nature communications

Article

https://doi.org/10.1038/s41467-025-65541-1

Neural decoding of autobiographical mental image features with a general semantic model

Received: 25 January 2025

Accepted: 17 October 2025

Published online: 12 December 2025

Check for updates

Andrew J. Anderson[1,2,3], Leonardo Fernandino[1,2], William L. Gross[1,2,4,5], Hernan G. Rey[2,3] & Jeffrey R. Binder[1,5]

To what extent does language comprehension engage the same brain areas that enable one to imagine autobiographical experiences? Traditionally, these two abilities have been thought to recruit largely separate neurocognitive systems, namely the semantic system – involving lateral aspects of the temporal, parietal, and frontal lobes – and the autobiographic/episodic system, primarily involving midline cortical areas and the hippocampus. Here, we provide evidence of shared representations elicited in sentence comprehension and self-generated autobiographical mental imagery that are most prominent in midline regions and present in both sets of cortical areas. Fifty participants who imagined their unique experiences of twenty common scenarios while undergoing functional Magnetic Resonance Imaging (fMRI) are analyzed. Self-reported, participant-specific experiential feature ratings of the mental images are reconstructed from fMRI activation in both midline and lateral surface cortical areas using a pre-trained semantic decoder. Critically, the semantic decoder was derived from fMRI activation recorded during sentence reading in a separate participant group. This finding demonstrates zero-shot decoding of participant-specific autobiographical feature ratings from fMRI data, across people and cognitive tasks. Moreover, it strongly suggests that the neural encoding of sentence meaning shares cortical areas and at least some representational codes with self-generated autobiographical mental images.

People often imagine autobiographical experiences—that is, situations internally simulated from ones' own perspective, such as what one saw, heard or felt, what one was doing, who was there, and so on. Through language comprehension, people can also understand situations they have never experienced themselves from verbal input. How are the two phenomena related in the brain? While both cognitive processes are anchored in human experience, the extent to which they engage common brain systems is unclear. Although there are clear differences in the subjective experience of autobiographical mental imagery and language comprehension[1,2], with the former linked to vivid multisensory simulations and the latter to more general conceptual knowledge, some representational overlap in the brain would appear to be essential to enable one to understand language in the context of one's own memories[3]. The current study seeks to expose and characterize such commonalities by computationally modeling autobiographical mental images and sentence comprehension in the same representational space. In so doing, it contributes to advancing the scientific understanding of human memory and imagination, which

[1]Department of Neurology, Medical College of Wisconsin, Milwaukee, WI, USA. [2]Department of Biomedical Engineering, Medical College of Wisconsin, Milwaukee, WI, USA. [3]Department of Neurosurgery, Medical College of Wisconsin, Milwaukee, WI, USA. [4]Department of Anesthesiology, Medical College of Wisconsin, Milwaukee, WI, USA. [5]Department of Biophysics, Medical College of Wisconsin, Milwaukee, WI, USA. e-mail: andanderson@mcw.edu

are principal goals of cognitive neuroscience. Also, by providing proof-of-principle that features associated with self-generated mental images and language comprehension can be recovered from brain activity with the same computational model, the work is relevant to initiatives to build AI language model-based brain decoders[4,5].

The study examines the roles of three different subsystems of the brain's default mode network[6] (DMN), which have been proposed to differentially contribute to the imagination of autobiographical experiences and to language/conceptual knowledge representation[7,8]. Specifically, the medial temporal (MT-DMN) and Core (Core-DMN) subsystems, together, have been strongly associated with autobiographical/episodic recollection and simulation[9–11]. MT-DMN has been informally characterized as the "mind's eye" and may enable one to mentally visualize scenes[8], whereas Core-DMN has been proposed to support self-reflection and to integrate perception and memory[7]. Differently, the fronto-temporal subsystem (FT-DMN) contains brain regions that have been associated with sentence comprehension[12–14] and conceptual knowledge representation[15], and FT-DMN has been informally characterized as "the mind's mind", encoding abstract verbal components of imaginative thought[8]. Alongside these specialisms, all three DMN subsystems have been implicated in encoding semantic features in language comprehension[16–23]. However, because studies have generally treated the imagination of autobiographical experiences and comprehension separately (though see ref. 24 for evidence that DMN activation patterns correlate during video watching and spoken video recall), it is unclear whether the same DMN representational features are self-activated during the imagination of autobiographical experiences and externally activated during language comprehension. We hypothesized that this would be the case, and each DMN subsystem would represent imagined autobiographical scenarios and comprehended language in a similar representational space, to some degree at least.

Our overarching hypothesis was that (some) participant-specific features of autobiographical mental images could be reconstructed from corresponding brain activity using a semantic decoder trained on fMRI data obtained from different participants during a sentence comprehension task[20,25–27]. This would not be possible if autobiographical mental images and sentence meaning were encoded in separate brain systems or used different representational codes. To evaluate this hypothesis, we needed to not only have recordings of brain activity related to imagining autobiographical experiences, but also to know the idiosyncratic characteristics of the experiences imagined, to determine whether these characteristics could be reconstructed from the fMRI data. This was approached[28] by preselecting a set of twenty generic scenarios (e.g., a wedding) for participants to imagine in terms of their own idiosyncratic experiences (i.e., a wedding they have experienced) and having them describe, in detail, the contents of their imagined scenarios prior to the fMRI experiment. During fMRI scanning, participants again imagined their autobiographical scenarios when cued by generic written prompts that did not describe any personal details (e.g., "a wedding scenario"). However, whilst this approach provided participant-specific information with which to evaluate feature reconstructions, the resulting fMRI data encoded not only self-generated mental images, but also semantic representations elicited by reading the stimulus prompt. To establish that the sentence decoder was not simply reconstructing the meaning of the prompt, we employed an individual differences analysis to test whether participant-specific mental image content could be reconstructed from fMRI data (Fig. 1).

Results

Experiment overview

We analyzed fMRI scans of brain activation taken from 50 healthy people[28,29]. The participants were from a diverse age range: 25 were young adults (mean ± SD age = 24 ± 3, 16 F, 9 M) and 25 were elderly adults (mean ± SD age = 73 ± 7 years, 16 F, 9 M). While age-related differences were not a focus of the study, we analyzed the data from each age group separately, in addition to the two groups combined, to determine whether findings generalized across ages.

The experiment[28] hinged on participants imagining their autobiographical experience of twenty common, experimenter-selected events (e.g., Wedding, Funeral, Festival, Exercising, Driving). To ensure there was a record of the participant-specific details of the mental images, experimentation began outside the scanner (Fig. 1A). Scenario prompts were read to participants, who were requested to imagine the scenario including their sensations, actions and feelings (e.g., on being told "Wedding" the participant might imagine a scene from their own wedding, including what they were doing, how they felt and who was there). To provide a participant-specific model of each mental image that would later be reconstructed from fMRI data, participants rated the importance of twenty sensory, motor, affective, social, cognitive and spatiotemporal features of experience to their mental image on a scale of [0 6] (see also ref. 30). Thus, each mental image scenario was modeled as a vector of 20 idiosyncratic numeric feature ratings (see the blue/yellow matrix in Fig. 1A). Supplementary Figs. 1 and 2 display mean and SD ratings across participants. The specific rating instructions are included in Supplementary Table 1.

To further characterize the nature of the imagined scenarios, and the extent to which they reflected episodic memories, specific to a single space and time[1,2], participants rated each scenario on whether it reflected a real-life event and whether it was vividly imagined. Ratings indicated that imagined scenarios largely reflected real events (overall mean across scenarios and participants = 5.5 and 5.6 for elderly and young, respectively [0 6]) that were vividly imagined (overall mean = 5.2 and 4.6 for elderly and young, [0 6]).

Additionally, participants verbally described the scenarios they had imagined (see word clouds in Supplementary Fig. 3). The verbal descriptions were of secondary importance to the analyses and were used to provide comparative decoding outcomes using a large language model (GPT-2[31]).

Participants then underwent fMRI as they reimagined the same set of autobiographical experiences, one at a time, when cued by written prompts (e.g., "a wedding scenario"). The entire set of scenario prompts was repeated five times, in different randomized orders to enable fMRI data for each scenario to be averaged across the presentations to increase signal-to-noise ratio. fMRI data were preprocessed using standard methods (including fMRIPrep[32]), which generated a single fMRI volume for each of the twenty mental images for each participant.

Because the subsequent analyses were based on cross-participant neural decoding/encoding regression mappings, and because individual voxels tend not to have one-to-one functional correspondences across participants, even when normalized to a common neuroanatomical space, we re-represented all individuals' fMRI data in a coarse-grained standard space. For this, we deployed the Schaefer-1000 cortical parcellation[33], which divides the cortex into 1000 parcels (aka regions-of-interest or ROIs) containing contiguous voxels that co-activate during resting-state (when people are prone to think about autobiographical experiences[34]). Critically, the Schaefer parcellation space also predefines Core-DMN (DefaultA), FT-DMN (DefaultB) and MT-DMN (DefaultC) as subsets of 82, 66, and 27 ROIs, respectively, which enabled analyses to be conducted within each DMN subsystem. To reduce the current fMRI data to the common Schaefer space, voxel activation values within each ROI were averaged, thus preserving coarse cortical activation patterns that we anticipated would replicate across participants and blurring fine-grained differences that may arise from individual differences in brain anatomy.

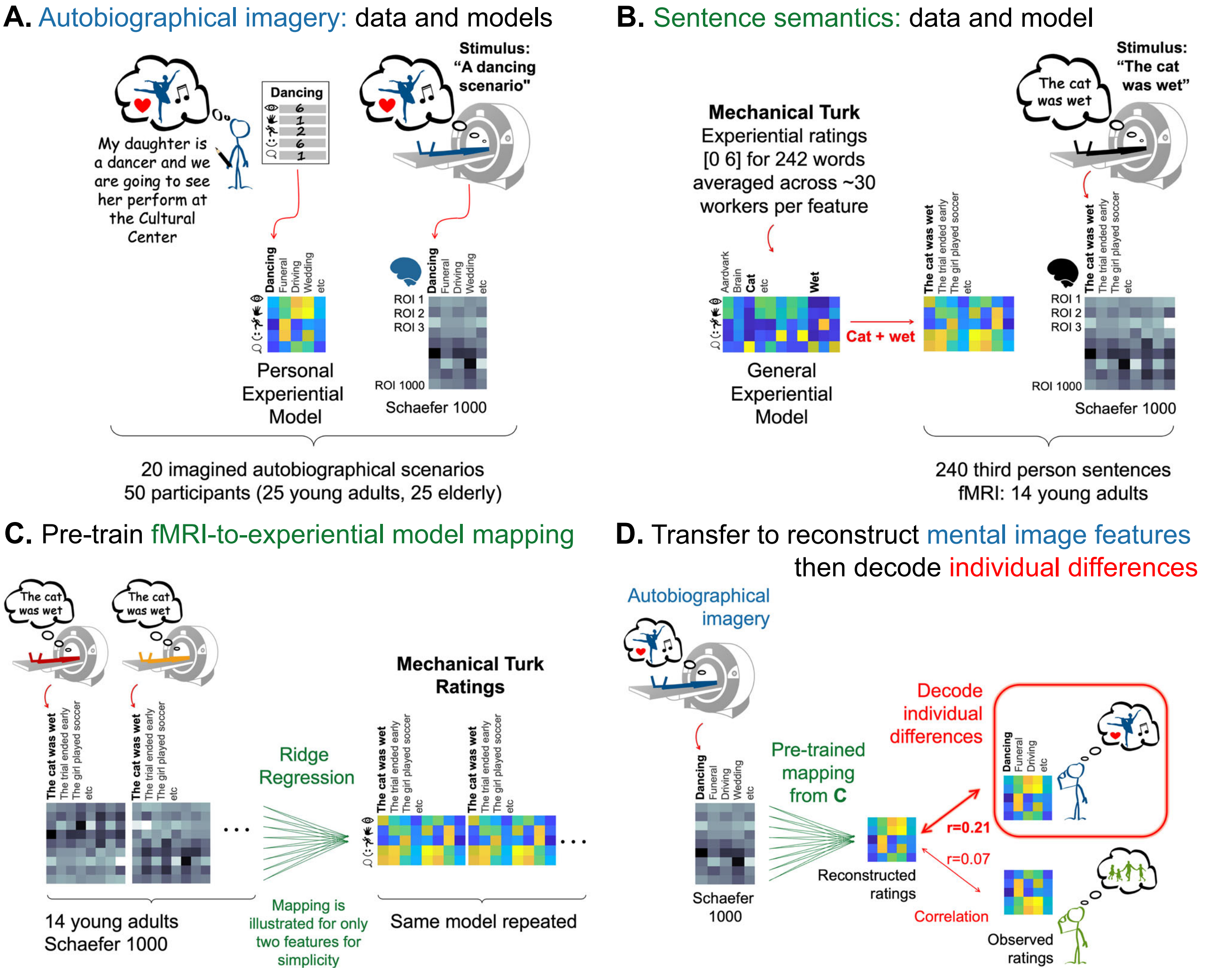
A. Autobiographical imagery: data and models
Stimulus: "A dancing scenario"
Dancing
6
1
2
6
1
My daughter is a dancer and we are going to see her perform at the Cultural Center
Dancing
Funeral
Driving
Wedding
etc
Personal Experiential Model
ROI 1
ROI 2
ROI 3
ROI 1000
Schaefer 1000
20 imagined autobiographical scenarios
50 participants (25 young adults, 25 elderly)
B. Sentence semantics: data and model
Stimulus: "The cat was wet"
The cat was wet
Mechanical Turk
Experiential ratings [0 6] for 242 words averaged across ~30 workers per feature
Aardvark
Brain
Cat
etc
Wet
Cat + wet
The cat was wet
The trial ended early
The girl played soccer
etc
General Experiential Model
Schaefer 1000
240 third person sentences
fMRI: 14 young adults
C. Pre-train fMRI-to-experiential model mapping
The cat was wet
Ridge Regression
Mechanical Turk Ratings
14 young adults
Schaefer 1000
Mapping is illustrated for only two features for simplicity
Same model repeated
D. Transfer to reconstruct mental image features then decode individual differences
Autobiographical imagery
Pre-trained mapping from C
Decode individual differences
Schaefer 1000
Reconstructed ratings
r=0.21
r=0.07
Correlation
Observed ratings

Fig. 1 | Decoding autobiographical mental image features from fMRI data with a general semantic model. A Autobiographical mental imagery data: data analyzed corresponded to fifty participants, who imagined themselves in 20 loosely defined common natural scenarios (e.g., "a dancing scenario"), mentally simulating their perception, action and feelings. Participants then rated each imagined scenario in terms of twenty sensory, motor, spatial, social, cognitive and affective experiential features. Later, they underwent fMRI as they reimagined the same 20 autobiographical scenarios, one by one, when cued by written prompts. **B** Sentence semantics data: to build a general semantic decoding model, we used an fMRI dataset acquired from a different group of 14 participants as they read 240 third-party sentences. Sentence semantics were modeled from crowdsourced ratings of the individual words on the same twenty experiential features as above. Each sentence's meaning was modeled as the feature-wise sum of content word ratings. All fMRI data from **A** and **B** were represented in a common neuroanatomical space (Schaefer-1000). **C** An fMRI decoding model was pre-trained to map the sentence reading fMRI data to the crowdsourced feature ratings using ridge regression. **D** The pre-trained fMRI decoder was transferred to reconstruct feature ratings from the mental image fMRI data (see results in Fig. 2A). To then evaluate whether the reconstructed ratings reflected idiosyncratic mental image content as opposed to generic group-level semantics, an individual differences analysis was run. This decoded participant identity by matching the reconstructed feature ratings to the observed ratings provided by the same and different participants, with the expectation that the strongest match would be for the same participant (see Figs. 2B and 3B). Some icons were adapted from public domain resources on ClipArtMax (https://www.clipartmax.com/).

Mental image features were reconstructed from fMRI data with a general semantic decoding model

To first establish whether participants' experiential feature ratings could be reconstructed from the autobiographical mental imagery fMRI data, we deployed a cortex-level decoding model that was pre-trained to reconstruct crowdsourced feature ratings from sentence fMRI data. To recap, feature reconstruction would only be successful if autobiographical imagery and sentence semantics overlap in their neural encoding. Conversely, if the two phenomena are encoded in separate brain systems or using distinct nonoverlapping representational codes, then a model that decodes fMRI sentence semantics will not reconstruct information from autobiographical brain systems.

The general semantic decoding model was built from a sentence fMRI dataset consisting of 14 participants (none from the autobiographical imagery dataset) who were scanned as they read 240 sentences that described simple situations e.g., "the child broke the glass in the restaurant". Following standard preprocessing, each sentence was represented as a single fMRI volume in Schaefer-1000 parcellation space to match the imagination fMRI data.

The general semantics of each sentence were modeled from group-averaged crowdsourced ratings of the same twenty experiential features that were used here to model autobiographical mental images. Experiential ratings were, however, collected for all 242 content words forming the 240 sentences rather than the sentences themselves. Sentences were modeled by pointwise summing together the rating vectors of all content words in each sentence (Fig. 1B).

To construct the cross-participant decoding model (Fig. 1C), we concatenated all participants' fMRI data to produce a matrix with 1000

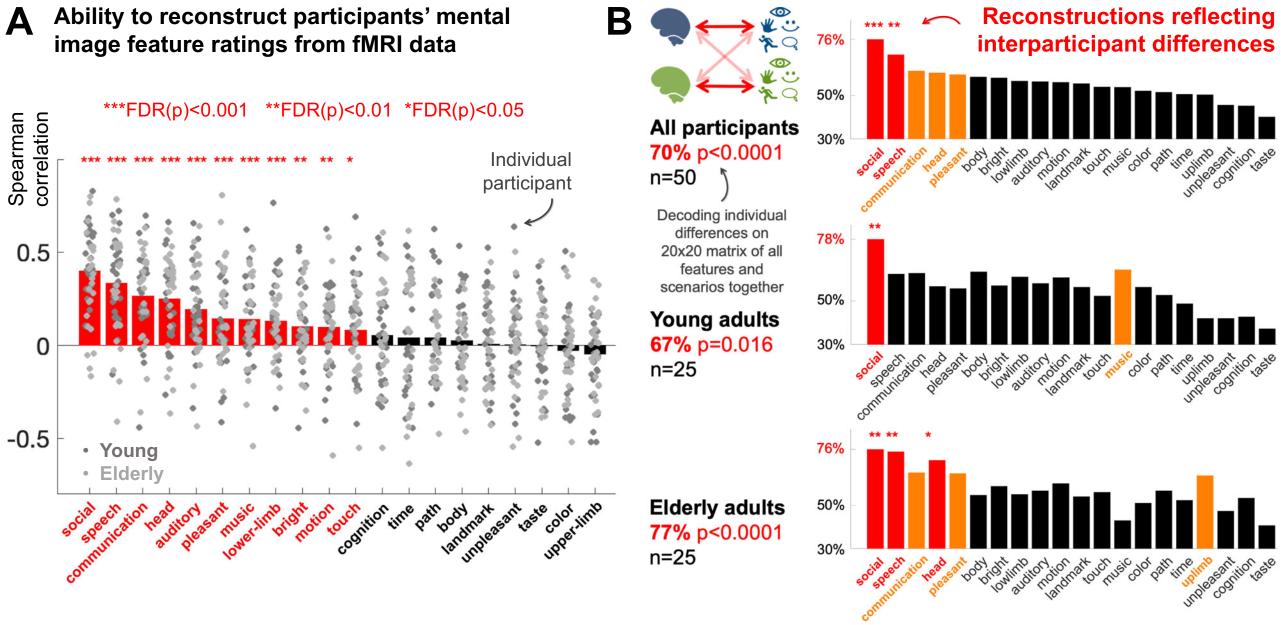

Fig. 2 | Autobiographical mental image features were reconstructed from fMRI data using a general semantic decoding model. A Reconstruction accuracy of autobiographical mental image features, decoded from fMRI activation across the entire cortex. Accuracy was computed as Spearman correlation against participants' feature ratings across the 20 autobiographical scenarios (Wedding, Exercising, etc). Bars indicate mean accuracy across participants. To estimate the statistical significance with which each feature could be reconstructed, one sample signed ranks tests were run to compare correlation coefficients to zero ($n = 50$ participants, one-tailed). *P*-values were adjusted according to the false discovery rate (FDR)[35]. 11/20 features had accuracies significantly greater than zero [FDR(p) < 0.05]. **B** Reconstructed mental image feature ratings reflect interparticipant differences: to evaluate whether reconstructed feature ratings reflected interparticipant differences, a two-alternative forced choice decoding analysis was run, testing how often participant pairs could be told apart by comparing the reconstructed feature ratings to each other's idiosyncratic feature ratings. This was evaluated on the entire vectorized matrix of 20 features for 20 scenarios, as well as on individual features (see plot bars). Red bars indicate significant decoding accuracies [FDR(p) < 0.05, permutation test, one-tailed]. Orange bars indicate decoding accuracies where permutation tests returned $p < 0.05$, but values did not survive FDR correction. The analyses were computed on all fifty participants as well as on each age group, as labeled beside the plots. Some icons were adapted from public domain resources on ClipArtMax (https://www.clipartmax.com/). Source data are provided as a Source Data file.

rows (all cortical Schaefer ROIs) and 240*14 columns (sentences*nparticipants). Feature-based representations of the sentences were likewise concatenated to align with the fMRI sentence representations (20 rows and 240*14 columns). An fMRI-to-model mapping (the general semantic decoding model) was then fit between the two matrices using ridge regression to predict, and thereby reconstruct ratings on each of the 20 individual features. The mapping, which corresponds to the regression beta coefficients, is illustrated schematically as green lines in Fig. 1C.

To evaluate whether the decoding model could reconstruct mental image feature ratings from fMRI data, we mapped the autobiographical fMRI data into experiential feature space. This was achieved with a matrix multiplication between the fMRI data matrix (20 scenarios by 1000 ROIs) and the decoding mapping matrix (1000 ROIs by 20 Features). This produced a matrix of reconstructed feature ratings for each participant (20 scenarios by 20 features by 50 participants). To evaluate how accurately the ratings of specific features were reconstructed, Spearman correlation was computed between the reconstructed and observed feature ratings corresponding to the twenty scenarios for each participant. This was repeated for each feature (20) and each participant (50). To statistically evaluate whether reconstructions were greater than would be expected by chance, signed-rank tests were applied to compare the set of 50 correlation coefficients to zero (no correlation). This produced a set of 20 *p*-values, which were corrected according to false discovery rate (FDR[35]).

As shown in Fig. 2A, ratings for 11 out of the 20 experiential features were reconstructed with statistically significant accuracy across participants [FDR(p) < 0.05]. In descending order of reconstruction accuracy, these were: Social, Speech, Communication, Head, Auditory, Pleasant, Music, Lower-limb, Bright, Motion, and Touch. As such, accuracies were highest for features associated with social interaction. However, 9/20 features were not reconstructed successfully. Besides the possibility that these features were represented differently in the autobiographical imagery and sentence comprehension fMRI datasets, the analyses in subsequent sections suggest that: (1) the current full-cortex regression analysis failed to get a good fit on some features that were reconstructed from analyses within DMN subsystems (specifically, Path, Unpleasant, Taste, Color, and Landmark). (2) Some features may have been relatively underrepresented in the two datasets, limiting the opportunity to reconstruct them.

To estimate how accurately feature ratings for specific scenarios could be reconstructed, an analogous analysis was conducted on scenario vectors (e.g., the twenty feature ratings corresponding to Funeral). This was evaluated by computing Spearman correlation between the reconstructed and the observed rating vectors, repeated for each scenario and participant (Supplementary Fig. 4A). Besides Restaurant and Reading, none of the scenarios appeared to be directly related to situations described in the sentence training dataset (Supplementary Fig. 4B). 12 out of 20 scenarios were reconstructed with statistically significant accuracies. In descending order of mean reconstruction accuracy (Spearman correlation in parentheses), the reconstructed scenarios were: Telephoning (0.22), Funeral (0.20), Cooking (0.18), Exercising (0.14), Party (0.13), Bathing (0.13), Barbecue (0.11), Shopping (0.10), Restaurant (0.08), Museum (0.08), Housework (0.08), and Internet (0.06). This selection is likely to echo the importance of the 11 reconstructed features in Fig. 2A to characterizing the scenarios. For example, Telephoning is characterized by high ratings on Speech, Communication and Social (Supplementary Fig. 1), which were the three most accurately reconstructed features.

To then evaluate whether the reconstructed scenario vectors distinguished between pairs of observed scenario vectors, within each participant, a two-alternative forced-choice scenario discrimination analysis was run (see Supplementary Fig. 5 caption for details). Statistical significance was evaluated in each participant with permutation tests. Scenario pairs were discriminated with a mean accuracy of 64%, with significant permutation tests (p < 0.05) in 25/50 of participants. There were no significant differences in accuracy between young and elderly adults. To estimate which specific scenario pairs were most accurately distinguished, sign tests were used to evaluate correct or incorrect (1 or 0) discriminations against 0.5, for the fifty participants. Forty-six percent of the 190 scenario pairs were discriminated with FDR[35](p) < 0.05. Mean pairwise discrimination accuracies for each scenario across all 50 participants were as follows, with the second number in parenthesis counting the number of sign tests (out of 19) with FDR(p) < 0.05: Telephoning (78%, 16), Funeral (71%, 14), Party (67%, 13), Exercising (65%, 11), Dancing (64%, 11), Wedding (63%, 11), Cooking (67%, 10), Driving (66%, 10), Barbecue (65%, 10), Shopping (65%, 10), Internet (64%, 8), Museum (63%, 8), Writing (61%, 7), Housework (60%, 7), Restaurant (62%, 6), Bathing (62%, 6), Reading (60%, 6), Resting (61%, 5), Movie (58%, 3), Festival (56%, 2). Similar patterns were observed for young and elderly adults (Supplementary Fig. 6). Besides Festival, socially related scenarios appeared to be the most discriminable, echoing the relatively strong reconstruction of socially related feature ratings. It remains unclear why Festival was weakly reconstructed.

In sum, by reconstructing experiential feature ratings that also discriminate between scenarios to some degree, the section has presented evidence that (some) experiential features are jointly encoded in fMRI data associated with autobiographical mental imagery and sentence comprehension. However, as they stand, these results do not indicate whether the reconstructions reflect inter-participant differences associated with idiosyncratic autobiographical experiences.

Reconstructed mental image features ratings reflect inter-participant differences

To estimate whether fMRI-based rating reconstructions reflected participant-specific autobiographical content, we ran an individual differences analysis (Fig. 1D). This was first evaluated by comparing the entire matrices of reconstructed feature ratings across the twenty scenarios to observed ratings in the same/different participants. This test was implemented using a two-alternative forced-choice participant discrimination algorithm, which repurposed ref. 36's word discrimination approach: two participants were selected, the corresponding two matrices of reconstructed feature ratings, and two matrices of observed ratings (20 scenarios by 20 features) were all vectorized to produce four vectors with 400 entries, and the two reconstructed and two observed vectors were cross-correlated with Spearman correlation. The participant pair were judged to be discriminated if the sum of congruent correlation coefficients (P1:Reconstructed vs P1:Observed + P2:Reconstructed vs P2:Observed) exceeded the sum incongruent coefficients (P1:R vs P2:O + P2:R vs P1:O). As there is a 50% chance of success by chance, pairwise decoding was repeated across all participant pairs. Permutation tests with 50,000 random shuffles were used to estimate statistical significance.

In support of the overarching hypothesis that reconstructed features would reflect participant-specific ratings—and therefore idiosyncratic autobiographical content—interparticipant differences were discriminated with a 70% accuracy (p < 1e-4) (Fig. 2B). Inter-participant differences were also detected when the analysis was restricted to each age group: elderly: 77% (p < 1e-4), young: 67% (p = 0.016).

To explore whether specific experiential features underpinned discrimination, the interparticipant decoding analysis was repeated on individual features (e.g., the vector of social ratings across twenty scenarios). As is illustrated in Fig. 2B, this analysis suggested that reconstructed social ratings played a dominant role, not only discriminating between all 50 participants (76%, p < 1e-4, FDR corrected), but also between young participants only (78%, p < 0.002, FDR corrected), and elderly participants only (76%, p < 0.008, FDR corrected). Reconstructed Speech ratings also discriminated between all 50 participants combined, and elderly but not young participants. There was also tentative evidence that reconstructed Music ratings may discriminate young but not elderly participants.

Repeating the analogous analyses on individual scenario vectors (e.g., the 20 feature ratings for Funeral) found that only Exercising reflected interparticipant differences, except when tests were restricted to elderly participants (All 50 participants: 66% FDR(p) = 0.032; young: 72%, FDR(p) < 0.042; elderly 61%, FDR(p) = 0.3 (Supplementary Fig. 4C). However, weaker outcomes were to be expected when analyzing individual scenarios, as opposed to features, because each analysis evaluated just 5% of the fMRI data (i.e., one rather than all twenty scenarios). Inspection of Supplementary Figs. 1–3 suggested that young people were more variable in their self-reports of exercising (incl. running, sports and music), whereas the elderly group tended to imagine going for a walk.

Critically, this section has presented: (1) evidence that elements of autobiographical mental images and third-person sentence semantics are represented in common brain systems; (2) proof-of-principle that autobiographical experience can be decoded from fMRI data with a pre-trained decoder that was not fit to the participant. However, these results did not identify how different DMN subsystems contribute to feature reconstruction. To recap, Core and MT-DMN have often been associated with episodic recollection and simulation, and FT-DMN is often associated with language comprehension and concept representation (see Supplementary Fig. 7).

DMN subsystems reconstruct interparticipant differences in feature ratings

To identify which experiential features could be reconstructed from the three DMN subsystems, fMRI-to-model decoding mappings were fit on the sentence comprehension data to each DMN subsystem in isolation. The mappings were then applied to reconstruct feature ratings from the autobiographical imagery fMRI data, which were subsequently evaluated for interparticipant differences (as in Fig. 2D). This revealed that participant-specific ratings could be reconstructed from each DMN subsystem (Fig. 3) and differences subsystems reconstructed different features (Fig. 4).

Collectively, the DMN subsystem analyses reconstructed 15 features in total and therefore recovered features that were not detected in the full-cortex analysis (Fig. 2), despite this also including DMN. We assert that this was due to the overfitting of fMRI noise outside of DMN, which drowned out some signal in DMN. In support of this, most of the newly reconstructed features came from MT-DMN, which is a relatively small subsystem, containing 27/1000 cortical ROIs, which in the full cortex analysis would have been vulnerable to noise contamination from the other 973 ROIs.

The pattern of features reconstructed from Core and FT-DMN strongly resembled one another, and the full cortex analysis (Fig. 3A vs Fig. 2A). Ratings associated with social interaction (Social, Communication and Speech) were recovered with relatively high accuracy. Social ratings were more accurately reconstructed from Core-DMN than FT-DMN, but besides this were no significant differences between these two subsystems (see Fig. 4 for test statistics).

Conversely, MT-DMN reconstructed ratings associated with locomotion (Path) and perception (Color, Taste) more accurately than either Core or FT-DMN. Although Social ratings were reconstructed,

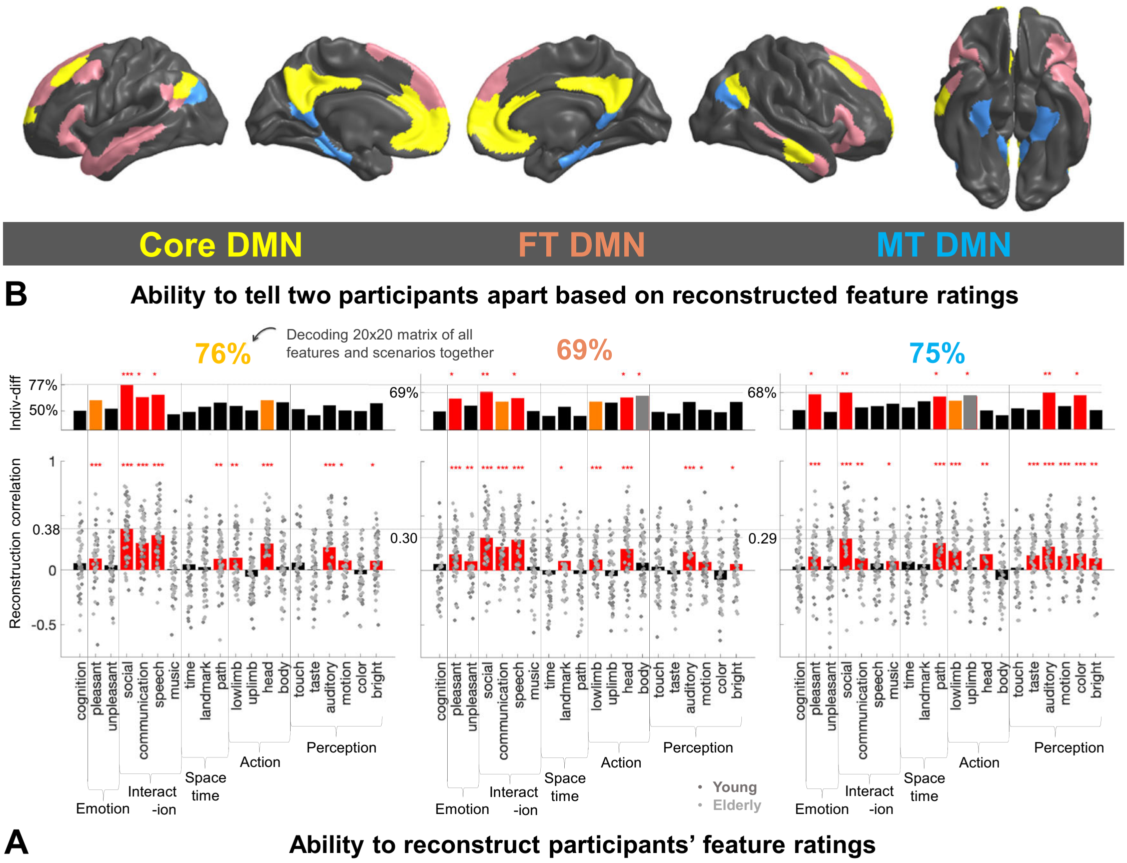

Fig. 3 | Autobiographical mental image feature ratings were reconstructed from each DMN subsystem. Feature reconstruction/individual differences outcomes replicating Fig. 2A, B within each default mode network (DMN) subsystem are displayed (please see Fig. 2 caption for analytic details). To recap, Core-DMN and medial temporal DMN (MT-DMN) contain brain areas that are often associated with autobiographical simulation, whereas fronto-temporal DMN (FT-DMN) contains areas that are often associated with language and conceptual representation (see Supplementary Fig. 7). However, the current results suggest language semantics and self-generated mental images share some common representations in each DMN subsystem. MT-DMN appears to be the most distinct in terms of information encoded (see Fig. 4). **A** Feature reconstruction accuracies. Statistical significances were determined using one-sample signed ranks tests to compare reconstruction accuracy correlation coefficients to zero ($n = 50$ participants, one-tailed). P-values were adjusted according to the false discovery rate (FDR)[35]. Red star notation for statistical significances is: ***FDR(p) < 0.001, **FDR(p) < 0.01, *FDR(p) < 0.05. **B** Individual differences analyses. Red bars indicate significant decoding accuracies where FDR(p) < 0.05 (permutation test, one-tailed). Orange bars indicate decoding accuracies where permutation tests returned $p < 0.05$, but values did not survive FDR correction. Gray bars indicate seemingly inconsistent outcomes where individual differences were detected for features that were not reconstructed with significant accuracy in (**A**). Please see main text for further details. Brain images were generated using BrainSpace v1.10[57]. Source data are provided as a Source Data file.

this was with significantly lower accuracy than Core or FT-DMN. Speech was less accurately reconstructed by MT-DMN than Core and FT-DMN, and Communication was less accurate than Core-DMN. (see Fig. 4 for test statistics).

Interparticipant differences in autobiographical imagery feature ratings were reflected in each DMN subsystem (Fig. 3B). When interparticipant differences were evaluated on the entire 20 feature*20 scenario rating matrices, participant pairs could be told apart with accuracies of: Core-DMN: 76%, $p = 2e\text{-}5$; FT-DMN: 69%, $p = 2e\text{-}4$; and MT-DMN: 75%, $p = 2e\text{-}5$. Interparticipant differences within each age group were also detected in each subsystem: Core-DMN: young = 69%, $p = 0.01$/elderly = 83%, $p = 2e\text{-}5$; FT-DMN: young = 66%, $p = 0.02$/elderly = 72%, $p = 0.003$; MT-DMN: young = 67%, $p = 0.02$/elderly = 82%, $p = 2e\text{-}5$. When each feature was evaluated in isolation (Fig. 3B bar plots), Core-DMN captured interparticipant differences in ratings associated with social interaction (social, communication, speech), and FT-DMN broadly reflected Core-DMN with qualitatively lower discrimination percentages. Differently, MT-DMN reconstructed interparticipant differences in ratings on path, color, and audition. Interparticipant differences in only the social ratings were reflected in all three DMN subsystems. Running the analogous interparticipant differences tests on individual scenario vectors yielded no statistically significant outcomes ($p < 0.05$) in any DMN subsystem, following FDR correction.

There were two apparent inconsistencies, where interparticipant differences were detected on weakly reconstructed features (see gray bars on Body/Upper limb in FT/MT-DMN in Fig. 3B). The precise cause is unknown and could reflect false positives, or the reconstruction of gross trait-level differences but not fine-grained patterns across scenarios. However, the underlying cause may be the same for both Body and Upper limb, because the two were strongly correlated in both the crowd-sourced sentence ratings (Spearman $r = 0.51$) and the participant-specific mental image ratings [mean Spearman $r = 0.59$ over fifty participants, following r-to-z (arctanh) transformation of correlation coefficients and z-to-r (tanh) back transformation of the mean].

Critically, this section has presented evidence that: (1) participant-specific internally generated autobiographical mental images and externally elicited general sentence semantics share representations in each DMN subsystem, to some extent at least. (2) DMN subsystems reconstruct different features, with Core and FT-DMN reconstructing social features more accurately than MT-DMN, and MT-DMN more

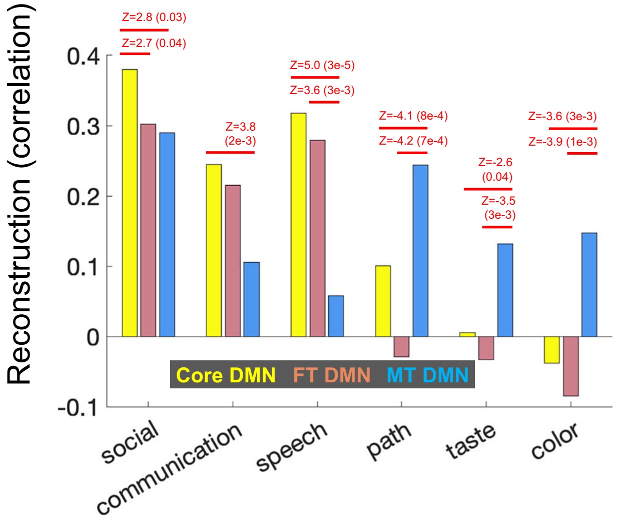

Fig. 4 | Differences in feature reconstruction accuracy between DMN subsystems. Differences between default mode network (DMN) subsystems were identified using two-tailed signed ranks tests to compare Core-DMN vs fronto-temporal DMN (FT-DMN), Core-DMN vs medial temporal DMN (MT-DMN) and FT-DMN vs MT-DMN in their ability to reconstruct each feature (50 participants). There were 60 signed ranks tests in total (3 DMN pairs*20 features). The 60 resultant *p*-values were adjusted according to false discovery rate (FDR)[35]. The figure displays only the comparisons yielding statistically significant differences [FDR(*p*) = 0.05]. Significant differences were also detected for Body and Time, but these are not displayed, because neither Body nor Time was reconstructed from any DMN subsystem with significant accuracy, according to the statistical significance thresholds used in Fig. 3A. Source data are provided as a Source Data file.

accurately reconstructing features associated with locomotion and perception.

The subset of experiential features that contributed to modeling fMRI data

Because experiential features may correlate (e.g., socialization often involves speech), the ability to reconstruct one feature, such as speech, may derive from fMRI activation associated with other concurrent aspects of interpersonal interaction (e.g., estimating other peoples' intent/knowledge/feelings, aka theory of mind). To help disentangle these effects and simultaneously pinpoint the cortical regions associated with the different experiential features, we deployed a predictive variance partitioning analysis[37].

For this purpose, rather than reconstructing individual features from fMRI data (Fig. 1C), an fMRI encoding model was fit to predict sentence activation in individual ROIs (1000) based on all twenty crowdsourced experiential features (see Supplementary Fig. 8A). The encoding model was then applied to predict each participants' mental image fMRI data, based on the corresponding participant-specific feature ratings (Supplementary Fig. 8B). Prediction accuracy was determined with Spearman correlation between predicted and observed ROI activation across the twenty autobiographical scenarios. Cortical maps of prediction accuracies are included in Supplementary Fig. 9 with ROIs in and around DMN encoding experiential features.

To then estimate whether each individual feature uniquely contributed to predicting fMRI activation, a second comparative regression mapping was fit, with that feature deleted, so that the model had nineteen features. The unique predictive value added by the feature was estimated as the improvement inprediction accuracy brought by incorporating the feature into the encoding model. This was determined statistically with two-sample signed ranks tests, evaluating whether the twenty feature model yielded more accurate predictions than the nineteen feature model (50 participants). To increase analytic power, the analysis was focused on ROIs within the DMN [175/1000 of the cortex] (see Fig. 5). Comparative results on all 1000 ROIs are included in Supplementary Fig. 10. To assist interpretation, cortical maps of the encoding model beta-coefficients are in Supplementary Fig. 11.

Seven experiential features uniquely contributed to predicting the autobiographical imagery fMRI data (Fig. 5). Most prominent was Social, which predicted a swathe of cortex spanning parts of all three DMN subsystems, including brain areas linked to theory of mind[38]. Also prominent was Path, which helped to predict MT-DMN. Color and Lower-limb also contributed to predicting MT-DMN. Otherwise, Unpleasant (inferior frontal and left temporal pole) and Motion (inferior frontal) contributed to FT-DMN, and Communication contributed to Core-DMN (ventromedial prefrontal cortex). Running the analysis on the entire cortex (Supplementary Fig. 10), rather than the DMN only, identified 6/7 features, excluding Unpleasant, with similar but more patchy cortical distributions, due to the stricter nature of FDR correction across 825 extra ROIs outside DMN, which often did not correlate with the model (Supplementary Fig. 8).

The finding that prediction was driven by seven features raised a question over why the remaining thirteen features did not contribute. This appears to be in part due to limitations in the encoding mapping's ability to distinguish these features, and by extension, the sentence comprehension fMRI data on which the mapping was fit. To come to this conclusion, we undertook a comparative analysis on the sentence comprehension dataset alone, where model-to-fMRI encoding mappings were trained and evaluated across sentences and participants. Here the same seven features accounted for all sentence variation in 93% (478/513) of brain regions that were predicted by all twenty features (Supplementary Fig. 12). To find out if the remaining unpredicted variance could be accounted for by any single one of the thirteen features, we iterated through those features and added each in turn to the seven-feature model, before fitting an encoding mapping to predict the sentence comprehension data (across sentences and participants). Features were added in order of feature reconstruction accuracy, which was computed in a separate analysis of the sentence comprehension data (Supplementary Fig. 13). Iteration was halted if/when all DMN activation was predicted. This analysis found that by including Time, all remaining variance was predicted in 99% (506/513) of brain regions, including all DMN (Supplementary Fig. 14). This suggested that the sentence comprehension fMRI dataset independently represented correlates of eight of the twenty features. It is possible that variation in Time ("occurrences at typical times", Supplementary Table 1) was underrepresented in the autobiographical imagery data, given that the scenarios were all occurrences associated with times. In contrast, the sentence data (Supplementary Table 2) had more variety in including examples of time-agnostic statements (e.g., The red pencil was on the desk.) as well as occurrences at typical times (e.g., The couple laughed at dinner.).

Another possibility was that the current experiential features neglected to model some critical information, which seemed possible given that it is now commonplace to model fMRI data with hundreds or thousands of features from language AI models.

Decoding outcomes achieved with experiential feature ratings were commensurate with GPT-2

To explore the possibility that the current set of experiential features had neglected to model critical information, we ran a comparative analysis with the language model GPT-2-medium[31]. GPT-2 was chosen because it has generated strong fMRI predictions of language comprehension in other work[39,40]. Both the 240 sentence stimuli and participants' verbal descriptions of mental images were modeled with layer 16 (of 24), given evidence that layers at ~2/3 depth are a good choice for modeling fMRI data[41]. Each sentence or autobiographical

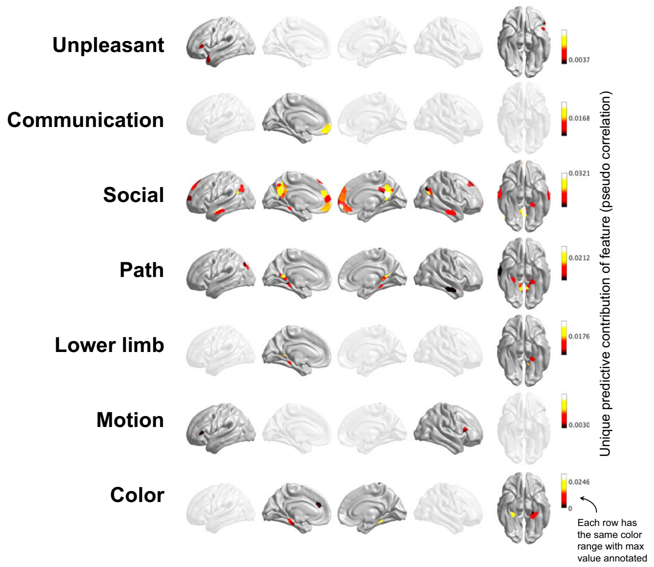
Unpleasant
Communication
Social
Path
Lower limb
Motion
Color
0.0037
0.0168
0.0321
0.0212
0.0176
0.0030
0.0246
0
Unique predictive contribution of feature (pseudo correlation)
Each row has the same color range with max value annotated

Fig. 5 | Experiential features' independent contributions to predicting mental image fMRI data. To estimate which experiential features independently contributed to fMRI prediction, and where in default mode network (DMN) was predicted, a variance partitioning method was used. To estimate unique feature contributions, an fMRI encoding model was first fit to the sentence dataset—namely, a set of regression mappings to predict the average activation of each DMN Schaefer ROI, based on the twenty crowd-sourced feature ratings (see Supplementary Fig. 8). The encoding model was transferred to predict the mental image fMRI data with the participant-specific mental image feature ratings. To then isolate the predictive contribution of a feature Y, a second encoding model was fit on the sentence fMRI data to the nineteen other features (excluding Y). The contribution of Y was estimated as the difference in predicted variance (estimated as r^2) between the full set of twenty featuresand the subset of nineteen that excluded Y. This value is presented on the cortical maps as a pseudo correlation estimate: sqrt[r^2(full) -r^2(subset)]. Predictive contributions are displayed only if prediction accuracies derived from all twenty features were significantly greater than the subset of nineteen. Significance was determined using signed-ranks tests on each DMN Schaefer ROI (fifty participants, one-tailed). *P*-values were corrected according to the false discovery rate (FDR)[35]. Values are displayed for ROIs where FDR(p) < 0.05. A comparative analysis of all 1000 ROIs is in Supplementary Fig. 10. To assist visual interpretation, the encoding model beta coefficients for the twenty-feature model are in Supplementary Fig. 11. Brain images were generated using BrainSpace v1.10[57].

mental image description was modeled with a single vector of 1024 features, derived by processing the sentence/description through GPT-2 and pointwise averaging L16 activation for all constituent tokens.

An fMRI-to-GPT-2 decoding mapping (equivalent to Fig. 1C) was fit on the sentence reading data in each DMN subsystem with ridge regression and then transferred to reconstruct GPT-2 features from the mental image data (equivalent to Fig. 1D). Reconstruction accuracy was evaluated against participant-specific GPT-2 models of mental images, and an individual differences analysis was run equivalent to Fig. 1D.

Because comparisons of individual feature reconstruction accuracies between GPT-2's 1024 features and the experiential model's twenty features had no clear interpretation, we evaluated the models' comparative ability to reconstruct mental image scenario similarities with representational similarity analysis[42] (RSA). fMRI reconstructions of GPT-2 and the experiential model were re-represented as inter-scenario correlation matrices (20*20), with each matrix entry reflecting the similarity between a pair of scenarios. The corresponding participant-specific self-report models were likewise re-represented as correlation matrices. To compare the reconstructed to the observed correlation matrices, the lower off-diagonal triangle of each (symmetric) correlation matrix was extracted and vectorized, and Spearman correlation was applied to the reconstructed vs observed vectors. This yielded an RSA correlation coefficient for each participant and model. To evaluate whether GPT-2 or the experiential model recovered scenario correlation structure more accurately, two-tailed signed ranks tests were deployed to the fifty RSA coefficients from each model. There were no significant differences between models, suggesting no major quantitative advantage to either approach (Fig. 6). Likewise, GPT-2 reconstructions were no better at discriminating between scenario pairs than the experiential model (both 64%), and 41% of scenario pairs were discriminated with statistical significance (Supplementary Fig. 15 vs Supplementary Fig. 5). Similar scenario pairs were discriminated in young and elderly participants (Supplementary Fig. 16). Repeating the interparticipant differences analyses of Fig. 3B with GPT-2 also yielded broadly similar outcomes to the experiential model: Core-DMN 74% ($p < 2e\text{-}5$); FT-DMN: 64% ($p < 0.01$); MT-DMN 73% ($p < 2e\text{-}5$). Thus, there appeared to be little to separate GPT-2 and experiential feature-based outcomes.

Nonetheless, it was also possible that the choice to average voxel activation within Schaefer atlas ROIs—which was undertaken to facilitate data integration across participants—may have blurred detailed voxel-level information content, and disadvantaged GPT-2. Whilst repeating the entire analyses at voxel level was beyond the scope of the study, we estimated the impact of averaging voxel activation within ROIs with a second RSA comparing the mental image fMRI data itself (rather than model reconstructions) to the participant-specific GPT-2 and experiential models. RSA was repeated when fMRI data were averaged within Schaefer ROIs, and then with raw voxel activation. Although both treatments of fMRI data yielded significant RSA coefficients in each analysis, RSA coefficients tended to be equivalent or greater for the Schaefer-averaging approach, whether using GPT-2 or the experiential model (Supplementary Fig. 17). Based on these tests, we do not believe that the current analyses were substantially disadvantaged by using experiential features, or by averaging activation within ROIs. However, in other circumstances, given larger fMRI datasets that span human experience more comprehensively (e.g., ref. 5), it would be natural to expect the potential to benefit from deploying voxel-wise approaches on models with more features.

Discussion

This study has shown that the human brain encodes elements of imagined autobiographical experiences and read sentence meaning in a shared representational space, where in imagination, the representational details were shaped from inside by autobiographical memory, and in reading, the representational details were shaped from outside by a sequence of words. This shared representational space spans Core and MT-DMN, which contain brain areas often associated with episodic recollection and simulation[9–11] as well as FT-DMN, which contains areas often associated with language[12–14] and general semantics[15]. These findings help to broaden a body of research that is increasingly viewing the boundaries between episodic and semantic memory systems as blurred[43–47]. The demonstration that cross-participant reconstructions of feature ratings could discriminate interparticipant differences—at least when evaluated over a set of autobiographical scenarios (Figs. 2B and 3B)—provides proof of principle that idiosyncratic features of self-generated mental states can be extracted from fMRI data without first tailoring a decoding model to those specific individuals.

There were some similarities and some differences in the feature ratings that could be recovered from the three DMN subsystems. For instance, social interaction ratings were recovered from all three subsystems with relatively high accuracy comparative to other features. This may reflect theory of mind processes associated with inferring other peoples' intent, feelings and knowledge, especially in FT and Core-DMN (see ref. 38). Conversely, MT-DMN was most

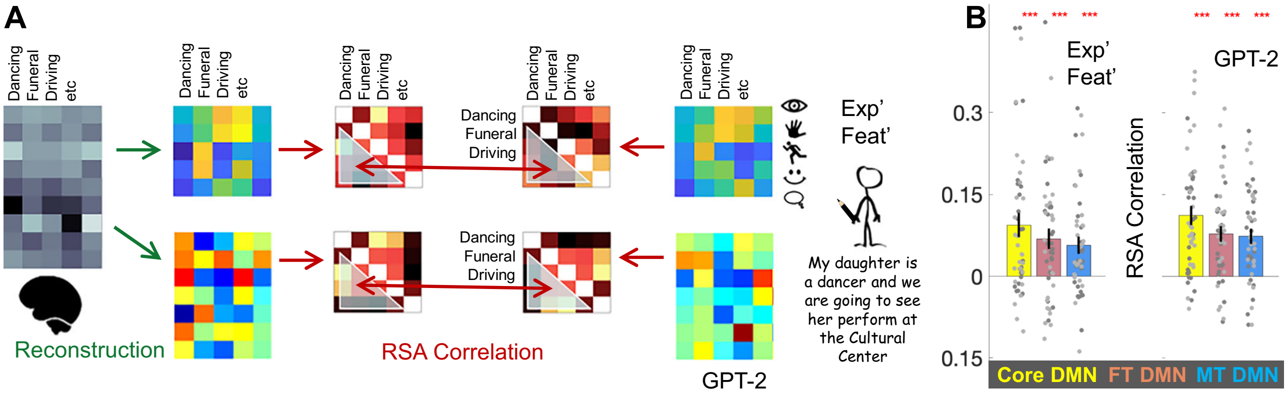

Fig. 6 | Reconstruction of experiential mental image scenario structure was commensurate with GPT-2. A To explore the possibility that the twenty experiential features were too few to comprehensively model the fMRI data, a comparative analysis was run using 1024 feature GPT-2 layer-16 embeddings derived from participants' autobiographical mental image descriptions and the 240 sentence stimuli. GPT-2 features were reconstructed from the autobiographical imagery data using an fMRI-to-GPT-2 mapping fit on the sentence comprehension data (equivalent to Fig. 1C). Because there was no clear way to interpret comparisons in individual feature reconstruction accuracies between GPT-2 vs the experiential model, representational similarity analysis (RSA) was used to compare the relative ability to reconstruct scenario similarity structure. **B** RSA correlation coefficients are illustrated for individual participants as dots in the bar plot (dark gray: young, light gray: elderly). Bars are mean accuracy, and error bars represent SEM. Both GPT-2 and the experiential model yielded RSA correlation coefficients that were significantly greater than zero in the Core default mode network (Core-DMN), fronto-temporal DMN (FT-DMN) and medial temporal DMN (MT-DMN), according to one-sample signed ranks tests against zero (fifty participants, one-tailed, corrected according to false discovery rate (FDR)[35]). Significance is illustrated above bars using star notation: ***FDR(p) < 0.001. Critically, there were no significant differences in RSA coefficient magnitude between GPT-2 and the experiential model in any DMN subsystem, according to two-sample signed ranks tests (fifty participants, two-tailed). Source data are provided as a Source Data file.

distinctive, and most accurately reconstructed ratings related to locomotion (Path) and perception (Color/Taste). This distinction may echo MT-DMN's proposed role in visuospatial mental simulation[7,8], however, future work would be needed to clarify this. On the one hand, the reconstruction of these features could reflect the mental visualization of scenes associated with people, travel and food. On the other hand, one might expect neural correlates of active visual imagery to be represented relatively weakly by the current decoding model, because it was trained on sentence comprehension data, when active imagery was not a task requirement.

Core and FT-DMN reconstructed a broadly similar set of features, despite their prior associations with autobiographical simulation[9–11] vs language comprehension[12–14], and their tendency to decouple in cognitive tasks that do not activate autobiographical memory[48,49], or do not require stimuli to be placed into recent semantic context[49]. However, areas within Core-DMN have also been implicated in semantic representation[16–23]. The current results contribute to this picture by helping to characterize the circumstances in which Core and FT-DMN do couple together, in the encoding of social information during both language comprehension and autobiographical imagery. This is consistent with theories that DMN operates to form context-dependent models of situations as they unfold[3] as multimodal simulations of phenomenological experience[50].

The demonstration that a pre-trained cross-participant fMRI decoder can reconstruct experiential feature ratings that both discriminate interparticipant differences across sets of scenarios (Figs. 2B and 3B) and different scenarios within participants (Fig. 6, and Supplementary Figs. 5,6,15,16) implies that in future research: (1) experiential features associated with self-generated thoughts and language semantics may be reconstructed, to first approximation, from new participants' fMRI data, using off-the-shelf decoding mappings, such as the one derived here (Fig. 1D and Fig. 2/Fig. 3); (2) reconstructed ratings may distinguish between different peoples' phenomenological experience of the same stimuli. With these ideas in mind, we make publicly available the decoding mapping used here, as well as a new mapping fit across the sentence and mental imagery datasets that may benefit from the broader experiential coverage. We hope these will be useful for inferring the experiential content of mental states underlying new fMRI datasets in both young and elderly adults.

Whilst we hope the current decoding approach will have application to interpreting external fMRI datasets, and especially for indexing social cognition, such applications will be limited to the features that could be reconstructed/predicted here. The critical features that drove cross-task/participant prediction/reconstruction were: Social, Communication, Unpleasant, Path, Lower limb, Color, and Motion (Fig. 5). Time could additionally be predicted/reconstructed in analyses conducted within the sentence comprehension dataset (Supplementary Fig. 14). Future work, using larger datasets with more comprehensive experiential coverage may lead to fMRI decoders that reconstruct more features more accurately. There also seems to be room for improvement in decoding the current autobiographical imagery dataset, given that: (1) RSA comparisons between fMRI data and the self-report models yielded stronger correlation values than comparisons between model reconstructions and self-report models (Fig. 6 vs Supplementary Fig. 17); (2) the experiential model and GPT-2 played complementary roles in capturing the representational structure of the autobiographical imagery fMRI (see Supplementary Fig. 17 caption).

To close, the study has demonstrated that experiential semantic content of sentences is encoded to some degree in the same cortical areas and using the same representational codes as autobiographical mental imagery. In so doing, the study has also revealed how general-purpose brain decoding models can be transferred across both people and cognitive tasks. We consider this an early step towards unifying brain models of long-term memory and language processing. Going forward, future models will also need to capture the vivid autonoetic recollections that distinguish episodic memories from conceptual processing[1,2]. In time, such brain models may lead towards exciting new brain-computer interface technologies.

Methods

Autobiographical mental imagery participants

The original data collection was approved by the University of Rochester research subject review board (RSRB00067540). Participants received monetary compensation and were required to understand the experimental procedure and give their consent by signing an informed consent form. 50 participants' data were reanalyzed. Fifty percent of the data were collected from healthy elderly adults (mean ± SD age was 73 ± 7 years, 16 F, 9 M)[28]. The other 25 were healthy young adults (mean ± SD age = 24 ± 3, 16 F, 9 M)[29]. Sex was self-reported by participants. Sex/gender differences were not critical factors in the study design, nor the hypotheses tested. One elderly participant was excluded from the current study because of discordant BOLD/structural fMRI co-registration when the data were preprocessed with fMRIPrep for the current study (SPM had been used in ref. 28).

Autobiographical mental imagery scenario stimuli

Twenty scenarios were preselected by ref. 28 to be diverse events that almost every participant would have personally experienced. Scenarios were always presented to participants in the following form: "A X scenario", or "An X scenario" where X is a placeholder for: resting, reading, writing, bathing, cooking, housework, exercising, internet, telephoning, driving, shopping, movie, museum, restaurant, barbecue, party, dancing, wedding, funeral, and festival.

Autobiographical mental imagery experimental procedure

At the beginning of the experiment, an experimenter went through the twenty scenario prompts asking participants to vividly imagine themselves experiencing each scenario, and to actively simulate their sensory experiences, actions and feelings. Once participants had formed a rich mental image, they provided a brief verbal description of their mental image, which was transcribed by the experimenter. After all the scenarios had been imagined and described, the experimenter went through the scenarios again, first reminding the participant of their scenario description, and then recording the participant's ratings of the scenario on 20 experiential features, as well as the likelihood that their mental image reflects a real personal experience (rather than being fictional) and the vividness of the mental image. Ratings were on the scale [0 6]. See Supplementary Table 1 for the specific instructions for rating each feature, adapted from ref. 30. The 20 features were a subset of the 65 features collected by ref. 30 that were chosen to broadly span the twelve domains of experience that were originally identified by ref. 30 along with the author's intuitions of which features were most relevant for the 20 scenarios[28]. The need to collect 20 rather than all 65 features was mandated by experimental time constraints[28].

Autobiographical mental imagery MRI data collection parameters

Imaging data were collected at the University of Rochester Center for Brain Imaging using a 3T Siemens Prisma scanner (Erlangen, Germany) equipped with a 32-channel receive-only head coil. The fMRI scan began with a MPRAGE scan (TR/TE = 1400/2344 ms, TI = 702 ms, flip angle = 8°, FOV = 256 mm, matrix = 256 × 256 mm, 192 sagittal slices, slice thickness = 1 mm, voxel size 1 × 1 × 1 mm^3). fMRI data were collected using a gradient echo-planar imaging (EPI) sequence (TR/TE = 2500 ms/30 ms, flip angle = 85°, FOV = 256 mm, 90 axial slices, slice thickness = 2 mm, voxel size 2 × 2 × 2 mm^3, number of volumes = 639).

Autobiographical mental imagery fMRI experiment

Prior to scanning, participants were reminded of their verbal descriptions of the 20 scenarios (see above) and were requested to vividly reimagine the same mental images in the scanner when prompted. Outside the scanner they then underwent a single dry run of the fMRI experiment below (a single viewing of all 20 stimuli) to familiarize them with the setup[28].

Participants then underwent a single uninterrupted fMRI session in which the 20 scenario stimuli were presented five times over (five runs). The five repeats enabled the computation of average mental image representations for each scenario, to counteract fMRI noise. Stimulus order was randomized within each run. Runs were separated by a 15s interval, in which a second-by-second countdown was displayed (e.g., "starting run 2 in 13 s"), which was followed by 7.5 s of fixation cross preceding the first stimulus of the run. All participants reported that they had been able to imagine the scenarios on prompt.

Stimulus prompts (e.g., "a dancing scenario") were presented one by one in random order, on a screen in black Arial font (size 50) on a gray background. Prompts remained on screen for 7.5 s (3TRs), during which time participants vividly imaged themselves in the scenario. After the prompt was removed, there was a 7.5 s delay prior to the next prompt, during which time a fixation cross was displayed. Participants were instructed to attempt to clear their minds when the prompts were removed.

Autobiographical mental imagery MRI data preprocessing

MRI data were preprocessed using fMRIPrep[32] to counteract head motion and spatially normalize images to a common neuroanatomical space (MNI152NLin2009cAsym). The boilerplate template detailing the procedure is in Supplementary Materials. To counteract nuisance signals and potential confounds in the fMRI data a comprehensive selection of nuisance regressors generated by fMRIPrep were regressed out from each voxels time series. These were 24 head motion parameters (including translation, rotation and their derivatives), white-matter and cerebrospinal fluid timeseries and cosine00, cosine01, cosine02, cosine03. Confound removal was implemented through computing a single multiple regression across the entire fMRI timeline: first, each voxel and each nuisance regressor's time-series was separately z-scored (by subtracting the mean and dividing by the standard deviation). Second, a separate multiple regression was fit for each voxel, mapping nuisance regressors to predict voxel activation. The residuals from the regression (computed separately for each voxel) were taken forward to further analysis.

For analyses, we reduced each participant's fMRI data to obtain a single volume for each imagined scenario. To do this, we first computed a single volume for each scenario replicate by computing the voxel-wise mean of 4 fMRI volumes (5–15 s) post stimulus onset. To reduce the five replicates of each scenario to a single scenario volume, we again took the voxel-wise mean. This left 20 scenario volumes per individual. The 5 s (2TR) offset was to accommodate hemodynamic response delay (4–5 s). The four-volume period spans the time until the next stimulus is presented on screen and was set to maximize our chances of capturing the mental image. The averaging approach was chosen in ref. 28 to accommodate the individual differences in hemodynamic response latency and duration associated with self-generating and maintaining mental images, which appeared not to be well suited to modeling with a canonical hemodynamic response function time-locked to stimulus presentation.

Sentence comprehension participants

Participants were 14 healthy, native speakers of English (5 males, 9 females; mean age = 32.5, range 21–55) with no history of neurological or psychiatric disorders. Sex was self-reported. All were right-handed according to the Edinburgh Handedness Inventory[51]. Participants received monetary compensation and gave informed consent in conformity with the protocol approved by the Medical College of Wisconsin Institutional Review Board.

Sentence comprehension stimuli

The stimuli, listed in full in Supplementary Table 2, consisted of 240 written sentences containing 3–9 words and 2–5 (mean ± SD = 3.33 ± .76) content words, formed from different combinations of 141 nouns, 62 verbs, and 39 adjectives (242 words). Sentences were in active voice and consisted of a noun phrase followed by a verb phrase in past tense, with no relative clauses. Most sentences (200/240) contained an action verb and involved interactions between humans, animals and objects, or described situations involving different entities, events, locations, and affective connotations. The remaining 40 sentences contained only a linking verb ("was"). All sentences were pre-selected as experimental materials for the Knowledge Representation in Neural Systems project[52] (www.iarpa.gov/index.php/research-programs/krns), sponsored by the Intelligence Advanced Research Projects Activity.

Crowd-sourced experiential feature rating for sentence stimuli

Experiential feature ratings were collected on Amazon Mechanical Turk for each of the 242 content words in the set of 240 experimental sentences[30] (Supplementary Table 2). Workers were asked to rate, on a scale of [0 6], the degree to which a given lexical concept was associated with a particular feature of experience (e.g., "to what degree do you think of a football as having a characteristic or defining color?"). The exact wording of these queries was tailored to the feature and the grammatical class of the word. A total of 7237 rating sessions were conducted, with approximately 30 complete ratings sets (all attributes for a given word) collected for each word. Mean ratings were computed for each word on each feature.

Sentence comprehension experimental procedure

Participants took part in multiple scanning visits, where they were instructed to read sentences and think about their overall meaning (as detailed in ref. 20). In each visit, the entire list of sentences was presented 1.5 times over 12 scanning runs, with each run containing 30 trials (one sentence per trial) and lasting approximately 6 min. The presentation order of each set of 240 sentences was randomly shuffled. Sentences were presented word-by-word using a rapid serial visual presentation paradigm. Nouns, verbs, adjectives, and prepositions were presented for 400 ms each, followed by a 200 ms interstimulus interval (ISI). Articles ("the") were presented for 150 ms followed by a 50-ms ISI. Mean sentence duration was 2.8 s (range [1.4–4.2 s]). Words subtended an average horizontal visual angle of approximately 2.5°.

Sentence comprehension MRI parameters and preprocessing

MRI data were acquired with a whole-body 3T GE 750 scanner at the Center for Imaging Research of the Medical College of Wisconsin using a GE 32-channel head coil. Functional T2*-weighted echoplanar images (EPI) were collected with TR = 2000 ms, TE = 24 ms, flip angle = 77°, 41 axial slices, FOV = 192 mm, in-plane matrix = 64 × 64, slice thickness = 3 mm, resulting in 3 × 3 × 3 mm voxels. T1-weighted anatomical images were obtained using a 3D spoiled gradient-echo sequence with voxel dimensions of 1 ×1 × 1 mm.

Sentence comprehension MRI data preprocessing

fMRI data are publicly available[27] and were preprocessed using AFNI[53]. EPI volumes were corrected for slice acquisition time and head motion. Functional volumes were aligned to the T1-weighted anatomical volume, transformed into a standardized space[54], and smoothed with a 6 mm FWHM Gaussian kernel. The data were analyzed using a general linear model with a duration-modulated HRF, and the model included one regressor for each sentence. fMRI activity was modeled as a gamma function convolved with a square wave with the same duration as the presentation of the sentence, as implemented in AFNI's 3dDeconvolve with the option dmBLOCK. Duration was coded separately for each individual sentence. Finally, a single sentence-level fMRI representation was created for each unique sentence by taking the voxel-wise mean of all replicates of the sentence.

Parcellation of all fMRI data into coarse-grained common neuroanatomical space

Because all primary analyses deployed cross-participant regression decoding/encoding models, we reduced all individuals' fMRI data (mental images and sentences) into the same coarse-grained neuroanatomical space. We used the Schaefer-1000 cortical parcellation[33], which segmented the cortex into 1000 parcels defined by regularities in functional connectivity measures of resting-state fMRI, and linked ROIs to the three DMN subsystems[7,55] analyzed here. This step aimed to preserve large scale cortical activation patterns that we anticipated would be common across people, and blur fine-grained differences in activation that may reflect individual differences in brain anatomy. Data were reduced by averaging voxel activation values within each of the 1000 parcels (aka regions-of-interest or ROIs), yielding a vector of 1000 values for each mental image or sentence per participant. DMN subsystems are illustrated in the context of meta analytic[56] renditions of language/semantic and autobiographical/episodic simulation networks in Supplementary Fig. 7.

Fitting the cross-participant fMRI decoding model

A single cross-participant fMRI-to-semantic model mapping was fit on the entire Sentence Comprehension dataset. To do this, the 14 participants fMRI data (240 sentences by 1000 ROIs) were normalized by z-scoring each ROI across sentences (subtracting the mean from each sentence and dividing by the standard deviation, within participant) and then concatenated into a long (14*240 row) matrix. The crowd-sourced semantic model (240 sentences by 20 features) was likewise z-scored and self-concatenated, 14 times over. Ridge regression was used to fit a mapping (i.e., the regression beta-weights) between the 1000 ROIs and each of the 20 individual semantic features, separately. The same regularization penalty (lambda = 1e5) was used for each feature. The regularization penalty was selected by running a series of cross-validated, cross-participant decoding analyses. In each analysis, the sentence data were randomly split into 120 training and 120 test sentences. The participants were divided into 13 training participants, and 1 test participant. The 13 training participants' fMRI and model data for the 120 training sentences were concatenated as above, and a cross-participant Ridge Regression mapping from ROIs to individual features was fit for each regularization penalty in the range: lambda = [1 1e1 1e2 1e3 1e4 1e5 1e6]. Feature ratings for the 120 test sentences were reconstructed with a matrix multiplication of the test participant's fMRI data (test sentences) and the fMRI-to-model decoding mapping (beta-weights). Reconstruction accuracy was evaluated for each feature separately by computing Spearman correlation between the 120 reconstructed ratings for that feature and the observed feature ratings. To derive a single summary metric of reconstruction accuracy, the mean correlation coefficient across the 20 features was taken. This yielded reconstruction accuracy metrics corresponding to each of the seven lambdas. The reconstruction analysis was repeated 14 times, treating each participant in turn as the test participant, and training with the other 13 participants. The entire process was repeated 30 times with different random selections of 120 training/120 test sentences. The 30 reconstruction accuracy metrics for each participant and lambda were averaged, yielding a 14 (participants) by 7 (lambdas) matrix of accuracy values. As low lambda parameters risk overfitting and high lambda parameters risk underfitting, we sought to identify the ideal lambda as the largest value before median prediction accuracy dropped significantly (measured by sign ranks tests, between consecutive lambda values, $n = 14$).

Fitting the cross-participant fMRI encoding model

A single cross-participant Model-to-FMRI mapping was fit using precisely the same procedure as the section above, except with the direction of the regression mapping reversed, such that it mapped from 20 model features to predict ROI activation.

Reporting summary

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

Data availability

Data are publicly available at: 10.17605/OSF.IO/D4KU8. Source data are provided with this paper.

Code availability

MATLAB R2023B code to reproduce analyses and figures are publicly available at: 10.17605/OSF.IO/D4KU8.

References

1. Tulving, E. *Episodic and Semantic Memory*. in *Organization of Memory* (eds Tulving, E. & Donaldson, W.) 381–403 (Academic Press, 1972).
2. Wheeler, M. A., Stuss, D. T. & Tulving, E. Toward a theory of episodic memory: the frontal lobes and autonoetic consciousness. *Psychol. Bull.* **121**, 331 (1997).
3. Yeshurun, Y., Nguyen, M. & Hasson, U. The default mode network: Where the idiosyncratic self meets the shared social world. *Nat. Rev. Neurosci.* **22**, 181–192 (2021).
4. Pereira, F. et al. Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.* **6,** 9–963 (2018).
5. Tang, J., LeBel, A., Jain, S. & Huth, A. G. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nat. Neurosci.* **1**, 1–9 (2023).
6. Raichle, M. E. et al. A default mode of brain function. *Proc. Natl. Acad. Sci. USA* **98**, 676–682 (2001).
7. Andrews-Hanna, J. R., Smallwood, J. & Spreng, R. N. The default network and self-generated thought: component processes, dynamic control, and clinical relevance. *Ann. N. Y. Acad. Sci.* **1316**, 29–52 (2014).
8. Andrews-Hanna, J. R. & Grilli, M. D. Mapping the imaginative mind: charting new paths forward. *Curr. Dir. Psychol. Sci.* **30**, 82–89 (2021).
9. Rugg, M. D. & Vilberg, K. L. Brain networks underlying episodic memory retrieval. *Curr. Opin. Neurobiol.* **23**, 255–260 (2013).
10. Benoit, R. G. & Schacter, D. L. Specifying the core network supporting episodic simulation and episodic memory by activation likelihood estimation. *Neuropsychologia* **75**, 450–457 (2015).
11. Ritchey, M., Libby, L. A. & Ranganath, C. Cortico-hippocampal systems involved in memory and cognition: the PMAT framework. *Prog. Brain Res.* **219**, 45–64 (2015).
12. Fedorenko, E., Hsieh, P. J., Nieto-Castañón, A., Whitfield-Gabrieli, S. & Kanwisher, N. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *J. Neurophysiol.* **104**, 1177–1194 (2010).
13. Fedorenko, E., Ivanova, A. A. & Regev, T. I. The language network as a natural kind within the broader landscape of the human brain. *Nat. Rev. Neurosci.* **12**, 1–24 (2024).
14. Braga, R. M., DiNicola, L. M., Becker, H. C. & Buckner, R. L. Situating the left-lateralized language network in the broader organization of multiple specialized large-scale distributed networks. *J. Neurophysiol.* **124**, 1415–1448 (2020).
15. Lambon Ralph, M. A., Jefferies, E., Patterson, K. & Rogers, T. T. The neural and computational bases of semantic cognition. *Nat. Rev. Neurosci.* **18**, 42–55 (2017).
16. Binder, J. R., Desai, R. H., Graves, W. W. & Conant, L. L. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* **19**, 2767–2796 (2009).
17. Honey, C. J., Thompson, C. R., Lerner, Y. & Hasson, U. Not lost in translation: neural responses shared across languages. *J. Neurosci.* **32**, 15277–15283 (2012).
18. Simony, E. et al. Dynamic reconfiguration of the default mode network during narrative comprehension. *Nat. Commun.* **7**, 12141 (2016).
19. Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
20. Anderson, A. J. et al. Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cereb. Cortex* **27**, 4379–4395 (2017).
21. Deniz, F., Nunez-Elizalde, A. O., Huth, A. G. & Gallant, J. L. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *J. Neurosci.* **39**, 7722–7736 (2019).
22. Tong, J. et al. A distributed network for multimodal experiential representation of concepts. *J. Neurosci.* **42**, 7121–7130 (2022).
23. Caucheteux, C., Gramfort, A. & King, J.-R. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat. Hum. Behav.* **7**, 430–441 (2023).
24. Chen, J. et al. Shared memories reveal shared structure in neural activity across individuals. *Nat. Neurosci.* **20**, 115–125 (2017).
25. Anderson, A. J. et al. Multiple regions of a cortical network commonly encode the meaning of words in multiple grammatical positions of read sentences. *Cereb. Cortex* **29**, 2396–2411 (2019).
26. Anderson, A. J. et al. An integrated neural decoder of linguistic and experiential meaning. *J. Neurosci.* **39**, 8969–8987 (2019).
27. Anderson, A. J. et al. Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *J. Neurosci.* **41**, 4100–4119 (2021).
28. Anderson, A. J. et al. Decoding individual identity from brain activity elicited in imagining common experiences. *Nat. Commun. 20* **11**, 1–4 (2020).
29. Wang, Y. et al. Decoding visual experience and mapping semantics through whole-brain analysis using fMRI foundation models. Preprint at https://doi.org/10.48550/arXiv.2411.07121 (2024).
30. Binder, J. R. et al. Toward a brain-based componential semantic representation. *Cogn. Neuropsychol.* **33**, 130–174 (2016).
31. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
32. Esteban, O. et al. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
33. Schaefer, A. et al. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb. Cortex* **29**, 3095–3114 (2018).
34. Smallwood, J. & Schooler, J. W. The science of mind wandering: empirically navigating the stream of consciousness. *Annu. Rev. Psychol.* **66**, 487–518 (2015).
35. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
36. Mitchell, T. M. et al. Predicting human brain activity associated with the meaning of nouns. *Science* **320**, 1191–1195 (2008).
37. de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L. & Theunissen, F. E. The hierarchical cortical organization of human speech processing. *J. Neurosci.* **37**, 6539–6557 (2017).
38. Molenberghs, P., Johnson, H., Henry, J. D. & Mattingley, J. B. Understanding the minds of others: a neuroimaging meta-analysis. *Neurosci. Biobehav. Rev.* **65**, 276–291 (2016).
39. Sun, J., Wang, S., Zhang, J. & Zong, C. Neural encoding and decoding with distributed sentence representations. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 589–603 (2020).
40. Schrimpf, M. et al. The neural architecture of language: integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci. USA* **118**, e2105646118 (2021).
41. Caucheteux, C., Gramfort, A. & King, J.-R. Deep language algorithms predict semantic comprehension from brain activity. *Sci. Rep.* **12**, 16327 (2022).

42. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **24**, 2–249 (2008).
43. Renoult, L., Davidson, P. S., Palombo, D. J., Moscovitch, M. & Levine, B. Personal semantics: at the crossroads of semantic and episodic memory. *Trends Cogn. Sci.* **16**, 550–558 (2012).
44. Grilli, M. D. & Verfaellie, M. Personal semantic memory: insights from neuropsychological research on amnesia. *Neuropsychologia* **61**, 56–64 (2014).
45. Renoult, L., Irish, M., Moscovitch, M. & Rugg, M. D. From knowing to remembering: the semantic–episodic distinction. *Trends Cogn. Sci.* **23**, 1041–1057 (2019).
46. Irish, M. & Vatansever, D. Rethinking the episodic-semantic distinction from a gradient perspective. *Curr. Opin. Behav. Sci.* **32**, 43–49 (2020).
47. Gilmore, A. W. et al. Dynamic content reactivation supports naturalistic autobiographical recall in humans. *J. Neurosci.* **41**, 153–166 (2021).
48. Zhang, M. et al. Perceptual coupling and decoupling of the default mode network during mind-wandering and reading. *Elife* **11**, e74011 (2022).
49. Shao, X. et al. Distinctive and complementary roles of default mode network subsystems in semantic cognition. *J. Neurosci*. **44**, e1907232024 (2024).
50. Fernandino, L. & Binder, J. R. How does the “default mode” network contribute to semantic cognition? *Brain Lang.* **252**, 105405 (2024).
51. Oldfield, R. C. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* **9**, 97–113 (1971).
52. Glasgow, K., Roos, M., Haufler, A., Chevillet, M. & Wolmetz, M. Evaluating semantic models with word-sentence relatedness. Preprint at https://doi.org/10.48550/arXiv.1603.07253 (2016).
53. Cox, R. W. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**, 162–173 (1996).
54. Talairach, J. & Tournoux, P. *Co-planar stereotaxic atlas of the human brain*. in: 3-*Dimensional Proportional System: An Approach to Cerebral Imaging,* pp. 122 (Thieme, 1988).
55. Yeo, B. T. et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 1125 (2011).
56. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665–670 (2011).
57. Vos de Wael, R. et al. BrainSpace: a toolbox for the analysis of macroscale gradients in neuroimaging and connectomics datasets. *Commun. Biol.* **3**, 103 (2020).

Acknowledgements

This research was part funded by the Advancing a Healthier Wisconsin Endowment, under Momentum Award #5520801 (AJA) and the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory under grant FA8650-14-C-7357 (JRB). Thanks to Vankee Lin, Colin Humphries, Kelsey McDermott, and Kaylin Cottone for coordination and original data acquisition.

Author contributions

A.J.A. conceived the project, designed the research, performed analyses and wrote the manuscript. A.J.A., L.F., W.L.G., H.G.R., and J.R.B. discussed the results. A.J.A., L.F., W.L.G., H.G.R., and J.R.B. edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-65541-1.

Correspondence and requests for materials should be addressed to Andrew J. Anderson.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.