced the patterns of prediction performances, we re-trained the valence models with the matched number of trials. For this, we randomly selected the trials for the group that had a larger number of trials between the low vs. high self-relevance trial bins. For example, if the number of one participant's high self-relevance trials was 100, and that of low self-relevance was 60, then we randomly selected 60 trials among the high self-relevance trials to match the number of trials between the two bins. We repeated this random selection procedure 100 times. The results showed that the prediction performances of the valence models with the matched number of trials were similar to the original results—i.e., the valence-lowself model showed a better and more significant prediction performance compared to the valence-highself model, for the valence-lowself model, mean prediction-outcome correlation $r = 0.349$ (originally, $r = 0.307$), for the valence-highself model, mean $r = 0.131$ (originally, $r = 0.031$). These results suggest that the lower prediction performance of the valence-highself model in comparison to the valence-lowself model was not driven by the different number of trials.

(C) The high self-relevance trials showed a significantly higher level of vividness than the low self-relevance trials, $t_{60} = 16.202$, $P < 0.0001$.

(D) To compare the semantic distances among the concepts (i.e., trials) within the high vs. low self-relevance bins, we implemented a word embedding model that transforms a word text into a single vector of multidimensional semantic space. We used the Korean Wikipedia corpus and built a Korean Word2Vec model with 90 dimensions. The results showed that the Word2Vec mean distance among the concepts within the high self-relevance bin was significantly shorter than the low self-relevance bin, $t_{60} = 2.184$, $P = 0.0329$, suggesting that the concepts within the high self-relevance bin were semantically closer to each other (i.e., high semantic density) compared to the low self-relevance bin. The findings of **(C)** and **(D)** could be due to the episodic memory component in the highly self-relevant concepts; episodic memory is known to have a high level of specificity, details, and concreteness (e.g., compared to semantic memory), which can lead to high semantic neighborhood density (*80*).

(E) To examine the relative frequency of each word in the high vs. low self-relevance trial bins across individuals, we first obtained the unique concepts that at least appeared either in the high or low self-relevance bins across more than two individuals. Then, we

counted and compared how many participants reported each concept for the high vs. low self-relevance bins. In more detail, there were 3,413 unique concepts in total (from 61 participants in the fMRI study), and 434 concepts appeared across at least 2+ individuals, either in the high or low self-relevance bins. For these 434 concepts, we compared the number of participants between the high vs. low self-relevance bins. The results showed that the high self-relevance bin showed a greater level of consistency across individuals than the low self-relevance bin, $t_{433} = -6.738$, $P < 0.0001$.

$^{*}P < 0.05$, $^{**}P < 0.01$, $^{****}P < 0.0001$, two-tailed, paired *t*-test.

Main findings: Overall, we observed a number of differences between the high vs. low self-relevance trial bins, including the distribution of valence and vividness scores, semantic distances, and the consistency of the reported concepts across individuals. However, it is unclear whether the observed differences are *causes* of our main findings, or whether all these observed differences (including our main findings) are the characteristics of self-relevant spontaneous thought. Therefore, we believe that this line of inquiry is rather fundamental and necessitates further, careful investigations that should be the focus of future studies.

Table S1. Stability of the features across different sets of seed words and test time points (7-week interval)

Dynamic features	Valence		Safety-threat		Time		Self-relevance		Vividness	
	r	*P*	*r*	*P*	*r*	*P*	*r*	*P*	*r*	*P*
Mean	0.543**	0.0019	0.772***	0.0000	0.626***	0.0002	0.609***	0.0004	0.677***	0.0000
Variance	0.721***	0.0000	0.754***	0.0000	0.367*	0.0462	0.693***	0.0000	0.621***	0.0002
Transition prob.										
Lv.1 → Lv.1	0.408*	0.0254	0.271	0.1467	0.447*	0.0133	0.671***	0.0000	0.596***	0.0005
Lv.2 → Lv.1	0.456*	0.0113	0.741***	0.0000	0.807***	0.0000	0.440*	0.0149	0.522**	0.0031
Lv.3 → Lv.1	0.589***	0.0006	0.197	0.2967	-0.075	0.6947	-	-	-	-
Lv.1 → Lv.2	0.567**	0.0011	0.110	0.5621	0.314	0.0915	0.671***	0.0000	0.596***	0.0005
Lv.2 → Lv.2	0.579***	0.0008	0.815***	0.0000	0.810***	0.0000	0.440*	0.0149	0.522**	0.0031
Lv.3 → Lv.2	0.489**	0.0061	0.498**	0.0051	-0.263	0.1607	-	-	-	-
Lv.1 → Lv.3	0.505**	0.0044	0.164	0.3871	0.466**	0.0095	-	-	-	-
Lv.2 → Lv.3	0.438*	0.0154	0.747***	0.0000	0.755***	0.0000	-	-	-	-
Lv.3 → Lv.3	0.554**	0.0015	0.661***	0.0001	0.023	0.9020	-	-	-	-
Steady state prob.										
Lv.1	0.690***	0.0000	0.682***	0.0000	0.812***	0.0000	0.565**	0.0011	0.591***	0.0006
Lv. 2	0.673***	0.0000	0.757***	0.0000	0.788***	0.0000	0.565**	0.0011	0.591***	0.0006
Lv. 3	0.566**	0.0011	0.791***	0.0000	0.467**	0.0092	-	-	-	-

Note. We examined the stability and test-retest reliability of the Markov chain-based dynamic features across different sets of seed words and across two different time points (7-week interval on average) with a subset of participants (n = 30). For the details of how we defined the states and calculated the transition and steady-state probabilities, please refer to **Methods**. Note that since the self-relevance and vividness dimensions had only two discrete states, their correlation values had the same values (e.g., for Lv.1 → Lv.1 and Lv.1 → Lv.2; because one probability is one minus the other probability). One-sample *t*-test for correlations was performed. Lv.1 represents negative, threat, past, low, and low for valence, safety-threat, time, self-relevance, and vividness, respectively. Lv.2 represents neutral, neutral, present, high, and high for valence, safety-threat, time, self-relevance, and vividness, respectively. Lv. 3

represents positive, safety, and future for valence, safety-threat, and time, respectively. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, *t*-test for Pearson's correlation, two-tailed.

Table S2. Test-retest reliability and factor loadings of self-report questionnaires

Questionnaires	Subscale (if subscales were used)	FAST-fMRI (n = 30) Test-retest reliability		FAST-fMRI (n = 62)		FAST-web (n = 117)	
				Factor 1	Factor 2	Factor 1	Factor 2
		r	P	Negative affect	Positive affect	Negative affect	Positive affect
PANAS	Positive affect	0.278	0.1371	0.081	**0.939**	0.261	**0.751**
PANAS	Negative affect	0.545 **	0.0018	**0.736**	0.264	**1.127**	0.496
CES-D	-	0.655 ***	0.0001	**0.667**	-0.329	**0.775**	-0.108
MASQ30	General distress	0.714 ***	0.0000	**0.870**	-0.073	-	-
MASQ30	Anhedonic depression[a]	0.415 *	0.0227	-0.094	**0.809**	-	-
MASQ30	Anxiety arousal	0.723 ***	0.0000	**0.571**	0.323	-	-
STAI-T	-	0.850 ***	0.0000	**0.794**	-0.249	**0.663**	-0.304
RRS	Brooding	0.563 **	0.0012	**0.747**	0.060	-	-
RRS	Depressive rumination	0.562 **	0.0012	**0.814**	0.070	**0.547**	-0.110
SIQ	Two representative items (12, 22)	-	-	-	-	**0.144**	-0.403
LS	-	-	-	-	-	**0.306**	-0.511
PWB	Environmental mastery	-	-	-	-	-0.192	**0.436**
PWB	Autonomy	-	-	-	-	0.081	**0.486**
PWB	Positive relations	-	-	-	-	-0.206	**0.574**
PWB	Purpose in life	-	-	-	-	0.173	**0.605**
PWB	Personal growth	-	-	-	-	0.102	**0.649**
PWB	Self-acceptance	-	-	-	-	-0.192	**0.675**
SWLS	-	-	-	-	-	-0.138	**0.545**

Note. Through the FAST-fMRI study, we examined the test-retest reliability of the self-report questionnaires with a 7-week interval using Pearson's correlations. All the questionnaires except for the PANAS-positive affect subscale showed medium to high levels of test-retest reliability, suggesting that these questionnaires provide trait measures. To obtain a general negative affectivity score that

can be used as an outcome variable in predictive modeling, we conducted factor analyses for a two-factor model. The factor analyses were done separately for the FAST-fMRI and FAST-web studies because these two studies conducted different sets of self-report questionnaires (e.g., we included more questionnaires related to positive affectivity in the FAST-web study). The values in bold indicate the higher factor loadings between two factors to show which factor the questionnaire belongs to. PANAS, Positive and Negative Affect Schedule; CES-D, Center for Epidemiologic Studies Depression; MASQ30, 30-item Mood and Anxiety Symptom Questionnaire; STAI-T, State-Trait Anxiety Inventory-Trait version; RRS, Rumination Response Scale; SIQ, Suicidal Ideation Questionnaire; LS, Loneliness Scale; PWB, Psychological Well-Being Scale; SWLS, Satisfaction With Life Scale. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, *t*-test for Pearson's correlation, two-tailed. [a] reverse coding.

Table S3. Descriptive statistics of self-report questionnaires

Questionnaire	Subscale (if subscales were used)	Test				Re-test				Possible range
		Mean	SD	Min	Max	Mean	SD	Min	Max	
fMRI study										
PANAS	Positive affect	22.177	6.880	9	36	22.767	5.894	11	36	9-45
PANAS	Negative affect	18.774	6.258	11	34	20.733	6.198	12	35	11-55
CES-D	-	12.661	8.686	0	43	14.933	11.095	2	36	0-60
MASQ30	General distress	19.661	7.767	10	38	21.767	9.261	10	41	10-50
MASQ30	Anhedonic depression	26.806	9.179	11	49	26.333	8.409	12	42	10-50
MASQ30	Anxiety arousal	14.290	5.391	10	33	16.800	7.481	10	40	10-50
STAI-T	-	21.355	10.217	5	49	24.533	11.048	10	51	0-60
RRS	Brooding	13.484	5.315	7	26	15.467	5.692	7	25	7-28
RRS	Depressive rumination	17.177	5.801	9	34	19.900	6.820	9	33	9-36
Web study										
PANAS	Positive affect	23.342	7.350	10	38	22.755	6.336	10	35	9-45
PANAS	Negative affect	19.393	7.375	11	47	19.367	8.416	11	47	11-55
CES-D	-	11.897	7.429	0	35	11.816	8.348	1	35	0-60
STAI-T	-	41.684	10.397	25	65	42.367	11.178	26	63	20-80
RRS	Depressive rumination	15.880	5.138	9	31	16.735	5.696	9	31	9-36
SIQ	Two representative items (12, 22)	1.009	1.941	0	11	1.082	1.956	0	9	0-12
LS	-	2.752	2.308	0	9	2.878	2.627	0	9	0-9
PWB	Environmental mastery	10.692	2.091	5	14	10.714	2.021	6	14	3-18
PWB	Autonomy	9.889	2.522	4	15	9.245	2.411	4	15	3-18
PWB	Positive relations	11.214	2.579	4	15	10.673	2.954	4	15	3-18
PWB	Purpose in life	12.316	2.250	6	15	12.327	2.212	7	15	3-18
PWB	Personal growth	12.812	1.903	7	15	12.755	1.820	9	15	3-18

PWB	Self-acceptance	10.821	2.351	3	15	10.633	2.464	3	15	3-18
SWLS	-	20.530	5.854	5	34	19.531	6.059	5	32	5-35

Note. Descriptive statistics of self-report questionnaires of the FAST-fMRI study and FAST-web study. The FAST-fMRI study and FAST-web study measured individual differences in personality and affectivity by using different sets of self-reported questionnaires. We conducted the factor analysis respectively for each study to calculate the general negative affectivity scores. Re-test data indicate the second sessions of the two studies. PANAS, Positive and Negative Affect Schedule; CES-D, Center for Epidemiologic Studies Depression; MASQ30, 30-item Mood and Anxiety Symptom Questionnaire; STAI-T, State-Trait Anxiety Inventory-Trait version; RRS, Rumination Response Scale; SIQ, Suicidal Ideation Questionnaire; LS, Loneliness Scale; PWB, Psychological Well-Being Scale; SWLS, Satisfaction With Life Scale.

Table S4. Correlation between general negative affectivity and Markov-chain dynamic features

Dynamic features	Valence		Safety-threat		Time		Self-relevance		Vividness	
	r	*P*	*r*	*P*	*r*	*P*	*r*	*P*	*r*	*P*
Mean	-0.432***	0.0004	-0.265*	0.0374	0.137	0.2880	0.348**	0.0056	0.270*	0.0338
Variance	0.343**	0.0063	0.337**	0.0074	0.159	0.2175	0.005	0.9702	0.140	0.2773
Transition prob.										
Lv.1 → Lv.1	0.294*	0.0203	0.202	0.1149	0.005	0.9682	-0.213	0.0962	-0.133	0.3042
Lv.2 → Lv.1	0.360**	0.0041	0.399**	0.0013	0.004	0.9767	-0.258*	0.0425	-0.133	0.3030
Lv.3 → Lv.1	0.416***	0.0008	0.344**	0.0062	-0.124	0.3381				
Lv.1 → Lv.2	-0.173	0.1796	-0.119	0.3561	-0.009	0.9432	0.213	0.0962	0.133	0.3042
Lv.2 → Lv.2	-0.045	0.7302	-0.200	0.1191	-0.057	0.6603	0.258*	0.0425	0.133	0.3030
Lv.3 → Lv.2	-0.207	0.1060	-0.156	0.2268	-0.005	0.9697				
Lv.1 → Lv.3	-0.221	0.0843	-0.109	0.4008	0.011	0.9312				
Lv.2 → Lv.3	-0.229	0.0739	-0.030	0.8189	0.114	0.3792				
Lv.3 → Lv.3	-0.180	0.1607	-0.068	0.5975	0.137	0.2892				
Steady state prob.										
Lv. 1	0.484***	0.0001	0.415***	0.0008	-0.027	0.8332	-0.278*	0.0288	-0.186	0.1486
Lv. 2	-0.193	0.1327	-0.249	0.0511	-0.028	0.8293	0.278*	0.0288	0.186	0.1486
Lv. 3	-0.232	0.0702	-0.001	0.9932	0.107	0.4058				

Note. Correlation values between the general negative affectivity scores from the factor analysis and the dynamic features from the Markov chain analysis (n = 62). Note that since the self-relevance and vividness dimensions had only two discrete states, some correlation values are the same (e.g., for Lv.1 → Lv.1 and Lv.1 → Lv.2; because one probability is one minus the other probability). One-sample t-test for correlations was performed. Lv.1 represents negative, threat, past, low, and low for valence, safety-threat, time, self-relevance, and vividness, respectively. Lv.2 represents neutral, neutral, present, high, and high for valence, safety-threat, time,

self-relevance, and vividness, respectively. Lv. 3 represents positive, safety, and future for valence, safety-threat, and time, respectively. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, *t*-test for Pearson's correlation, two-tailed.

Table S5. Hierarchical regression analysis to compare the contributions of non-Markov chain features versus Markov chain features

	Adjust R-squared	R-squared
A. Reduced model (features: mean and variance)	0.111	0.213
B. Full model (features: mean, variance, and Markov chain features	0.404	0.521
A-B	0.293	0.308

Note. We calculated the contributions of non-Markov chain features vs. Markov chain features of our predictive model using a hierarchical regression approach. We conducted this analysis in the training dataset with the final model (i.e., without further fitting). We did not use the cross-validation approach in this analysis because the goal of this analysis is not for a generalization but for an understanding of the model. The non-Markov chain features, including mean and variance, explained 21.3% of the total variance of y, while the full model explained 52.1% of the total variance. Thus, the additional variance explained by the Markov-chain features was 30.8%. When we used the adjusted R-squared, the additional variance explained by the Markov-chain features was 29.3%, suggesting that the Markov chain features (i.e., transition dynamics) play an important role in predicting negative affectivity.

Table S6. Optimal number of state segments and representational connectivity for the 26 regions-of-interest (ROIs)

ROIs	Relative frequency of state transition			Representational connectivity (Kendall's τ_A)			
				region-level averages		bootstrap test results	
	Average z-score	*z* (boot)	*P* (boot)	high self-relevance	low self-relevance	*z*	*P*
Stimulus-driven cluster							
Visual early(B)	0.065	2.808 **	0.0050	0.078	0.079	-0.298	0.7657
dACC	0.044	0.719	0.4719	0.068	0.065	0.956	0.3391
MFG(B)	0.051	1.568	0.1168	0.107	0.102	1.959	0.0501
MTS(R)	0.028	0.576	0.5643	0.092	0.090	0.712	0.4763
STS(R)	0.007	0.171	0.8640	0.101	0.100	0.631	0.5279
aINS(B)	0.044	1.251	0.2109	0.111	0.107	1.355	0.1755
Caudate head(R)	0.017	0.229	0.8188	0.043	0.043	-0.070	0.9441
PAG	0.028	0.636	0.5250	0.058	0.054	1.839	0.0660
Thalamus	-0.069	-1.762	0.0780	0.058	0.057	0.501	0.6164
Meaning-related cluster							
VMPFC	0.166	4.239 ***	0.0000	0.101	0.101	0.161	0.8717
DMPFC	-0.015	-0.452	0.6514	0.131	0.129	0.628	0.5298
DLPFC(L)	-0.179	-4.795 ***	0.0000	0.104	0.102	0.918	0.3586
TPJ(L)	-0.088	-1.965 *	0.0495	0.121	0.115	2.385 *	0.0171
PCC	-0.128	-2.221 *	0.0264	0.085	0.082	1.168	0.2426
IFG(B)	0.042	1.334	0.1821	0.120	0.118	0.781	0.4349
LOC	0.140	3.906 **	0.0001	0.051	0.057	-2.446 *	0.0144
TP(B)	0.148	2.846 **	0.0044	0.077	0.073	1.510	0.1310
Caudate Body(B)	-0.048	-1.238	0.2159	0.065	0.062	1.589	0.1121
Putamen mid(B)	-0.006	-0.226	0.8215	0.071	0.075	-1.668	0.0953
HC/AMY	0.064	2.465 *	0.0137	0.101	0.095	2.178 *	0.0294
Sensorimotor cluster							
AMPFC	-0.007	-0.101	0.9192	0.090	0.090	-0.043	0.9658
Sensorimotor(B)	-0.133	-6.389 ***	0.0000	0.100	0.099	0.501	0.6161
S2(B)	-0.047	-1.379	0.1679	0.088	0.087	0.281	0.7791
dpINS(B)	0.030	1.225	0.2207	0.098	0.098	0.086	0.9314
ITG(B)	0.058	1.572	0.1159	0.096	0.093	1.212	0.2255
Putamen posterior(B)	-0.030	-0.718	0.4725	0.063	0.064	-0.508	0.6115

Note. ROI-level analyses of the Hidden Markov Model (HMM) and representational connectivity. For details of these analyses, please refer to **Methods** (HMM) and **fig. S14** (representational connectivity analysis). $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, bootstrap tests with 10,000 iteration, two-tailed.

Table S7. Correlations with verbal fluency

Variables	*r*	*P*	Valence		Safety-threat		Time		Self-relevance		Vividness	
			r	*P*	*r*	*P*	*r*	*P*	*r*	*P*	*r*	*P*
Model prediction (predicted negative affectivity)	-0.135	0.2942										
# of unique words	0.299 *	0.0183										
Mean			-0.060	0.6439	-0.151	0.2399	-0.082	0.5268	-0.178	0.1652	-0.144	0.2658
Variance			-0.081	0.5308	-0.104	0.4204	-0.002	0.9895	0.176	0.1716	0.202	0.1147
Transition prob.												
Lv.1 → Lv.1			0.123	0.3420	0.066	0.6108	0.053	0.6827	0.083	0.5225	0.039	0.7649
Lv.2 → Lv.1			-0.077	0.5538	-0.016	0.9015	-0.031	0.8109	0.306 *	0.0154	0.177	0.1684
Lv.3 → Lv.1			-0.038	0.7722	-0.012	0.9274	0.008	0.9495				
Lv.1 → Lv.2			-0.090	0.4863	0.019	0.8823	0.023	0.8615	-0.083	0.5225	-0.039	0.7649
Lv.2 → Lv.2			0.058	0.6567	0.064	0.6188	0.021	0.8717	-0.306 *	0.0154	-0.177	0.1684
Lv.3 → Lv.2			0.197	0.1249	0.169	0.1891	0.109	0.3986				
Lv.1 → Lv.3			-0.061	0.6354	-0.142	0.2714	-0.233	0.0681				
Lv.2 → Lv.3			-0.016	0.8991	-0.084	0.5184	0.011	0.9352				
Lv.3 → Lv.3			-0.141	0.2730	-0.159	0.2171	-0.132	0.3075				
Steady state prob.												
Lv. 1			0.014	0.9166	0.033	0.7964	0.001	0.9913	0.209	0.1033	0.146	0.2565
Lv. 2			0.086	0.5072	0.110	0.3941	0.054	0.6767	-0.209	0.1033	-0.146	0.2565
Lv. 3			-0.110	0.3950	-0.167	0.1938	-0.111	0.3892				

Note. We assessed participants' verbal fluency prior to the fMRI experiment to examine whether the individual differences in verbal fluency were related to their FAST performance and results. For the verbal fluency test, we asked participants to produce as many words as possible that start with a given letter of the Korean alphabet within a one minute window. We tested with three Korean alphabet letters: ㄱ, ㅇ, and ㅅ. We also asked participants to produce as many animals as possible within one minute, regardless of the letter the word begins with. We used the total number of words the participants produced as a verbal fluency score. The verbal fluency

score did not show significant correlations with most of the variables except for two variables: 1) the number of unique words in the FAST response and 2) the transition probability of Lv.2 → Lv.1 (or Lv.2) on the self-relevance dimension (P = 0.0183 and 0.0154, respectively; non-significant after the correction for multiple comparisons). $^{*}P < 0.05$, t-test for Pearson's correlation, two-tailed.

REFERENCES AND NOTES

1. E. Bleuler, in *Studies in Word-Association*, M. D. Eder, Ed. (Moffat, Yard & Company, 1919), pp. 4–5.

2. K. Christoff, Z. C. Irving, K. C. Fox, R. N. Spreng, J. R. Andrews-Hanna, Mind-wandering as spontaneous thought: A dynamic framework. *Nat. Rev. Neurosci.* **17**, 718–731 (2016).

3. J. Smallwood, J. W. Schooler, The science of mind wandering: Empirically navigating the stream of consciousness. *Annu. Rev. Psychol.* **66**, 487–518 (2015).

4. M. A. Killingsworth, D. T. Gilbert, A wandering mind is an unhappy mind. *Science* **330**, 932 (2010).

5. W. James, F. Burkhardt, F. Bowers, I. K. Skrupskelis, *The Principles of Psychology* (Macmillan London, 1890), vol. 1.

6. I. Marchetti, E. H. W. Koster, E. Klinger, L. B. Alloy, Spontaneous thought and vulnerability to mood disorders: The dark side of the wandering mind. *Clin. Psychol. Sci.* **4**, 835–857 (2016).

7. K. A. McLaughlin, S. Nolen-Hoeksema, Rumination as a transdiagnostic factor in depression and anxiety. *Behav. Res. Ther.* **49**, 186–193 (2011).

8. Z. V. Segal, Appraisal of the self-schema construct in cognitive models of depression. *Psychol. Bull.* **103**, 147–162 (1988).

9. N. Marupaka, L. R. Iyer, A. A. Minai, Connectivity and thought: The influence of semantic network structure in a neurodynamical model of thinking. *Neural Netw.* **32**, 147–158 (2012).

10. M. L. Dixon, J. J. Gross, Dynamic network organization of the self: Implications for affective experience. *Curr. Opin. Behav. Sci.* **39**, 1–9 (2021).

11. J. Joormann, S. M. Levens, I. H. Gotlib, Sticky thoughts: Depression and rumination are associated with difficulties manipulating emotional material in working memory. *Psychol. Sci.* **22**, 979–983 (2011).

12. J. R. Andrews-Hanna, C.-W. Woo, R. Wilcox, H. Eisenbarth, B. Kim, J. Han, E. A. R. Losin, T. D. Wager, The conceptual building blocks of everyday thought: Tracking the emergence and dynamics of ruminative and nonruminative thinking. *J. Exp. Psychol. Gen.* **151**, 628–642 (2022).

13. J. R. Andrews-Hanna, R. H. Kaiser, A. E. J. Turner, A. E. Reineberg, D. Godinez, S. Dimidjian, M. T. Banich, A penny for your thoughts: Dimensions of self-generated thought content and relationships with individual differences in emotional wellbeing. *Front. Psychol.* **4**, 900 (2013).

14. A. D'Argembeau, Mind-wandering and self-referential thought, in *The Oxford Handbook of Spontaneous Thought: Mind-Wandering, Creativity, and Dreaming* (Oxford Univ. Press, 2018), pp. 181–191.

15. E. Klinger, Goal commitments and the content of thoughts and dreams: Basic principles. *Front. Psychol.* **4**, 415 (2013).

16. B. Medea, T. Karapanagiotidis, M. Konishi, C. Ottaviani, D. Margulies, A. Bernasconi, N. Bernasconi, B. C. Bernhardt, E. Jefferies, J. Smallwood, How do we decide what to do? Resting-state connectivity patterns and components of self-generated thought linked to the development of more concrete personal goals. *Exp. Brain Res.* **236**, 2469–2481 (2018).

17. J. N. Mildner, D. I. Tamir, Spontaneous thought as an unconstrained memory process. *Trends Neurosci.* **42**, 763–777 (2019).

18. J. Smallwood, J. W. Schooler, D. J. Turk, S. J. Cunningham, P. Burns, C. N. Macrae, Self-reflection and the temporal focus of the wandering mind. *Conscious. Cogn.* **20**, 1120–1126 (2011).

19. D. Stawarczyk, S. Majerus, P. Maquet, A. D'Argembeau, Neural correlates of ongoing conscious experience: Both task-unrelatedness and stimulus-independence are related to default network activity. *PLOS ONE* **6**, e16997 (2011).

20. J. Smallwood, A. Fitzgerald, L. K. Miles, L. H. Phillips, Shifting moods, wandering minds: Negative moods lead the mind to wander. *Emotion* **9**, 271–276 (2009).

21. L. Koban, P. J. Gianaros, H. Kober, T. D. Wager, The self in context: Brain systems linking mental and physical health. *Nat. Rev. Neurosci.* **22**, 309–322 (2021).

22. M. F. Mason, M. I. Norton, J. D. van Horn, D. M. Wegner, S. T. Grafton, C. N. Macrae, Wandering minds: The default network and stimulus-independent thought. *Science* **315**, 393–395 (2007).

23. V. Axelrod, G. Rees, M. Bar, The default network and the combination of cognitive processes that mediate self-generated thought. *Nat. Hum. Behav.* **1**, 896–910 (2017).

24. J. R. Binder, R. H. Desai, W. W. Graves, L. L. Conant, Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* **19**, 2767–2796 (2009).

25. R. L. Buckner, D. C. Carroll, Self-projection and the brain. *Trends Cogn. Sci.* **11**, 49–57 (2007).

26. F. Galton, Psychometric experiments. *Brain* **2**, 149–162 (1879).

27. W. M. Wundt, *Outlines of Psychology* (Wilhelm Engelmann, 1897).

28. C. G. Jung, F. Riklin, in *Studies in Word-Association*, M. D. Eder, Ed. (Moffat, Yard & Company, 1904), chap. 2, pp. 8–172.

29. K. Gray, S. Anderson, E. E. Chen, J. M. Kelly, M. S. Christian, J. Patrick, L. Huang, Y. N. Kenett, K. Lewis, “Forward flow”: A new measure to quantify free thought and predict creativity. *Am. Psychol.* **74**, 539–554 (2019).

30. T. R. Marron, M. Faust, 15 Free association, divergent thinking, and creativity: Cognitive and neural perspectives, in *The Cambridge Handbook of the Neuroscience of Creativity* (Cambridge University Press, 2018), pp. 261–280.

31. R. A. Poldrack, G. Huckins, G. Varoquaux, Establishment of best practices for evidence for prediction: A review. *JAMA Psychiat.* **77**, 534–540 (2020).

32. G. Varoquaux, *P. R. Raamana, D. A. Engemann, A. Hoyos-Idrobo, Y. Schwartz, B. Thirion*, Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *Neuroimage* **145**, 166–179 (2017).

33. L. J. Chang, P. J. Gianaros, S. B. Manuck, A. Krishnan, T. D. Wager, A sensitive and specific neural signature for picture-induced negative affect. *PLOS Biol.* **13**, e1002180 (2015).

34. H. Y. Chan, A. Smidts, V. C. Schoots, A. G. Sanfey, M. A. S. Boksem, Decoding dynamic affective responses to naturalistic videos with shared neural patterns. *Neuroimage* **216**, 116618 (2020).

35. J. Chikazoe, D. H. Lee, N. Kriegeskorte, A. K. Anderson, Population coding of affect across stimuli, modalities and individuals. *Nat. Neurosci.* **17**, 1114–1122 (2014).

36. B. T. Denny, H. Kober, T. D. Wager, K. N. Ochsner, A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *J. Cogn. Neurosci.* **24**, 1742–1752 (2012).

37. C. N. Macrae, J. M. Moran, T. F. Heatherton, J. F. Banfield, W. M. Kelley, Medial prefrontal activity predicts memory for self. *Cereb. Cortex* **14**, 647–654 (2004).

38. G. Northoff, F. Bermpohl, Cortical midline structures and the self. *Trends Cogn. Sci.* **8**, 102–107 (2004).

39. D. I. Tamir, J. P. Mitchell, Disclosing information about the self is intrinsically rewarding. *Proc. Natl. Acad. Sci.* **109**, 8038–8043 (2012).

40. M. Gilead, C. Boccagno, M. Silverman, R. R. Hassin, J. Weber, K. N. Ochsner, Self-regulation via neural simulation. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 10037–10042 (2016).

41. S. Freud, *Introductory lectures on psychoanalysis*. (WW Norton & Company, 1977).

42. M. Roy, D. Shohamy, T. D. Wager, Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends Cogn. Sci.* **16**, 147–156 (2012).

43. M. L. Dixon, J. R. Andrews-Hanna, R. N. Spreng, Z. C. Irving, C. Mills, M. Girn, K. Christoff, Interactions between the default network and dorsal attention network vary across default subsystems, time, and cognitive states. *Neuroimage* **147**, 632–649 (2017).

44. L. J. Chang, E. Jolly, J. H. Cheong, K. M. Rapuano, N. Greenstein, P. H. A. Chen, J. R. Manning, Endogenous variation in ventromedial prefrontal cortex state dynamics during naturalistic viewing reflects affective experience. *Sci. Adv.* **7**, eabf7129 (2021).

45. M. Babo-Rebelo, C. G. Richter, C. Tallon-Baudry, Neural responses to heartbeats in the default network encode the self in spontaneous thoughts. *J. Neurosci.* **36**, 7829–7840 (2016).

46. K. C. Berridge, Affective valence in the brain: Modules or modes? *Nat. Rev. Neurosci.* **20**, 225–234 (2019).

47. M. Catani, F. Dell'acqua, M. Thiebaut de Schotten, A revised limbic system model for memory, emotion and behaviour. *Neurosci. Biobehav. Rev.* **37**, 1724–1737 (2013).

48. C. Tallon-Baudry, F. Campana, H. D. Park, M. Babo-Rebelo, The neural monitoring of visceral inputs, rather than attention, accounts for first-person perspective in conscious vision. *Cortex* **102**, 139–149 (2018).

49. D. Azzalini, I. Rebollo, C. Tallon-Baudry, Visceral signals shape brain dynamics and cognition. *Trends Cogn. Sci.* **23**, 488–509 (2019).

50. A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, J. L. Gallant, Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).

51. R. B. Ebitz, B. Y. Hayden, The population doctrine in cognitive neuroscience. *Neuron* **109**, 3055–3068 (2021).

52. L. Kohoutova, L. Kohoutová, J. Heo, S. Cha, S. Lee, T. Moon, T. D. Wager, C.-W. Woo, Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nat. Protoc.* **15**, 1399–1435 (2020).

53. F. J. Ruby, J. Smallwood, H. Engen, T. Singer, How self-generated thought shapes mood–The relation between mind-wandering and mood depends on the socio-temporal content of thoughts. *PLOS ONE* **8**, e77554 (2013).

54. J. Smallwood, R. C. O'Connor, Imprisoned by the past: Unhappy moods lead to a retrospective bias to mind wandering. *Cogn Emot* **25**, 1481–1490 (2011).

55. G. L. Poerio, P. Totterdell, E. Miles, Mind-wandering and negative mood: Does one thing really lead to another? *Conscious. Cogn.* **22**, 1412–1421 (2013).

56. D. Szucs, J. P. Ioannidis, Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. *Neuroimage* **221**, 117164 (2020).

57. D. Watson, L. A. Clark, G. Carey, Positive and negative affectivity and their relation to anxiety and depressive disorders. *J. Abnorm. Psychol.* **97**, 346–353 (1988).

58. L. S. Radloff, The CES-D scale: A self-report depression scale for research in the general population. *Appl. Psychol. Measur.* **1**, 385–401 (1977).

59. S. Lyubomirsky, N. D. Caldwell, S. Nolen-Hoeksema, Effects of ruminative and distracting responses to depressed mood on retrieval of autobiographical memories. *J. Pers. Soc. Psychol.* **75**, 166–177 (1998).

60. R. Spielberger, R. Gorsuch, R. Lushene, *STAI Manual for the State-Trait Anxiety Inventory 1970* (Consulting Psychologists Press, 1970).

61. K. J. Wardenaar, T. van Veen, E. J. Giltay, E. de Beurs, B. W. J. H. Penninx, F. G. Zitman, Development and validation of a 30-item short adaptation of the mood and anxiety symptoms questionnaire (MASQ). *Psychiatry Res.* **179**, 101–106 (2010).

62. W. M. Reynolds, *Suicidal Ideation Questionnaire (SIQ)* (Psychological Assessment Resources, 1987).

63. M. E. Hughes, L. J. Waite, L. C. Hawkley, J. T. Cacioppo, A short scale for measuring loneliness in large surveys: Results from two population-based studies. *Res. Aging* **26**, 655–672 (2004).

64. E. Diener, R. A. Emmons, R. J. Larsen, S. Griffin, The satisfaction with life scale. *J. Pers. Assess.* **49**, 71–75 (1985).

65. C. D. Ryff, C. L. Keyes, The structure of psychological well-being revisited. *J. Pers. Soc. Psychol.* **69**, 719–727 (1995).

66. H. Lee, K.-H. Kim, Validation of the Korean version of the mood and anxiety symptom questionnaire (K-MASQ). *Korean J. Clin. Psychol.* **33**, 395–411 (2014).

67. S. J. Kim, J. H. Kim, S. C. Youn, Validation of the Korean-ruminative response scale (K-RRS). *Korean J. Clin. Psychol.* **29**, 1–19 (2010).

68. J. Rissman, A. Gazzaley, M. D'Esposito, Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage* **23**, 752–763 (2004).

69. B. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, B. Fischl, H. Liu, R. L. Buckner, The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 1125–1165 (2011).

70. R. L. Buckner, F. M. Krienen, A. Castellanos, J. C. Diaz, B. T. Yeo, The organization of the human cerebellum estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 2322–2345 (2011).

71. E. Y. Choi, B. T. Yeo, R. L. Buckner, The organization of the human striatum estimated by intrinsic functional connectivity. *J. Neurophysiol.* **108**, 2242–2263 (2012).

72. J. N. Rouder, P. L. Speckman, D. Sun, R. D. Morey, G. Iverson, Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* **16**, 225–237 (2009).

73. L. Kasper, S. Bollmann, A. O. Diaconescu, C. Hutton, J. Heinzle, S. Iglesias, T. U. Hauser, M. Sebold, Z. M. Manjaly, K. P. Pruessmann, K. E. Stephan, The physIO toolbox for modeling physiological noise in fMRI data. *J. Neurosci. Methods* **276**, 56–72 (2017).

74. T. Yarkoni, R. A. Poldrack, T. E. Nichols, D. C. Van Essen, T. D. Wager, Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665–670 (2011).

75. D. S. Margulies, S. S. Ghosh, A. Goulas, M. Falkiewicz, J. M. Huntenburg, G. Langs, G. Bezgin, S. B. Eickhoff, F. X. Castellanos, M. Petrides, E. Jefferies, J. Smallwood, Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 12574–12579 (2016).

76. W. M. Pauli, R. C. O'Reilly, T. Yarkoni, T. D. Wager, Regional specialization within the human striatum for diverse psychological functions. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 1907–1912 (2016).

77. C. Baldassano, J. Chen, A. Zadbood, J. W. Pillow, U. Hasson, K. A. Norman, Discovering event structure in continuous narrative perception and memory. *Neuron* **95**, 709–721.e5 (2017).

78. Y. Lerner, C. J. Honey, L. J. Silbert, U. Hasson, Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* **31**, 2906–2915 (2011).

79. B. Thompson, G. M. Borrello, The importance of structure coefficients in regression research. *Educ. Psychol. Meas.* **45**, 203–209 (1985).

80. M. Reilly, R. H. Desai, Effects of semantic neighborhood density in abstract and concrete words. *Cognition* **169**, 46–53 (2017).