CORRECTED 24 AUGUST 2023; SEE FULL TEXT

COGNITIVE NEUROSCIENCE

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

Byeol Kim Lux[1,2,3], Jessica R. Andrews-Hanna[4,5], Jihoon Han[1,2,6], Eunjin Lee[1,2,6], Choong-Wan Woo[1,2,6]*

Self-relevant concepts are major building blocks of spontaneous thought, and their dynamics in a natural stream of thought are likely to reveal one's internal states that are important for mental health. Here, we conducted a functional magnetic resonance imaging experiment ($n = 62$) to examine brain representations and dynamics of self-generated concepts in the context of spontaneous thought using a newly developed free association–based thought sampling task. The dynamics of conceptual associations were predictive of individual differences in general negative affectivity, replicating across multiple datasets ($n = 196$). Reflecting on self-generated concepts strongly engaged brain regions linked to autobiographical memory, conceptual processes, emotion, and autonomic regulation, including the medial prefrontal and medial temporal subcortical structures. Multivariate pattern–based predictive modeling revealed that the neural representations of valence became more person-specific as the level of perceived self-relevance increased. Overall, this study sheds light on how self-generated concepts in spontaneous thought construct inner affective states and idiosyncrasies.

INTRODUCTION

> In the activity of association there is mirrored the whole psychical essence of the past and of the present, with all their experiences and desires. It thus becomes an index of all the psychical processes which we have but to decipher in order to understand the complete man.
>
> *by Eugen Bleuler* (*1*)

Our thoughts continuously come and go, transitioning from one topic to another even in the absence of overt task demands. This constant change and continuous flow are the key features of spontaneous thought (*2*, *3*). Previous studies have found that adults spend a considerable amount of time engaging in spontaneous, perceptually decoupled thought, a phenomenon commonly referred to as mind wandering (*4*). Since W. James featured "the stream of thought" as a major subject of psychology (*5*), researchers have posited the contents and dynamics of spontaneous thought to be important factors that can explain personality traits and mental health (*2*, *6*). For example, recurrent negative thoughts are considered a transdiagnostic phenomenon given their links to many mood and anxiety disorders (*7*). Adopting a complex dynamic systems view, spontaneous thought could be regarded as a random walk on a semantic network, in which the nodes represent autobiographical and semantic concepts, and the edges are the associations among the nodes established through past experiences (*8*–*10*). In this framework, recurrent thoughts can be regarded as sticky nodes or strong attractors of the network (*11*, *12*).

Self-relevant concepts are well-known attractors of spontaneous thought. Previous studies that examined the contents of spontaneous thought have found that the spontaneous thought contents are by no means random (*2*, *3*). Rather, thought content tends to be self-relevant in nature, encompassing personal concerns, past memories, personal goals and planning, thinking about close others, etc. (*3*, *13*–*17*). On the basis of these observations, a number of studies have suggested that self-referential processes are key functions of spontaneous thought (*6*, *13*, *14*, *18*, *19*). Moreover, self-related spontaneous thought is known to be important for long-term health and psychological well-being (*3*, *4*, *20*, *21*).

The past decade has brought increased interest in the neuroscience of spontaneous and task-unrelated thought, revealing associations with the brain's default mode network (DMN) (*22*). Although the DMN was initially featured as a signature of the brain's "resting state," it is now more broadly appreciated for its role in multiple internally guided cognitive and affective processes spanning spontaneous thought, conceptual processing, memory and future thinking, mentalizing, self-referential processes, and autonomic and visceral modulation (*21*, *23*–*25*). These findings suggest that our brain is dynamically and continuously constructing our mental and bodily life. The quantitative assessment of brain representations and cognitive underpinnings of these dynamic processes would therefore help us better understand how and why the brain generates particular patterns of spontaneous thought, resulting in a healthy or unhealthy body and mind. In the current study, we examined the spontaneous thought dynamics and their brain representations using functional magnetic resonance imaging (fMRI).

Despite the significance of dynamic, spontaneous thought in human psychology and psychopathology, few quantitative tools and methods are currently available for neuroimaging studies. To overcome this challenge, we adapted a recently developed task—the Free Association Semantic Task (FAST) (*12*)—to use in conjunction with fMRI. The FAST integrates ideas from free word association, experience sampling, and naturalistic tasks and, when adapted to a neuroimaging context, shows promise in revealing the dynamically

[1]Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon, South Korea. [2]Department of Biomedical Engineering, Sungkyunkwan University, Suwon, South Korea. [3]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA. [4]Department of Psychology, University of Arizona, Tucson, AZ, USA. [5]Cognitive Science, University of Arizona, Tucson, AZ, USA. [6]Department of Intelligent Precision Healthcare Convergence, Sungkyunkwan University, Suwon, South Korea.
*Corresponding author. Email: waniwoo@g.skku.edu

unfolding signatures of spontaneous thought. The history of using free word association as a psychological test goes back to Galton (*26*), Wundt (*27*), and Jung and Riklin (*28*), but modern psychology has largely ignored the method because of its questionable validity. However, the free association method has recently begun to receive attention again (*12*, *29*, *30*), especially for its potential to be combined with computational methods, such as natural language processing and dynamic modeling, to offer quantitative metrics for the emergence and unfolding of thought. Although early word association tasks recorded only one or two words in response to a seed word (*26*–*28*), we used a "chain" free association to better evaluate the dynamic characteristics of the stream of thought (*29*). We focused on the affective and personal aspects of participants' responses because emotionally charged and self-relevant thought topics are major ingredients of mind wandering and spontaneous thought (*4*, *13*, *17*). The free association method is known to be effective in revealing an individual's emotional and autobiographical concepts (*12*, *26*, *28*). Thus, we expect the FAST to provide a new way to probe the dynamic characteristics of affective and self-relevant thoughts, revealing personal inner affective states and idiosyncrasies that are important for human behaviors and mental health.

More specifically, this study aims to answer the following two research questions (Fig. 1A). First, are dynamic characteristics of spontaneous thought assessed with the FAST predictive of individual differences in emotional traits such as negative affectivity? Second, can we identify and decode the brain representations and dynamics of spontaneous thought? To answer these questions, we conducted an fMRI experiment (n = 62) in which participants completed the FAST involving three distinct phases (Fig. 1B). The first phase involved a "concept generation" phase, in which we asked participants to generate a concept as a word or phrase that came to their mind associated with the previous response every 2.5 s starting from a given seed word, leading to a self-generated associative concept chain. The responses were collected through an MR-compatible microphone, and participants generated 40 consecutive concepts for each seed word. The second phase involved "concept reflection," in which participants reflected on pairs of self-generated concepts in sequence for 15 s while undergoing fMRI scanning. This phase aimed to bring to mind the nature of the associative linkage between concepts generated by reflecting on the personal meaning of the concepts. This phase was our main target for the fMRI data analysis. The third part was the "postscan survey." After the fMRI scans, participants viewed their self-generated concepts once more and rated each concept using a multidimensional content scale (*13*), which consisted of items evaluating emotional valence (how positive or negative is the concept?), self-relevance (how much is the concept relevant to yourself?), time (is the concept related to the past, present, or future?), vividness (does the concept involve vivid imagery?), and safety-threat (how much is the concept safe or threatening?). For more details of the task procedure, please see Materials and Methods.

Overall, the current study aimed to develop a new experimental method that allows us to examine the dynamic characteristics of spontaneous thought and their brain representations, paving the way for quantitative modeling of spontaneous thought dynamics. In addition, our findings would provide a deeper understanding of where in the brain self-generated, endogenous thoughts are represented and how self-relevance modulates the brain's affective representations.

RESULTS

Dynamics of spontaneous thought probed with the FAST

Figure 1 (C and D) shows data from a representative participant to illustrate how FAST responses can reveal the characteristic topics and dynamic features of an individual's spontaneous thought. As shown in Fig. 1C, FAST responses can be viewed in the context of a high-dimensional state space of some phenomenological characteristics. This participant's concept association initially flowed from a given seed word, "tear," to negatively valence concepts ("cry," "sadness," and "suicide"), then moved to societal and less self-relevant thought topics ("society," "Bitcoin," "bubble," "fail," and "bank"), and transitioned to personal topics ("house," "company," and "brother"). This example highlights that the FAST can reveal topics of spontaneous thought that range from personal narratives to societal events and issues (e.g., the data were collected in early 2018 when the Bitcoin crash occurred).

FAST responses can also be viewed as a directed graph in which the nodes are the response concepts, and the directed edges are the connections from the previous concepts to the following associative concepts (Fig. 1D). Key hubs of this participant's graph, including "sadness," "brother," and "classroom," can be seen as an attractor that is densely connected to other related concepts. FAST responses tended to come back to these nodes as if they had strong gravity in this personal semantic state space.

Free association dynamics predictive of individual differences in negative affectivity

To answer our first research question (Q1 in Fig. 1A, i.e., whether the affective dynamics of spontaneous thought assessed with the FAST are predictive of individual differences in negative affectivity), we used the Markov chain analysis to create input features for machine learning to build a predictive model of general negative affectivity based on the transitional dynamics on the content dimensions. Before the analysis, we ensured that the postscan survey ratings reflected participants' in-scanner experience by showing that (i) the in-scanner heart rates were substantially modulated by the levels of postscan valence ratings and (ii) the valence ratings were consistent with the emotion ratings intermittently obtained during the fMRI scans (see figs. S1 and S2 for detail).

We first defined discrete states for the Markov chain analysis by dividing each content dimension into two or three discrete states, as shown in Fig. 2A. We divided valence, time, and safety-threat, which ranged from −1 to 1, into three discrete states (−1 to −0.33, −0.33 to 0.33, and 0.33 to 1; for valence, the three discrete states were negative/neutral/positive; for time, past/present/future; and for safety-threat, threatening/neutral/safe). For the self-relevance and vividness dimensions that ranged from 0 to 1, we divided them into two discrete states (0 to 0.5 and 0.5 to 1, which corresponded to low and high for both dimensions, respectively). We then calculated the transition probability, defined as the probability of making transitions from one to another discrete state on each dimension. We also calculated the steady-state probability, defined as the probability of converging to one state when the transition processes were sufficiently repeated. In addition to these dynamic features from the Markov chain analysis, we used each affective dimension's mean and variance as predictor variables for the subsequent predictive modeling. Many of these dynamic features were relatively stable over a 7-week interval (for their test-retest reliability, see table S1).

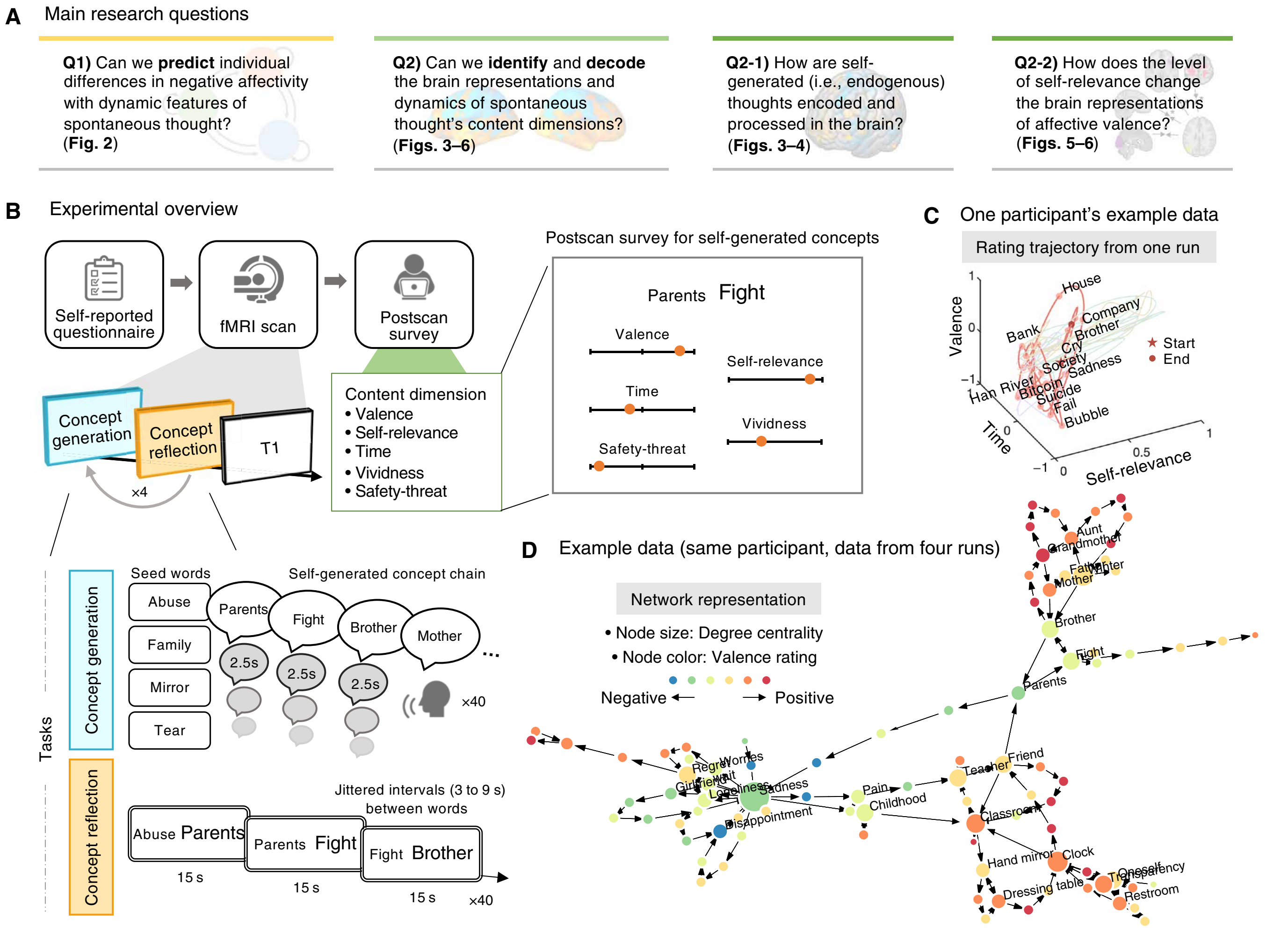
A
Main research questions
Q1) Can we predict individual differences in negative affectivity with dynamic features of spontaneous thought? (Fig. 2)
Q2) Can we identify and decode the brain representations and dynamics of spontaneous thought's content dimensions? (Figs. 3–6)
Q2-1) How are self-generated (i.e., endogenous) thoughts encoded and processed in the brain? (Figs. 3–4)
Q2-2) How does the level of self-relevance change the brain representations of affective valence? (Figs. 5–6)
B
Experimental overview
Self-reported questionnaire
fMRI scan
Postscan survey
Postscan survey for self-generated concepts
Parents Fight
Valence
Self-relevance
Time
Vividness
Safety-threat
Content dimension
• Valence
• Self-relevance
• Time
• Vividness
• Safety-threat
Concept generation
Concept reflection
T1
×4
Tasks
Seed words
Abuse
Family
Mirror
Tear
Self-generated concept chain
Parents
Fight
Brother
Mother
2.5s
×40
Jittered intervals (3 to 9 s) between words
Abuse Parents
Parents Fight
Fight Brother
15 s
×40
C
One participant's example data
Rating trajectory from one run
Valence
Time
Self-relevance
Start
End
House
Company
Bank
Brother
Cry
Society
Sadness
Han River
Bitcoin
Suicide
Fail
Bubble
D
Example data (same participant, data from four runs)
Network representation
• Node size: Degree centrality
• Node color: Valence rating
Negative
Positive
Aunt
Grandmother
Father
Mother
Brother
Fight
Parents
Regret
Worries
Girlfriend
Loneliness
Sadness
Disappointment
Pain
Childhood
Teacher
Friend
Classroom
Hand mirror
Clock
Dressing table
Transparency
Restroom

Fig. 1. Research questions and experimental overview. (**A**) Main research questions and their corresponding result figures. (**B**) An experimental overview. Participants completed a battery of self-report questionnaires before the fMRI scans. During the fMRI scans, participants underwent the FAST, which consisted of three main parts—concept generation, concept reflection, and postscan survey. The fMRI experiment had four runs, each of which included the concept generation and concept reflection tasks. For the concept generation task, we asked participants to report a word or phrase that came to their mind in response to the previous concept every 2.5 s starting from a given seed word. The seed words for the first session included family, tear, mirror, and abuse. After participants generated 40 concept-chain responses, we showed the two consecutive responses and asked them to think about the target (i.e., the second) concept's personal meaning for 15 s. After the scan, we asked participants to complete a postscan survey on the 160 self-generated concepts. We showed the two consecutive concepts again and asked participants to rate the target concept in terms of their valence, self-relevance, time, vividness, and safety-threat. (**C**) One participant's example data are shown on the three-dimensional space of valence, self-relevance, and time. The dots indicate self-generated responses, and the pink line indicates data from one run. A red star indicates the start of the run, and the red dot indicates the ending point. (**D**) A network representation of one participant's data. The dots represent the self-generated responses, and the arrows show the direction of the concept generation. The dot colors represent the averaged valence scores of each concept, and the dot size indicates the degree centrality (i.e., how many edges are connected to the node).

For the predictive modeling, we used least absolute shrinkage and selection operator (LASSO) regression to predict individual differences in general negative affectivity with leave-one-subject-out cross-validation (LOSO-CV). The number of predictor variables for the final model was determined on the basis of the cross-validation performance with the training data (n = 62; one participant was excluded because of excessively few responses). We then tested the final model on three testing datasets to evaluate two different generalizability types—seed words and response modality. First, we tested the model on retest data with an average 7-week interval, in which a different set of seed words were used on a subset of participants from the training dataset (n = 30). Second, we tested the model on an independent test data (n = 117) collected from a new set of participants using a web-based FAST experiment, which collected the response through typing instead of speaking and with a longer time limit for a response (for more detail of the web-based FAST, please see Materials and Methods). Third, we tested the model on retest web-based FAST data with an average 7-week interval, in which, again, a different set of seed words were used on a subset of participants from the web-based FAST dataset (n = 49). As the outcome variable, we used factor scores from a factor analysis of subscales from self-report questionnaires measuring multiple aspects of general negative affectivity (for the details of the factor analysis results and self-report questionnaires, please refer to tables S2 to S4).

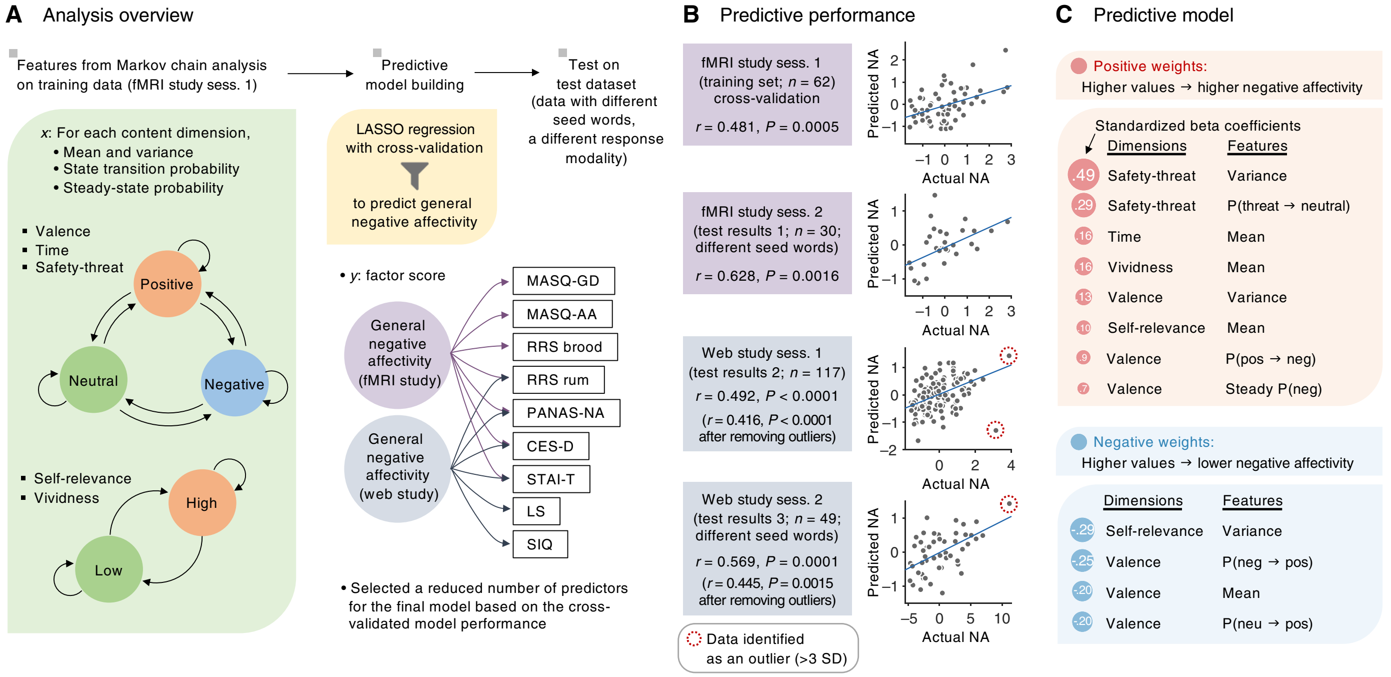
A Analysis overview
Features from Markov chain analysis on training data (fMRI study sess. 1)
Predictive model building
Test on test dataset (data with different seed words, a different response modality)
x: For each content dimension,
• Mean and variance
• State transition probability
• Steady-state probability
▪ Valence
▪ Time
▪ Safety-threat
Positive
Neutral
Negative
▪ Self-relevance
▪ Vividness
High
Low
LASSO regression with cross-validation
to predict general negative affectivity
• y: factor score
General negative affectivity (fMRI study)
General negative affectivity (web study)
MASQ-GD
MASQ-AA
RRS brood
RRS rum
PANAS-NA
CES-D
STAI-T
LS
SIQ
• Selected a reduced number of predictors for the final model based on the cross-validated model performance
B Predictive performance
fMRI study sess. 1 (training set; n = 62) cross-validation
r = 0.481, P = 0.0005
fMRI study sess. 2 (test results 1; n = 30; different seed words)
r = 0.628, P = 0.0016
Web study sess. 1 (test results 2; n = 117)
r = 0.492, P < 0.0001
(r = 0.416, P < 0.0001 after removing outliers)
Web study sess. 2 (test results 3; n = 49; different seed words)
r = 0.569, P = 0.0001
(r = 0.445, P = 0.0015 after removing outliers)
Data identified as an outlier (>3 SD)
Predicted NA
Actual NA
−1 0 1 2 3
2 1 0 −1
1 0 −1
−2
0 2 4
−5 0 5 10
C Predictive model
Positive weights:
Higher values → higher negative affectivity
Standardized beta coefficients
Dimensions Features
.49 Safety-threat Variance
.29 Safety-threat P(threat → neutral)
.16 Time Mean
.16 Vividness Mean
.13 Valence Variance
.10 Self-relevance Mean
.9 Valence P(pos → neg)
.7 Valence Steady P(neg)
Negative weights:
Higher values → lower negative affectivity
Dimensions Features
-.29 Self-relevance Variance
-.25 Valence P(neg → pos)
-.20 Valence Mean
-.20 Valence P(neu → pos)

Fig. 2. Markov chain–based predictive modeling of negative affectivity. (**A**) Analysis overview. The input features for the predictive modeling included the state transition and steady-state probabilities estimated with the Markov chain analysis and the mean and variance of the content dimension ratings. With these features, we developed predictive models of general negative affectivity, which we modeled with factor analyses. Although the actual factor models also included self-report questionnaires for a factor of general positive affectivity, here, we show the questionnaires only for the general negative affectivity factor for brevity. We used a LASSO regression as a fitting algorithm. For details about the analysis and questionnaires, please see Materials and Methods. (**B**) Model performance. From top to bottom, the plots show (i) 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 (ii) the second session retest data of the fMRI study with different seed words ($n = 30$), (iii) the first session data of the FAST web study ($n = 117$), and (iv) the second session retest 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 a robust correlation. After removing outliers identified with three SDs, the prediction-outcome correlation remained significant. NA, negative affectivity. (**C**) To interpret the final model, we examined the standardized beta coefficients of the input features of the model. From the LASSO regression, a total of 12 features were selected. The features in red indicate positive weights, whereas the features in blue indicate negative weights.

As shown in Fig. 2B, the final predictive model showed significant prediction performance across four datasets; for the training dataset ($n = 62$) with LOSO-CV, the prediction-outcome correlation between the actual and predictive values was $r = 0.481$, $P = 0.0005$, two-tailed, one-sample t test; for the retest data ($n = 30$) with different seed words, $r = 0.628$, $P = 0.0016$ (with LOSO-CV, $r = 0.501$, $P = 0.0105$); for the web-based FAST experiment ($n = 117$), $r = 0.492$, $P < 0.0001$; and for the web-based FAST retest ($n = 49$) with different seed words, $r = 0.569$, $P = 0.0001$. After removing outliers identified with three SDs, the prediction-outcome correlation remained to be significant ($r = 0.416$, $P < 0.0001$, $n = 115$ after removing two outliers from the web-based FAST; $r = 0.445$, $P = 0.0015$, $n = 48$ after removing one outlier participant from the web-based FAST retest). Given that these three independent test datasets had slightly different experimental parameters, such as seed words and response modality, these results demonstrated the robustness of the task and the predictive model.

We then examined the standardized beta coefficients of the 12 behavioral dynamic features to determine which predictors contributed significantly to the final model of general negative affectivity (Fig. 2C). The beta coefficients indicated that the participants who showed (i) a higher variance of the safety-threat and valence scores; (ii) a higher transition probability from threatening to neutral states, (iii) a higher mean score on the time, vividness, and self-relevance dimensions; (iv) a higher transition probability from positive to negative states; and (v) a higher steady-state probability for the negative state were likely to report a higher level of general negative affectivity. Conversely, participants who showed (i) a higher variance on the self-relevance score, (ii) a higher transition probability from the negative or neutral to positive states, and (iii) a higher mean score on the valence dimension (i.e., more positive) tended to report a lower level of general negative affectivity. An additional analysis comparing the relative contributions of Markov chain–based features versus non–Markov chain features suggested that the Markov chain–based features (i.e., state transition dynamics) explained a substantial amount of variance above and beyond the non–Markov chain features (table S5).

We also trained an additional predictive model with a subset of the content dimension—valence, self-relevance, and time, and thus, we called this reduced model a VST model—to see whether these three dimensions were enough to predict the level of general negative affectivity. We chose these three dimensions because valence and self-relevance were highly correlated with safety-threat and vividness, respectively (fig. S3A). The VST model also showed significant predictions across four datasets with $r = 0.409$ to 0.677, $P = 0.0042$ to $P < 0.0001$, and seven of eight selected features overlapped with the original full model features (fig. S4).

Brain activation patterns during the concept reflection phase

To answer our second research question (Q2-1 in Fig. 1A; identifying the brain representations of the content dimensions of spontaneous thought), we first examined brain activation patterns while participants were reflecting on the self-generated concepts in the context of their conceptual associations. We decided to make the concept reflection phase our main target for the fMRI analyses before we conducted the experiment because of the possibility of high levels of head motion during the concept generation phase (due to speaking, but see fig. S5 for the comparisons of head motion between the two phases). In addition, the concept reflection phase, in which each trial was 15 s long, could generate the brain representations of self-generated concepts more effectively than the concept generation phase, in which each trial was 2.5 s long including speaking.

As shown in Fig. 3A, brain regions spanning the hippocampus, amygdala, and parts of the somatomotor network and default mode network engaged to a greater degree during conceptual self-reflection (warm color) compared to fixation baseline. In contrast, the visual network was engaged to a greater degree during baseline (cool color) than during conceptual self-reflection and thus appeared "deactivated" during reflection (see fig. S6 for the large-scale network definitions used for identification purposes).

Further investigation into the temporal shape of these hemodynamic response patterns using a finite impulse response (FIR) model revealed that the visual cortex "deactivation" was driven by a transient increase in activity around 3 s after the stimulus onset, followed by a large decrease afterward (Fig. 3B). *K*-means clustering on the FIR signal across the brain showed that the visual cortex, some brain regions within the ventral attention network, and the thalamus formed a cluster. This cluster (purple in Fig. 3B) showed a transient activity right after the stimulus onset, likely reflecting perceptually guided and attentional orienting processes. Two additional clusters (green and yellow in Fig. 3B) emerging from the clustering analysis mainly consisted of default mode and limbic network, lateral prefrontal cortex, and hippocampus and amygdala regions. Both clusters showed a delayed peak of brain activity around 7 to 10 s after the stimulus onset. Another cluster (red in Fig. 3B) that had a large overlap with the somatomotor network showed a negative peak around 5 s after the stimulus onset. Given that the concept reflection did not involve any actual sensorimotor experience, this deactivation seemed reasonable. However, as shown in Fig. 3B, the brain activation level within this cluster showed a slow recovery and turned into positive activation toward the end of the trial. Further characterization of the clusters with meta-analysis database and cortical hierarchy suggested that our task strongly engaged brain

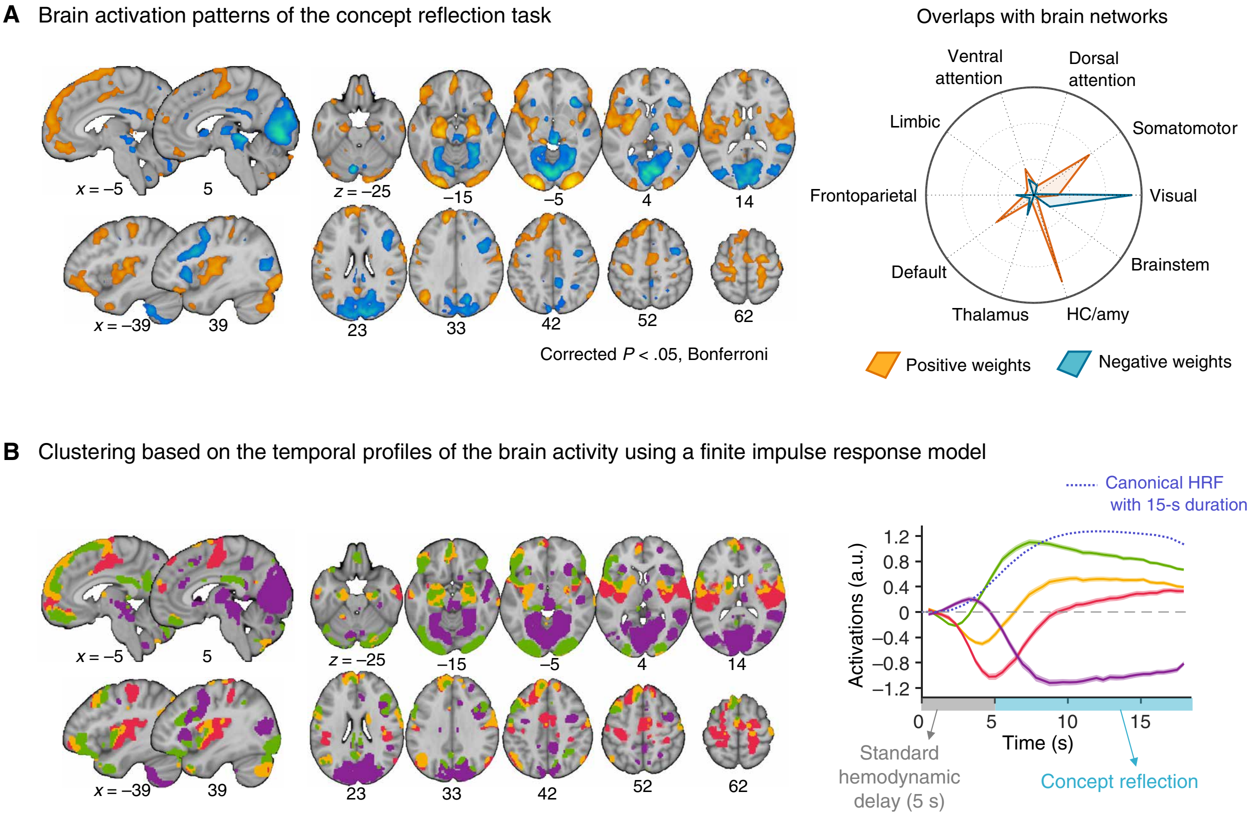

Fig. 3. Brain activation patterns during the concept reflection task. (**A**) The basic contrast map for concept reflection versus baseline thresholded at $P < 0.05$ with Bonferroni correction. The radial plot shows the relative proportions of overlapping voxels between the thresholded map and large-scale networks (or regions), given the total number of voxels within each network (or each region). For the definitions of large-scale networks and regions, see fig. S6. HC, hippocampus. (**B**) To better understand the temporal patterns of brain activity during concept reflection, we conducted a clustering analysis based on a finite impulse response (FIR) model. We identified four clusters, and the plot on the right shows the time-course of averaged activity across voxels of each cluster. Shading represents the SEM. On the *x* axis, zero represents the onset of the trial, and the gray area represents the standard hemodynamic delay, which peaks at around 5 s. The blue dotted line shows the canonical hemodynamic response function (HRF) for an event with a 15-s duration. a.u., arbitrary units.

regions linked to autobiographical memory, conceptual processes, emotion, and autonomic regulation, which largely overlapped with the transmodal end in the principal gradient of cortical hierarchy (figs. S7 to S9).

Multivariate pattern–based predictive models of self-relevance and valence

To further investigate our second research question ("can we identify and decode the brain representations of affective qualities of spontaneous thought?" in Fig. 1A), we developed whole-brain multivariate pattern–based predictive models for the content dimension ratings. To prepare training data, we grouped trials into quartiles representing four levels of each content dimension scale (for each participant) and then averaged the brain and rating data, resulting in four brain maps and four averaged rating scores per person for each dimension. After concatenating all participants' data ($n = 61$), we trained principal components regression (PCR) models for each content dimension and estimated model performance using two types of cross-validation methods—LOSO-CV and random-split cross-validation (RS-CV) (*31*, *32*). The cross-validated prediction performance was significant for self-relevance [correlation between actual and predicted ratings: with LOSO-CV, mean $r = 0.304$, $z = 4.400$, $P < 0.0001$, two-tailed, bootstrap tests, mean squared error (mse) = 0.155; with RS-CV, $r = 0.276$, mse = 0.156; Fig. 4A], while other dimensions showed relatively poor prediction performance, with LOSO-CV, mean $r = 0.185$, mse = 0.399 for valence; mean $r = 0.166$, mse = 0.319 for safety-threat; mean $r = -0.064$, mse = 0.228 for time; and $r = -0.015$, mse = 0.182 for vividness. With RS-CV, mean $r = 0.152$, mse = 0.427 for valence; mean $r = 0.147$, mse = 0.323 for safety-threat; mean $r = -0.012$, mse = 0.223 for time; and $r = -0.002$, mse = 0.183 for vividness.

Among the dimensions that showed poor prediction performance, the valence result was unexpected because previous studies have shown reasonable performance in predicting positive versus negative emotional valence. For example, Chang *et al.* (*33*) reported that a whole-brain pattern–based marker could predict negative emotion ratings induced by pictures with high prediction performance. Other studies also reported that regional brain activity patterns could classify the positive versus negative valence with significant classification accuracy (*34*, *35*). However, unlike the previous studies, which used exogenous stimuli to evoke emotions, such as pictures (*33*), movies (*34*), or tastants (*35*), the current study used self-generated, endogenous stimuli, which could have a potential impact on the semantic representations of emotional valence in the brain. Thus, we hypothesized that if we trained a predictive model only with the data from trials with low self-relevance scores, then we might be able to achieve a significant prediction performance similar to the previous studies. To test this hypothesis, we separately trained two models of valence, one for the low self-relevance trials (self-relevance scores ≤ 0.5) and the other for the high self-relevance trials (self-relevance scores > 0.5). Other analysis procedures were the same as the previous.

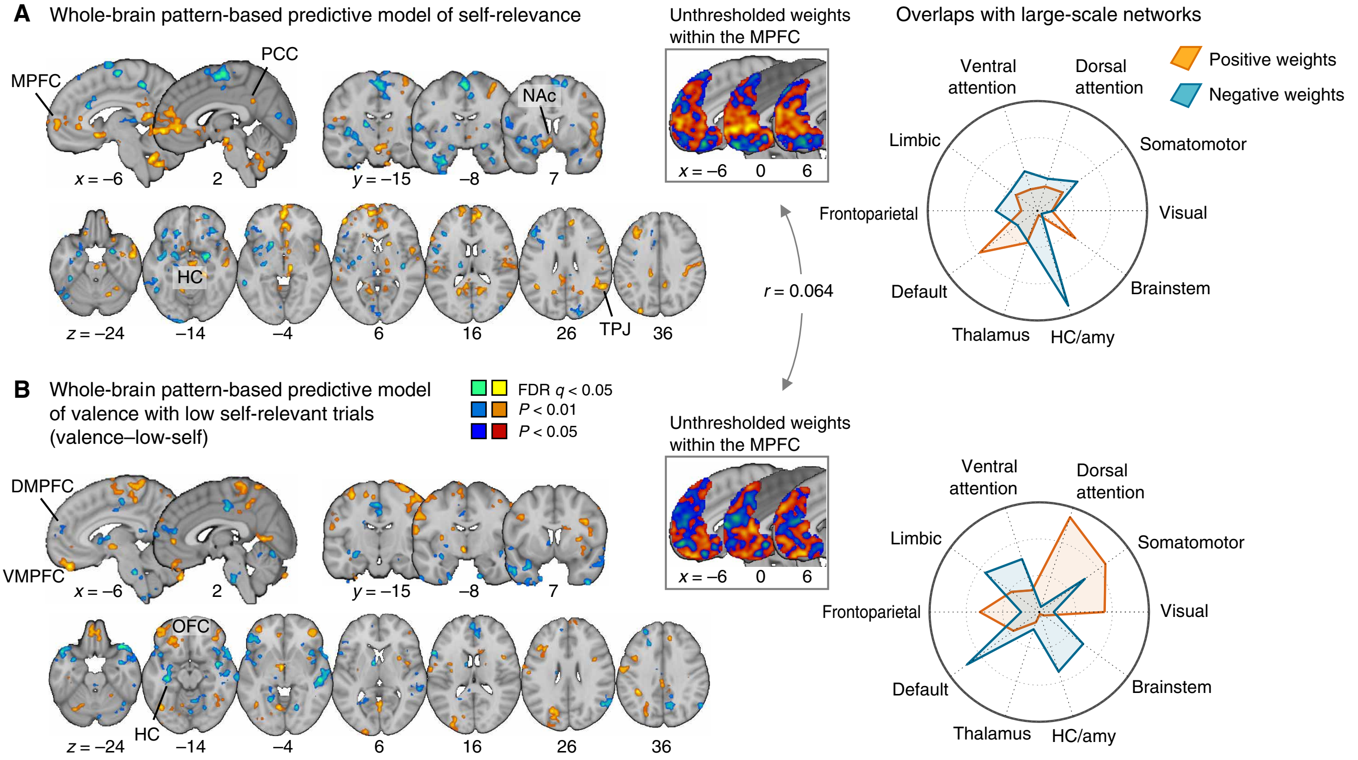

Fig. 4. Multivariate pattern–based predictive models of self-relevance and valence. (**A**) Self-relevance model. The map shows the voxels that reliably contributed to the prediction of self-relevance scores based on bootstrap tests (thresholded at FDR $q < 0.05$, two-tailed). Thresholding was performed for the purpose of display; all weights were used in the prediction. We also pruned the map using two additional more liberal thresholds, uncorrected $P < 0.01$ and $P < 0.05$, two-tailed, to show the extent of activation clusters. The radial plot shows the relative proportions of overlapping voxels between the thresholded map and large-scale networks, given the total number of voxels within each network. (**B**) Valence model for the trials with low self-relevance scores (valence–low-self model). Insets show the two models' unthresholded predictive weights within the MPFC. The correlation value ($r = 0.064$) indicates the pattern similarity of the MPFC weights between the two models.

As hypothesized, we found that the valence model trained only on the low self-relevance trials (named the "valence–low-self" model; Fig. 4B) showed a better and significant prediction performance, mean $r = 0.307$, $z = 3.808$, $P < 0.0001$, bootstrap test, mse = 0.362 with LOSO-CV and $r = 0.303$, mse = 0.364 with RS-CV, than the valence model trained on the data with high self-relevance, mean $r = 0.031$, $z = 0.403$, $P = 0.6872$, mse = 0.448 with LOSO-CV and $r = 0.060$, mse = 0.426 with RS-CV. To further validate this valence–low-self model, we tested the model on an independent study dataset from Chang *et al.* (*33*). We chose this study dataset because it used exogenous emotional stimuli to induce emotions [i.e., the International Affective Picture System (IAPS) pictures] and was publicly available from NeuroVault (https://identifiers.org/neurovault.collection:503). As shown in Fig. 5A, when we applied the valence–low-self model on the beta images corresponding to five-point negative emotion ratings ranging from 1 (neutral) to 5 (strongly negative), our model showed significant predictions across two independent datasets. The first dataset was the training data in the original study ($n = 121$, mean $r = 0.225$, $P < 0.00001$, two-tailed, bootstrap tests) (*33*), and the second dataset was the testing data in the original study ($n = 61$, $r = 0.256$, $P = 0.0002$). These results provided evidence for the generalizability of our valence–low-self model to emotions evoked with exogenous visual stimuli.

To further understand the neurobiological meaning and validity of the predictive models, we visualized the thresholded predictive maps of self-relevance (Fig. 4A) and valence–low-self models (Fig. 4B) based on bootstrap tests with 10,000 iterations and the false discovery rate (FDR) $q < 0.05$, identifying brain voxels that made reliable

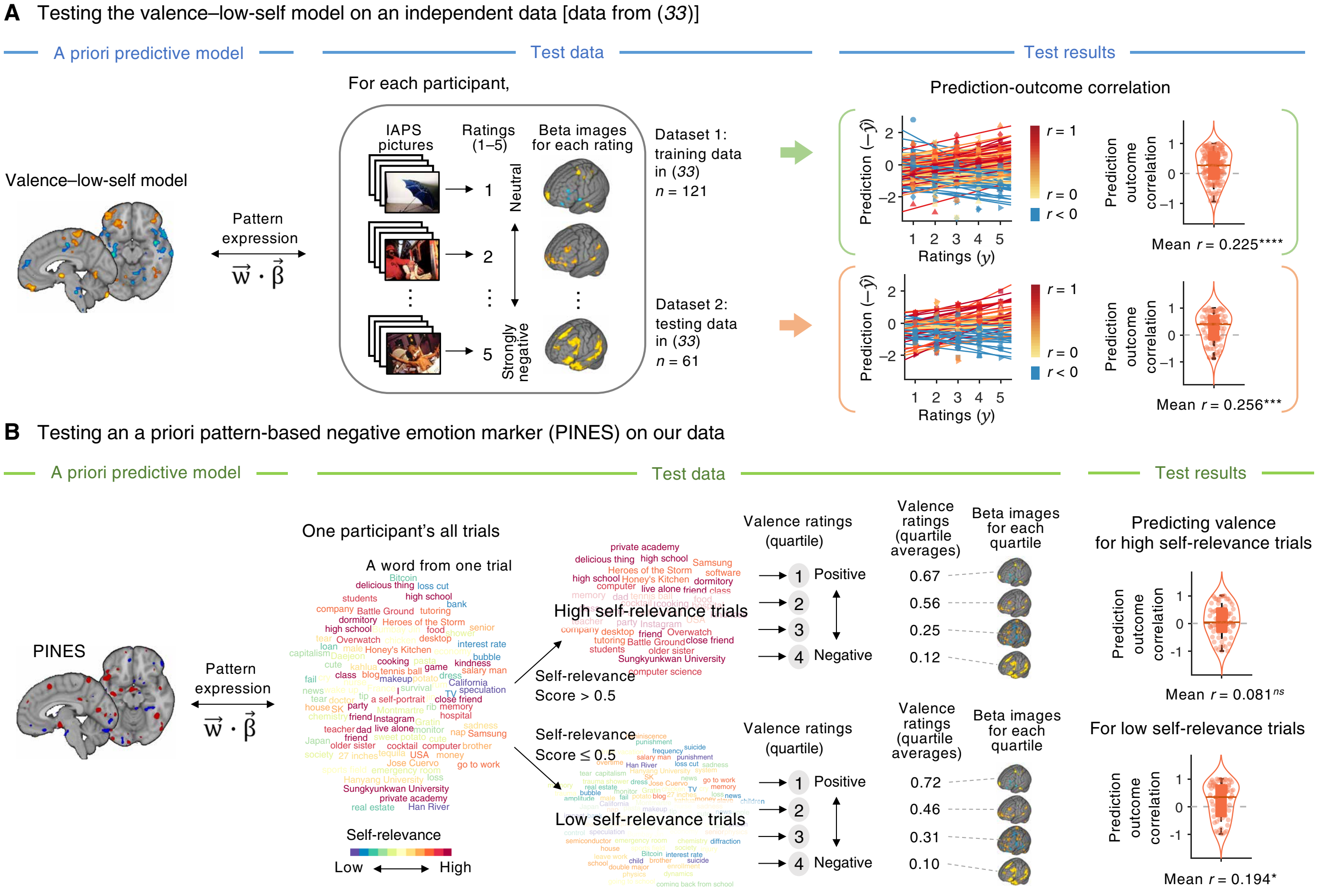

Fig. 5. Cross-testing of two a priori models of valence on two independent datasets. (**A**) To further validate our valence–low-self model, we tested the model on an independent study dataset, in which negative emotions were induced using the IAPS pictures (*33*). To this end, we applied the valence–low-self model to the beta images from the two independent datasets corresponding to five-point negative emotion ratings ranging from 1 (neutral) to 5 (strongly negative). We obtained predicted ratings by calculating a dot product of each vectorized test image data with the model weights. The plot on the right shows the actual versus predicted ratings. For convenience, we added the negative sign to the predicted ratings to make the expected prediction into positive correlations. Each colored line in the plot represents an individual participant's data (red, higher *r*; yellow, lower *r*; blue, $r < 0$). The violin and box plots display the distributions of within-participant prediction-outcome correlations. ***$P < 0.001$ and ****$P < 0.0001$, bootstrap tests, two-tailed. (**B**) To test whether the level of self-relevance modulated the brain representation of valence, we tested an a priori neuroimaging emotion marker, PINES (*33*) on our data, separately for trials with high self-relevance scores (>0.5) versus low self-relevance scores (≤0.5). First, we divided one individual's high and low self-relevance trials into valence quartiles. Then, we averaged beta images and valence ratings within each quartile to use as test data. We found pattern expression values with a dot product between the PINES weights and the test data of high and low self-relevant trials. Thus, the model performance was based on the quartile data based on the valence scores, and the violin and box plots on the right show the distributions of within-participant prediction-outcome correlations. ns, not significant; *$P < 0.05$, bootstrap tests, two-tailed.

contributions to the prediction. For the self-relevance model, multiple brain regions within the default mode and limbic networks appeared to be important, including the medial prefrontal cortex (MPFC), posterior cingulate cortex (PCC), temporoparietal junction (TPJ), temporal pole (TP), hippocampus, and nucleus accumbens (NAc), consistent with previous literature (*16*, *36*–*39*). Similarly, the valence–low-self model also identified important predictors within the default mode and limbic networks, such as the dorsomedial prefrontal cortex (DMPFC) and ventromedial prefrontal cortex (VMPFC), orbitofrontal cortex (OFC), and hippocampus. However, the predictive weight patterns within these regions were quite different between the self-relevance and valence–low-self models. For example, as shown in the insets of Fig. 4, which presented the unthresholded weights of the self-relevance and valence–low-self models within the MPFC, the self-relevance model showed a negative → positive → negative gradient from dorsal to ventral parts of the MPFC. In contrast, the valence–low-self model showed a negative → positive gradient from dorsal to ventral MPFC. The pattern similarity of the unthresholded predictive weights within the MPFC between the two models was low, $r = 0.064$. In addition to the default mode and limbic network regions, many voxels within the somatomotor and ventral and dorsal attention networks were among the important features of the models, suggesting that the information about the levels of self-relevance and valence involves many brain regions distributed across multiple brain systems. z-scoring the outcome variables (i.e., self-relevance and valence scores) yielded similar predictive maps and results (fig. S10), suggesting that the within-subject variance was the main driver of the results.

Note that we did not further examine the predictive models of the other three dimensions, i.e., vividness, safety-threat, and time, given that the principal component analysis (PCA) results suggested three main principal components in the content dimensions. As shown in fig. S3A, the valence and safety-threat dimensions were highly correlated, and the self-relevance and vividness were also highly correlated. Therefore, by modeling valence and self-relevance, we should be able to cover the first two principal components. Regarding the time dimension, its predictive model did not perform well, and thus, we did not further examine the model here. In fig. S3B, we presented the univariate general linear model results with the three principal components.

Idiosyncratic brain representations of emotional valence for high self-relevance trials

One of the intriguing observations in the previous section was the poor prediction performance of the valence model when it was trained on the high self-relevance trials. We hypothesized that valence information for the high self-relevance trials would be represented with more idiosyncratic brain activity patterns than for the low self-relevance trials. To test the hypothesis, we first tested whether an a priori multivariate pattern–based emotion marker provided a similar pattern of results. We used the picture-induced negative emotion signature (PINES) (*33*), which has shown its sensitivity and specificity in predicting the level of negative emotions across multiple studies (*33*, *40*). As presented in Fig. 5B, the results were consistent with our findings in the previous section—the PINES was able to predict the valence ratings only for the data from the low self-relevance trials, mean $r = 0.194$, $P = 0.0171$, two-tailed, bootstrap tests, but not for the high self-relevance trials, $r = 0.081$, $P = 0.2813$, although the difference was not significant, $z = 1.260$, $P = 0.208$, 95% confidence interval = −0.163 to 1.266, two-tailed, bootstrap test with Fisher z-transformation. These results suggest that the data from low self-relevance trials had some shared pattern information for valence common across participants, which the PINES could capture. In contrast, the valence information from high self-relevance trials cannot be decoded with the population-level emotion marker.

We then used an idiographic predictive modeling approach to quantifying the between-subject variability of predictive weights for high versus low self-relevance data. We first split the trials into high and low self-relevance groups (using 0.5 as a cutoff score). We then trained two valence models per person—one for high self-relevance trials and the other for low self-relevance trials using PCR with 5-fold cross-validation. Although the prediction performances were not different between the high versus low self-relevance predictive models ($t_{60} = 1.886$, $P = 0.0641$, two-tailed, paired t test; the middle panel of Fig. 6A), the SDs of the predictive weights across participants were significantly higher in the high self-relevance models than the low self-relevance models ($t_{211362} = 867.59$, $P < 0.0001$; the right panel of Fig. 6A). Moreover, a group-level model trained on the high self-relevance data from all participants showed significantly worse prediction performance than the idiographic model built individually with the same data ($t_{60} = 10.005$, $P < 0.0001$, two-tailed, paired t test; the middle panel of Fig. 6A). These results provide additional converging evidence that the brain representations of emotional valence are shared across people when the stimulus is less self-relevant, but they become idiosyncratic across people when the stimulus is highly self-relevant.

We then examined where in the brain showed similar or distinct patterns of predictive weights between the valence models for high versus low self-relevance using a searchlight-based pattern similarity analysis method. As shown in Fig. 6B, the brain areas that showed low pattern similarity (in blue, Bayes factor in favor of null hypothesis $BF_{01} > 6$) were larger and more widely distributed across the whole brain than the brain areas that showed high pattern similarity (in red, Bayes factor in favor of alternative hypothesis $BF_{10} > 6$). Brain regions with low pattern similarity included cortical and subcortical regions within the default mode and limbic networks, such as the VMPFC, perigenual anterior cingulate cortex, PCC, TP, hippocampus, and amygdala, and regions within the somatomotor network, such as supplementary motor area, right insula, and thalamus. Brain regions that showed high pattern similarity included the subgenual anterior cingulate cortex, left dorsal posterior insula, and right dorsal lateral prefrontal cortex. To summarize, these findings supported our hypothesis that the valence representations in the brain become more diverse and idiosyncratic across individuals as the stimuli become more self-relevant.

DISCUSSION

In this study, we replicated and significantly expanded our previous work using the FAST to assess the dynamic characteristics of the natural stream of thought (*12*). Through an fMRI experiment ($n = 62$) combined with the FAST, we aimed to test whether we could predict individual differences in negative affectivity with dynamics of spontaneous thought and whether we could identify and decode the brain representations and dynamics of phenomenological characteristics of spontaneous thought. Our main findings can be summarized as follows: (i) We developed a Markov chain–based predictive model of negative affectivity that generalized across

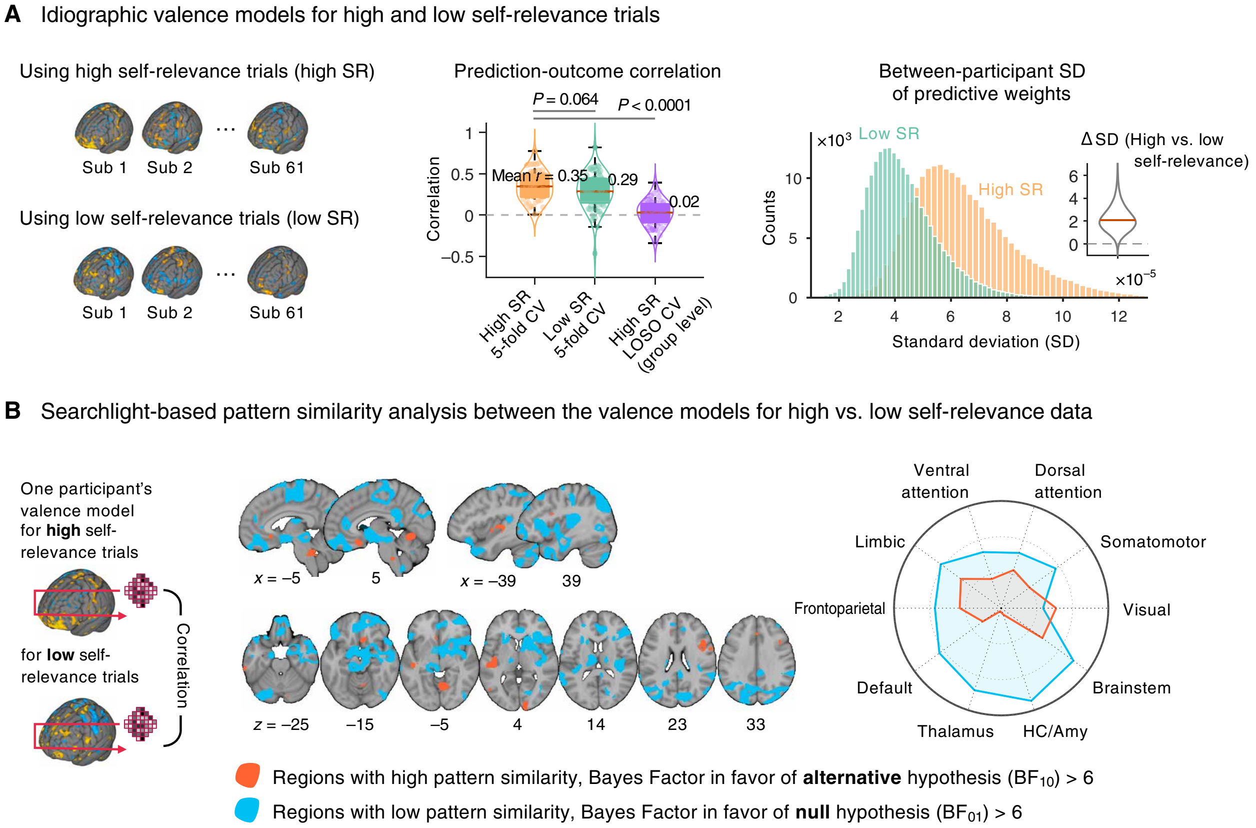

Fig. 6. Idiographic predictive modeling of valence. (**A**) We used an idiographic predictive modeling approach to quantifying the between-participant variability of predictive weights for high versus low self-relevance data. We trained two valence models per person—the first model used data from the trials with the high self-relevance scores (>0.5; high SR) and the other used data from the trials with low self-relevance scores (≤0.5; low SR). In addition, we trained a group-level valence model by concatenating all participants' high self-relevance data with LOSO-CV for comparison. The violin and box plots in the middle show the prediction performance of two idiographic models with 5-fold cross-validation and the group-level model with LOSO-CV. Each dot represents each participant. The histograms on the right show the distributions of the between-participant SD of predictive weights, and the inset violin plot shows the differences in the SD between two valence models across voxels. (**B**) We examined which brain regions displayed similar or distinct spatial patterns of predictive weights between two valence models with a searchlight-based pattern similarity analysis. For this, we first created a searchlight with a radius of five voxels and scanned it throughout the whole brain with a step size of four voxels. We then calculated correlation coefficients between the two models' predictive weights within each searchlight. The red regions showed Bayes factor values in favor of having distinct patterns between two models ($BF_{10} > 6$), whereas the blue regions were the opposite, i.e., in favor of having similar patterns ($BF_{01} > 6$). The radial plot shows the relative proportions of overlapping voxels between the thresholded map and large-scale networks, given the total number of voxels within each network.

multiple independent datasets. (ii) Reflecting on one's associative concepts strongly activated brain regions related to autobiographical memory, emotion, and internal and conceptual processing. (iii) Predictive modeling of content dimension ratings of the self-generated concepts revealed that the brain representations of valence became more idiosyncratic as the level of self-relevance increased.

First, we successfully validated our new dynamic FAST, highlighting its great potential as a versatile thought-sampling research tool for psychological assessment and neuroimaging studies. Here, we showed that the FAST provided information about personally important thought topics and their semantic networks that could reveal each individual's unique cognitive and phenomenological characteristics. The history of using free association as a psychological method to reveal one's internal thoughts and emotional states goes back to the late 1800s (*26*–*28*). The method gained its increasing popularity since Freud (*41*), who used free association as the primary technique for his psychoanalysis, but modern psychology abandoned the method because of its questionable scientific validity and reliability. However, recent advances in computational tools and techniques provide an opportunity to revive the free association method with many potential use cases. For example, as shown in the current study, the dynamic modeling combined with machine learning has the potential to be used as an assessment tool for depression and anxiety in adjunct to self-report. Even in the absence of participant's own self-report ratings, we previously showed in a behavioral FAST study that the affective dynamics of thought predicted individual differences in trait rumination (*12*), a common symptom of mood and anxiety disorders. Thus, if we can implement an automated sentiment analyzer into the analysis pipeline, we can shorten the task time markedly, providing a possibility to use the FAST as a web- or mobile-based monitoring tool for depression and anxiety. The FAST has the potential to be used for other clinical conditions, e.g., assessing spontaneous cognition in neurodegenerative disorders (e.g., Alzheimer's disease), thought disturbances in psychosis (e.g., derailment, flight of ideas, perseveration, etc.), or intrusive and repetitive thoughts in anxiety, obsessive-compulsive, or posttraumatic stress disorders. Furthermore, the FAST fully embraces ideas from emerging trends in neuroimaging studies, such as naturalistic and personalized approaches to high-dimensional neuroimaging data that emphasize multivariate representations and network-level dynamics. Together, our study provides a new research tool that will be useful for both behavioral and neuroimaging studies, creating a new possibility of capturing psychological and neurobiological processes previously challenging to study.

Second, our task's primary brain targets were regions related to autobiographical memory, self-referential and emotion processing, and the monitoring and modulation of visceral and autonomic activity, including the MPFC and the medial temporal lobe structures within the default mode and limbic systems. While participants were reflecting on the self-generated concepts, these brain regions (referred to as "meaning-related" in fig. S8) showed a delayed but strong activation after the transient stimulus-driven activity within the early visual and attentional orienting networks. These distinct temporal patterns of brain activity suggest that participants paid attention to the stimulus at the beginning of the trial but then turned their attention inward and initiated endogenous cognitive and affective processes to reflect on the stimulus' personal semantic meaning from a first-person perspective. These findings have significant basic science implications for the dynamic interplay between perceptually coupled and internally guided (i.e., "imaginative") thought—processes that are often assumed to be antagonistic but which must work together when attaching personal meaning to external stimuli, as we show here (*21*, *42*–*44*).

Third, predictive modeling of content dimension ratings revealed that the brain representations of emotional valence became more idiosyncratic (i.e., person specific) as the level of self-relevance increased. When we trained the population-level predictive models, which capitalized on multivariate fMRI pattern information conserved across individuals, we could only predict the self-relevance dimension ratings. The important predictors of the self-relevance model included multiple regions within the default mode and limbic system, including the MPFC, PCC, TPJ, TP, NAc, and hippocampus. Most of these brain regions have been implicated in self-referential, mentalizing, autobiographical memory, visceral monitoring, and autonomic regulation in the previous literature (*16*, *36*–*39*, *45*). For the valence dimension, we failed to develop a well-performing prediction model of valence when all data were combined for the modeling, but when we used only the trials with low self-relevance, we were able to develop a population-level prediction model of valence. This valence–low-self model also showed significant generalization to two independent datasets that used exogenous emotional stimuli (i.e., IAPS pictures), suggesting the existence of the generalizable valence codes in the brain when the emotional stimuli were less self-relevant. However, individuals exhibited idiosyncratic brain representations of valence when stimuli were highly self-relevant.

The findings of the idiosyncratic valence representations induced by self-relevant stimuli have important implications for emotion research. First, the results highlight the importance of the choice of stimuli and tasks for the study of emotion. Suppose that we only use exogenous stimuli to induce emotions, such as movies, music, or pictures generated or selected by researchers. In that case, we might not be able to fully capture the brain representations and mechanisms of endogenous affective experiences. For this reason, we call for future research to develop and use experimental paradigms that focus more on self-generated and naturalistic stimuli to target endogenous and first-person experiences of emotions. Second, our results suggest that the valence codes in the brain can be modulated by the stimulus types and contexts, supporting the "affective modes" hypothesis (a brain subsystem can have different valence codes depending on affective states and contexts) rather than the "affective module" hypothesis (a brain subsystem is dedicated to a single valence code) (*46*). Our study showed that the self-relevance level could serve as a crucial affective context that can produce significant changes in affective modes in the brain. Third, although the affective mode change appeared to occur in multiple brain regions distributed across the whole brain, regions within the default mode network and the limbic system, such as the TPJ, hippocampus/amygdala, VMPFC, and TP, seem to play a central role in this mode change. This is likely due to their involvement in episodic and semantic memories and emotion processes (*17*, *24*, *47*). Episodic memory–related brain regions may provide rich autobiographical and personal contextual information to the semantic representations of valence, producing idiosyncratic representations of valence—i.e., turning a simple representation of good and bad (i.e., valence) into more complicated, nuanced, and personally unique valence representations. In addition to the interaction between memory and emotion processes, visceral monitoring and autonomic modulation in these DMN and limbic regions may also play an essential role in modulating valence representations in the brain (*21*, *45*). These DMN and limbic areas have been proposed to provide a subject-centered reference frame by integrating various visceral inputs (*48*) and thereby serve as a basis for the subjective experience of "self-relatedness" (*45*) [or "mineness" (*49*)]. The importance of these regions in processing self-generated concepts was further supported by the supplementary analyses shown in figs. S11 to S14 and table S6, but it still needs to be further examined in future studies [e.g., using the semantic encoding model (*50*) or state-space modeling (*51*)].

There are some considerations and limitations in the current study. First, some participants reported that the task was challenging to perform. One factor that can influence task difficulty is verbal fluency. Although the verbal fluency scores did not show significant correlations with the model prediction or input features (see table S7), it may be because our participants were mostly college students with similar levels of verbal fluency. Therefore, to generalize our findings to a general population, future studies should further investigate the influences of experimental parameters and verbal fluency on task performance and difficulty. Second, although we obtained the content dimension ratings on a continuous scale, our Markov chain analysis was based on the dynamics on the discrete state space, which could result in information loss. Although simple analysis methods, such as discrete state-space Markov chain analysis, could allow us to achieve better generalizability (based on the bias-variance tradeoff) and better interpretability (*52*), future studies should test other analysis methods that can use the dynamic information about the continuous state space. Third, our Markov chain–based predictive model included a positive predictive weight for the time-mean variable, and this could seem inconsistent with the previous literature (*53*–*55*), which suggested a negative mood is associated with more frequent past-oriented spontaneous thought. Through further investigation (fig. S15), however, we found that the positive weight for the time-mean variable implied that the negative affectivity was associated with more recent-past oriented thoughts (rather than distant-past oriented thoughts). Therefore, our results were not inconsistent with previous studies. Instead, our results suggest that a more careful treatment for the time dimension is required in future studies on spontaneous thought (e.g., using a continuous time scale). Fourth, through supplementary analyses (fig. S16), we found several differences between the high versus low self-relevance trial bins, including the distribution of valence and vividness scores, semantic distances, and the consistency of the reported concepts across individuals. On the basis of the data of the current study, we cannot determine which are more important factors that cause other observed differences, and therefore, future studies should aim to characterize self-relevant spontaneous thought in more depth. In particular, it would be important to modulate the level of self-relevance while keeping other factors the same to better understand the effects of self-relevance. Thus, more controlled experimentation could be helpful. Fifth, although we analyzed the fMRI data from the concept reflection period to minimize the issues related to motion confounds, the brain dynamics and activity patterns during the actual free association (i.e., the concept generation period) could be different from the concept reflection period. To address this issue, we will need creative methods to mitigate motion confounds during free association (but see fig. S5). Sixth, future studies should also test features and dimensions other than the five content dimensions to examine whether our findings on the effects of self-relevance on the valence prediction performance are specific to valence or generalize to other features of self-relevant concepts (e.g., visual or semantic features).

Overall, the current study opens up a new possibility of a quantitative assessment of the spontaneous thought dynamics and their brain representations. Our behavioral task and predictive models have the potential to be clinically useful because they can provide rich information about an individual's cognitive and conceptual dynamic signatures. Furthermore, our findings suggest that neural representations of affective processes become more idiosyncratic when self-relevant stimuli are used to induce emotions, highlighting the importance of targeting endogenous cognitive and affective processes in the study of emotion. Together, our study leads us to a deeper understanding of how self-relevance modulates the affective representations in the brain—that is, what happens in our brain when the self comes to a wandering mind.

MATERIALS AND METHODS

Participants

For the FAST-fMRI study, 63 healthy, right-handed participants participated [age = 23.0 ± 2.5 years (means ± SD), 30 females]. Because the current study was the first fMRI study that used the FAST, we could not determine the sample size based on statistical power calculation. However, this sample size was larger than the top 90% sample sizes among the experimental fMRI studies published between 2017 and 2018 (*56*). The preliminary eligibility of participants was determined through an online screening questionnaire. We did not include participants with psychiatric, neurological, systemic disorders, or MRI contraindications. After the experiment, we excluded behavioral (and fMRI) data from one participant who generated too few responses. Thus, we used data from 62 participants in the behavioral data analysis. We also excluded one participant's fMRI data due to insufficient MRI coverage. Thus, we used data from 61 participants in the fMRI analysis. For the retest session, 30 participants (age = 22.8 ± 2.3 years, 15 females) revisited the experiment about 7 weeks (mean = 51.0 ± 16.8 days) after their first visit. For the web-based FAST study that was used for an independent test data of the Markov chain analysis, 117 participants (age = 22.6 ± 2.6 years, 56 females) completed the web-based behavioral FAST experiment in the behavioral experiment room. Among them, 49 participants (age = 23.0 ± 3.1 years, 24 females) revisited and conducted the second session. Their revisits were about 7 weeks (mean = 54.2 ± 8.5 days) after their first visits.

We recruited participants from the Suwon area in South Korea, and the experiments were conducted at the Center for Neuroscience Imaging Research, Sungkyunkwan University in Suwon, South Korea. The institutional review board of Sungkyunkwan University approved these studies. All participants provided written informed consent and were paid for their participation.

Self-report questionnaires

All participants completed a battery of self-report questionnaires to assess individual differences in mental health and affective traits and states. In the fMRI study, we included the 20-item Positive and Negative Affect Schedule (PANAS) (*57*) (which had two subscales—positive affect and negative affect), the 20-item Center for Epidemiologic Studies Depression (CES-D) (*58*), the 22-item Rumination Response Scale (RRS) (*59*) (which consisted of two subscales—brooding and depressive rumination), the 20-item trait version of the State-Trait Anxiety Inventory (STAI-T) (*60*), and the 30-item Mood and Anxiety Symptom Questionnaire-D30 (*61*) (which had three subscales—general distress, anhedonic depression, and anxiety arousal) in the battery. Participants completed their responses to the questionnaires before the fMRI scan.

In the web study, we included the 20-item PANAS, the 20-item CES-D, a subset of the RRS (nine items from the "depressive rumination" subscale), and the 20-item STAI-T. These questionnaires overlapped with the fMRI study. In addition to these, we also included the Suicidal Ideation Questionnaire (SIQ) (*62*) (we only used two items, "I wished I were dead" and "I thought that life was not worth living"), the 3-item Loneliness Scale (LS) (*63*), the 5-item Satisfaction with Life Scale (*64*), and the 18-item Psychological Well-Being Scale (*65*) (which had six subscales—autonomy, environmental mastery, personal growth, positive relations with others, purpose in life, and self-acceptance) in the battery. We used the Korean versions of these questionnaires that showed similar psychometric properties to the original questionnaires (*66*, *67*). Descriptive statistics of the self-report questionnaires are reported in table S3.

FAST for the fMRI experiment

The FAST for the fMRI experiment comprised three parts—(i) concept generation, (ii) concept reflection, and (iii) postscan survey. For the concept generation phase, we asked participants to report a word or phrase that came to mind in response to the previous concept every 2.5 s starting from a given seed word inside the MRI scanner. The responses were collected through an MR-compatible microphone. Participants were asked to generate a total of 40 consecutive concepts for each seed word, and we used four seed words for four runs total. We made the number of associations for each seed word much longer than our previous study (*12*), in which we collected only 10 consecutive concepts to obtain a larger number of personal concepts. The four seed words were "family," "tear," "mirror," and "abuse" for the first session and "love," "fantasy," "heart," and "pain" for the retest session. The seed words were selected on the basis of a presurvey of valence and self-relevance with a large set of candidate words. We selected these seed words to make them evenly distributed on the valence and self-relevance dimensions. The orders of the seed words were fully randomized across participants.

During the concept reflection phase, we showed participants two consecutive concepts they generated in sequence. We made the second concept bigger than the first one to make it clear for the second concept to be the target word. Then, we asked participants to think about the target concept and personal context relevant to the association between the two concepts for 15 s. The stimuli remained on the screen for the whole duration (i.e., 15 s). We tried to give them enough time to think about their personal context, such as memory, that gave the concept a personal meaning. Between the trials, we showed a fixation cross with a jittered duration between 3 and 9 s (i.e., interstimulus interval "baseline"). Intermittently, we showed 14 emotion words on the screen after the word presentation and asked participants to select one emotion descriptor closest to their current feeling. The emotion words were joy, distress, hope, fear, satisfaction, disappointment, pride, embarrassment, remorse, gratitude, anger, love, hate, and neutral. We conducted this emotion rating five times per run and collected 160 self-generated words per participant. We used these emotion rating data to validate the postscan survey results (fig. S2).

After the fMRI scan, participants completed a postscan survey on the 160 self-generated concepts in the behavioral experiment room. We showed the self-generated concepts again and asked participants to rate them on multiple content dimensions (*13*). The content dimensions evaluated emotional valence (how much positive or negative feelings does the concept evoke?), self-relevance (how much is the concept relevant to yourself?), time (which time point is most relevant to the concept, ranging from the past to the future?), vividness (how much vivid imagery does the concept induce?), and safety-threat (does the concept give rise to the feeling of safety or threat?). Similar to the concept reflection task, we showed two consecutive concepts in sequence with five content dimension questions and asked participants to answer the questions about the target concept (i.e., the second one). We used the visual analog scale. The valence, time, and safety-threat ratings were coded between −1 (negative, past, and threat, respectively) and 1 (positive, future, and safety), and 0 indicated neutral or present. The self-relevance and vividness ratings were coded between 0 (not self-relevant, not vivid) and 1 (highly self-related, highly vivid).

FAST for the web experiment

The web-based FAST consisted of the concept generation and concept survey tasks. The concept generation task was similar to that of the fMRI experiment, but this time, we provided 10 s for the concept generation because typing usually took longer than speaking. The seed words for the first and second sessions were the same as those of the fMRI experiment. After the concept generation task, participants completed the concept survey on three content dimensions—valence, self-relevance, and time. We chose these three dimensions based on the PCA results shown in fig. S3A. The PCA results suggested that the valence and safety-threat dimensions were highly correlated, and the self-relevance and vividness were also highly correlated.

Markov chain–based predictive modeling of negative affectivity

To build predictive models of general negative affectivity, we used features obtained from a Markov chain analysis on the content dimensions. First, we divided the dimension scores into multiple discrete states. Our decision on how to define the discrete states of each dimension was determined a priori. We defined two discrete states for the dimensions that use a unipolar scale (i.e., self-relevance and vividness) and three discrete states for the dimensions that use a bipolar scale (i.e., valence, time, and safety-threat). In these bipolar dimensions, we assumed the middle range (i.e., near zero) to be psychologically meaningful and interpretable—for example, in the case of valence and safety-threat dimensions, we can interpret the middle range as "neutral," and for the time dimension, we can interpret the middle range as "present." In more detail, for the valence, time, and safety-threat dimensions, which ranged from −1 to 1, we divided them into three discrete states using −0.33 and 0.33 as the boundaries for defining discrete states (i.e., −1 to −0.33, −0.33 to 0.33, and 0.33 to 1; for valence, the three discrete states were negative/neutral/positive; for time, past/present/future; and for safety-threat, threatening/neutral/safe). The self-relevance and vividness dimensions, ranging from 0 to 1, were divided into two discrete states (0 to 0.5 and 0.5 to 1, which corresponded to low and high for both dimensions, respectively).

We then calculated the state transition and steady-state probabilities for each dimension and each participant. The state transition probability refers to the probability of making transitions from one to another discrete state on each dimension. The steady-state probability refers to the probability of converging to one state when the transitions are sufficiently repeated. We obtained the steady-state probability by multiplying the transition probability matrix by itself 10,000 times, which was always converged to one probability vector. In addition to these dynamic features from the Markov chain analysis, we also used each content dimension's mean and variance as input features. Together, we used nine (=3 × 3) transition probability values and three steady-state probabilities for the valence, time, and safety-threat dimensions, four (=2 × 2) transition probabilities and two steady-state probabilities for the self-relevance and vividness dimension, and mean and variance of all dimensions, creating a total of 58 input features (i.e., predictor variables). Most of these features showed a good level of consistency across different sets of seed words and different time points (table S1).

For the outcome variable, we conducted factor analyses to calculate general negative affectivity scores from a combination of self-report questionnaires. As an input for the factor analyses, we used *z*-scored subscale scores because all questionnaires used different scales. We used a target oblique rotation for a two-factor model, given that we had a priori knowledge about what each questionnaire measures between positive versus negative affectivity. The general negative affectivity factor consisted of questionnaires related to negative affect, general distress, anxiety, and depression (table S2), which was then used as the outcome variable for the predictive modeling.

With these predictor and outcome variables, we developed a predictive model of the negative affectivity using LASSO regression. We used the first session data of the fMRI study (n = 62) as a training dataset and tested the developed model on the second session data of the fMRI study (n = 30) and two session (test and retest) datasets of the web study (n = 117 and 49, respectively). We determined the number of the predictor variables for the final model based on the LASO-CV performance in the training set with the LASSO regularization.

To test the prediction model on independent test datasets, we calculated the predicted levels of general negative affectivity using a dot product between the Markov chain features calculated from new datasets and the model weights. To evaluate the model performance, we used a robust correlation between the actual and predicted levels of general negative affectivity. We did not use mean squared error or R-squared to evaluate model performance as recommended in (*31*), given that the scale of outcome variables was different between the training and test datasets because they used different sets of self-report questionnaires.

fMRI data acquisition and preprocessing

We collected MRI data using a 3T Siemens Prisma scanner at Sungkyunkwan University. We acquired high-resolution T1-weighted structural images and functional echo-planar imaging (EPI) with repetition time (TR) = 460 ms, echo time (TE) = 27.2 ms, multiband acceleration factor = 8, field of view = 220 mm, 82 by 82 matrix, 2.7-mm by 2.7-mm by 2.7-mm voxels, 56 interleaved slices, and number of volumes = 2608. Stimulus presentation and behavioral data acquisition were controlled using MATLAB (MathWorks, Natick, MA) and Psychtoolbox (http://psychtoolbox.org/).

Data preprocessing was performed with SPM12 (Wellcome Trust Centre for Neuroimaging) and FSL (the Oxford Centre for Functional MRI of the Brain). For structural T1-weighted images, we coregistered the T1 images to the functional image using the first single-band reference (SBRef) image and segmented and normalized them to the MNI space. For EPI images, we removed the initial volumes (20 images) of fMRI data to allow for image intensity stabilization. We also identified outliers for each image and all slices based on Mahalanobis distances and the root mean square of successive differences to remove intermittent gradient and severe motion-related artifacts that are present to some degree in all fMRI data. For the Mahalanobis distance–based outlier detection, we computed Mahalanobis distances for the matrix of concatenated slice-wise mean and SD values by volumes across time. Then, we identified the images that exceed 10 mean absolute deviations based on moving averages with full width at half maximum (FWHM) of 20 images kernel as outliers. With the root mean square of successive differences across volumes, images that exceeded three SDs from the global mean were identified as outliers. Each time point identified as outliers by either outlier detection method was included as nuisance covariates.

We then conducted (i) motion correction (realignment) using the SBRef image as a reference, (ii) distortion correction using FSL's topup function, (iii) normalization to the MNI space using the parameters from the T1 normalization with the interpolation to 2-mm by 2-mm by 2-mm voxels, and (iv) smoothing with a 5-mm FWHM Gaussian kernel. Because data from two participants showed poorer quality images after the distortion correction, we used distortion-uncorrected images for these two participants.

fMRI single-trial analysis

We used the single-trial analysis approach. We estimated single-trial response magnitudes for each brain voxel using a general linear model design matrix with separate regressors for each trial, as in the "beta series" approach (*68*). We constructed each trial regressor for the concept reflection duration with a boxcar convolved with SPM12's canonical hemodynamic response function. We also included a regressor for intermittent emotion ratings for each run. We concatenated four runs' data during the concept reflection task for each participant and added the run intercepts. In the design matrix, we also included the nuisance covariates, including (i) "dummy" coding regressors for each run (an intercept for each run), (ii) linear drift across time within each run, (iii) 24 head motion parameters (six movement parameters including *x*, *y*, *z*, roll, pitch, and yaw, their mean-centered squares, their derivatives, and squared derivative) for each run, (iv) indicator vectors for outlier time points, and (v) five principal components of white matter and cerebrospinal fluid signal. With this design matrix, we ran the first-level analysis using SPM12 with a high-pass filter of 180 s. Because single-trial estimates could be strongly affected by acquisition artifacts that occur during that trial (for example, sudden motion, scanner pulse artifacts, etc.), we calculated trial-by-trial variance inflation factors (VIFs; a measure of design-induced uncertainty due to collinearity with nuisance regressors) using the design matrix, and any trials with VIFs that exceeded three were excluded from the following analyses. The average number of trials excluded because of high VIFs was 2.902, with the SD of 3.081.

Large-scale functional network overlap analysis

The radial plots in Figs. 3, 4, and 6 show the relative proportions of the number of overlapping voxels between the thresholded map and each network (or region), given the total number of voxels within each network (or region). We used the Buckner group's parcellations to define large-scale functional brain networks, including seven 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 and the brainstem region, as shown in fig. S6.

Clustering analysis using the FIR model

For the clustering analysis based on the temporal patterns of brain activity, we estimated the TR-level brain activity patterns using the FIR model (Fig. 3, B and C). We modeled 35 TRs (=16.1 s) from the concept reflection trial onset. We used the thresholded univariate contrast map (at $P < 0.05$, Bonferroni correction) for the concept reflection duration as a mask. We then conducted *k*-means clustering on the voxels within the mask based on the temporal patterns of brain activity across 35 TRs. We selected the number of clusters *k* that maximized the silhouette value of the clustering solution using "evalclusters.m" function included in the MATLAB statistical toolbox. To show the time course of the cluster brain activity, as in Fig. 3B, we plotted the averaged beta weights within each cluster at each time point.

Predictive modeling of content dimensions

To develop multivariate predictive models of each content dimension, we used the whole-brain beta images from the single-trial analysis after excluding trials with VIFs that exceeded three and self-report ratings of content dimensions. We first divided the trials into quartiles based on the content dimension scores and averaged the content dimension ratings and fMRI data for each level, creating the quartile training data for each participant. In these data, the outcome variable was still continuous because we used the quartile averages. For each dimension, a total of 244 images [4 (images per participant) × 61 (number of participants)] were created for model training. Some dimensions had less than 244 images due to the skewed distribution of some participants' ratings or the removal of some trials with high VIFs. We used PCR to train multivariate pattern-based predictive models. To obtain unbiased estimates of model performance, we used two different types of cross-validation methods. The first was the LOSO-CV, in which we derived a predictive map from all participants' data except one participant and used the hold-out participant's data for the model testing. The other method was the RS-CV (*32*), in which we randomly chose 20% of participants' data as the hold-out data for each iteration. We then used 80% of randomly selected participants' data to derive a predictive map and the 20% hold-out data for the model testing for each iteration. We repeated this procedure 50 times. We evaluated model performance with (i) averaged within-participant correlation between actual and predicted ratings and (ii) averaged within-participant mean squared error. To test whether the mean within-participant prediction-outcome correlation was significantly larger than zero, we conducted bootstrap tests with participant-level prediction-outcome correlation values with 10,000 iterations. For the predictive model of valence only for low self-relevance trials (i.e., the valence–low-self model; Fig. 4B), we divided the data into trials with high versus low self-relevance scores using 0.5 as a threshold before making the quartile data. After that, all the model training and testing steps were the same as above.

To help the feature-level interpretation of the predictive models, we conducted bootstrap tests with 10,000 iterations. We randomly sampled participants with replacement for each iteration and trained a PCR model using the resampled dataset. On the basis of the sampling distribution of bootstrapped predictive weights, we identified features that consistently contributed to the prediction with P values. For display, as in Fig. 4, we thresholded the map with the FDR $q < 0.05$ and pruned the results using two additional more liberal thresholds, uncorrected voxel-wise $P < 0.01$ and $P < 0.05$, two-tailed.

To test our valence–low-self model on an independent study dataset, as in Fig. 5, we obtained predicted ratings using a dot product of each vectorized test image data of the previous study with model weights. We used averaged within-participant correlation between actual and predicted ratings to evaluate the prediction performance. We did not use the mean squared error because the units differed between the training and testing datasets. Similarly, to test the PINES on our data, we calculated pattern expression values with a dot product between the PINES weights and the test data of high and low self-relevant trials separately. As the modeling approach we mentioned above, we first separated one individual's trials into high or low self-relevant data using 0.5 as a threshold first and divided them into quartiles again based on the valence ratings. Then, we applied averaged beta images and averaged valence scores within each quartile to the PINES. We used averaged within-participant prediction-outcome correlation as the evaluation measure.

Idiographic predictive modeling of valence

To quantify the between-subject variability of predictive weights for data across the whole brain with high versus low self-relevance, we used the idiographic predictive modeling approach. Similar to the group-level predictive modeling, we used the beta images obtained from a single-trial analysis and content dimension ratings. First, we divided trials into high and low self-relevance groups using 0.5 as a cutoff score. For each participant, we trained two predictive models of valence with these two groups of trials separately using PCR with 5-fold cross-validation.

We also conducted a searchlight-based pattern similarity analysis to identify brain areas that showed similar or distinct patterns of predictive weights between two types of valence prediction models. For this, we created a searchlight with a radius of five voxels and scanned it throughout the whole brain with a step size of four voxels. We then calculated correlation coefficients between two models' predictive weights within each searchlight and transformed r to z using Fisher's transformation. We added the z values to cubes that had one side of four voxels and were located at the center of each searchlight across the whole brain with no overlapping voxels between cubes. We smoothed the map with a 3-mm FWHM Gaussian kernel and performed a one-sample t test, treating participants as a random effect. We also calculated the JZS Bayes factors using the method proposed in (*72*).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abn8616

View/request a protocol for this paper from *Bio-protocol*.

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 Univ. 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).

Acknowledgments: We thank M. Jung, J. Park, and S. Song for help in conducting the experiments. **Funding:** This work was supported by IBS-R015-D1 (Institute for Basic Science; to C.-W.W.); 2019R1C1C1004512, 2021M3A9E4080780, and 2021M3E5D2A01022515 (National Research Foundation of Korea; to C.-W.W.); and 2E30410-20-085 (KIST Institutional Program; to C.-W.W.). **Author contributions:** Conceptualization: B.K.L., J.R.A.-H., and C.-W.W. Methodology: B.K.L. and C.-W.W. Investigation: B.K.L., J.H., E.L., J.R.A.-H., and C.-W.W. Visualization: B.K.L. and C.-W.W. Supervision: C.-W.W. Writing—original draft: B.K.L. and C.-W.W. Writing—review and editing: B.K.L., J.R.A.-H. and C.-W.W. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The predictive models, data, and codes to generate main figures are available at https://doi.org/10.5281/zenodo.6825023.

Submitted 29 December 2021
Accepted 14 July 2022
Published 31 August 2022
10.1126/sciadv.abn8616