Narrative 'twist' shifts within-individual neural representations of dissociable story features

Clara Sava-Segal[1], Clare Grall[1], Emily S. Finn[1]

[1] Department of Psychological and Brain Sciences,
Dartmouth College,
Hanover, NH 03755, USA

Corresponding authors:

Clara Sava-Segal
6207 Moore Hall
Department of Psychological and Brain Sciences,
Dartmouth College
Hanover, NH 03755, USA
Clara.A.Sava-Segal.gr@dartmouth.edu

Emily S. Finn
6207 Moore Hall
Department of Psychological and Brain Sciences,
Dartmouth College
Hanover, NH 03755, USA
Emily.S.Finn@dartmouth.edu

Abstract

Given the same external input, one's understanding of that input can differ based on internal contextual knowledge. Where and how does the brain represent latent belief frameworks that interact with incoming sensory information to shape subjective interpretations? In this study, participants listened to the same auditory narrative twice, with a plot twist in the middle that dramatically shifted their interpretations of the story. Using a robust within-subject whole-brain approach, we leveraged shifts in neural activity between the two listens to identify where latent interpretations are represented in the brain. We considered the narrative in terms of its hierarchical structure, examining how global situation models and their subcomponents–namely, episodes and characters–are represented, finding that they rely on partially distinct sets of brain regions. Results suggest that our brains represent narratives hierarchically, with individual narrative elements being distinct and dynamically updated as a part of changing interpretations of incoming information.

Introduction

Identical sensory inputs can evoke different interpretations. Rather than being fully predictable from properties of the information itself (i.e., "stimulus-computable"), our experiences of external information are flexibly shaped by how that information interacts with our internal expectations, prior knowledge, and mental state.

Perceptual malleability has been widely studied using diverse types of stimuli[1]. For instance, bistable illusions induce rapid and reversible shifts between different valid percepts (e.g. a face versus a vase in the Rubin face-vase illusion). Beyond lower-level perceptual phenomena, more complex stimuli such as conversations or narratives can also exhibit ambiguity, often leading to more consequential and "stickier"—i.e., less reversible—interpretive shifts. (For example, once you realize that the character you believed to be the villain is actually the hero, you fundamentally alter your understanding of their actions and motivations throughout a story [2]). While there has been extensive behavioral and neuroimaging work studying these perceptual experiences, it is unclear how the *subjective* interpretations of ambiguous information are instantiated within individual brains.

Differences in brain activity *across* participants to the same external input (e.g. movies, auditory narratives, animations) have frequently been used to index different internal experiences of that input. In these studies, variability in neural activity across participants is often attributed to differences in interpretation. These differences in interpretation are shown to arise from various sources including experimentally-imposed or endogenously-generated contexts and beliefs [3–8], life-long experiences [9], and stable personality traits [10,11], which are assumed to color how people process identical sensory input. While across-subject analyses are informative, they can be confounded by idiosyncratic factors such as variations in functional brain anatomy, unmeasured traits and states, or experiential differences, making it challenging to fully attribute differences in brain activity to differences in interpretation. Instead, a within-subject approach, which compares the same individual to themselves before and after an interpretational shift, inherently controls for these factors and is a much more robust test for identifying where and how interpretations are represented in the brain.

Beyond licensing stronger inferences about the neural basis of interpretations, adopting a within-individual approach enables us to address two additional gaps in prior research. First, most neuroimaging studies treat narratives—and their corresponding interpretations—as monolithic entities. However, narratives are hierarchical, consisting of nested subcomponents such as characters and episodes [12,13]. These subcomponents form dynamic, bidirectional relationships with the overarching mental framework that tracks the 'gist' of a narrative, or global "situation model"[14,15]: global situation models provide a scaffold for interpreting and organizing narrative subcomponents, while updates to these subcomponents (e.g., changes in character motivations or new events) reshape the broader situation model. Forming, maintaining, and updating these hierarchical representations of narratives is likely to engage brain regions spanning the cortical hierarchy (from sensory to transmodal areas [16,17]). Thus, a second, related limitation of past empirical work is that it has tended to focus on a handful of *a priori* chosen networks or brain regions, such as the default mode network (DMN) [13] and the medial prefrontal cortex (mPFC) [18], which may have obscured important effects elsewhere in the brain.

Here, we aimed to identify where the brain represents distinct aspects of narrative interpretations. We used a unique narrative stimulus that contained a major twist halfway through that prompted participants to substantially shift their interpretations of the events preceding the twist. Participants then listened to the narrative a second time with this updated interpretation. Importantly, we held both the participant and the stimulus constant, enabling us to leverage within-subject shifts in neural activity between the first and second listen to understand how and where latent interpretive frameworks, independent of external sensory input, are reflected in the brain. Furthermore, by taking a whole-brain approach, we found evidence that elements at different levels of the interpretation hierarchy—i.e., global situation models, episodes, and characters—are represented in dissociable sets of brain regions. This work highlights how latent subjective interpretations of narratives are instantiated in the brain hierarchically.

Results

Our overarching goal was to identify where and how the brain represents interpretations of narratives and their subcomponents using a robust within-subjects approach. Thirty-six healthy adults listened to an auditory narrative twice in a row during functional magnetic resonance imaging (fMRI) scanning. The narrative featured a twist in the middle that recontextualized the earlier segments of the story. Initially perceived as a straightforward dialogue between a curmudgeonly dress-shopper (Steve) and a friendly, if pushy, shopkeeper (Lucy), the story later reveals a radically different reality: Steve is struggling to survive an apocalypse, and Lucy is a robot undermining his survival (See Methods section *Stimulus Description* for further detail).

The dramatic shift induced by the twist required listeners to update their global situation model, reevaluate specific episodes, and reassess the characters in light of the new context. We captured sets of within-subject "*shifts*" – defined as between-listen changes in neural representations – that reflect updates to each narrative element to identify where each is represented in the brain.

Representations of global situation models.

We first investigated where global situation models are represented. To this end, we compared within-subject neural and behavioral responses between the two listens. Behavioral responses were derived from the continuous rating task participants did while listening to the narrative in which they were tasked with rating one of the characters (Lucy).

Given the twist in the middle, we split the narrative into three segments: pre-twist, twist, and post-twist. The twist changes the interpretation of everything that came before it, prompting participants to shift to a new global situation model that persists during the remainder of the first listen and throughout the second listen. As a result, the pre-twist segment, which is interpreted under a different situation model in the first (L1) versus second listen (L2), should be processed most differently between listens. In turn, the post-twist segment, which is processed with the same situation model across both listens, should exhibit more consistent neural and behavioral patterns (**Fig. 1A**). To identify where global situation models are represented, we therefore compared

within-subject neural and behavioral shifts in each segment between listens (pre-twist$_{L1\text{-}L2}$ to post-twist$_{L1\text{-}L2}$), expecting greater shifts in the pre-twist segment than in the post-twist segment in both patterns of neural activity and behavioral ratings as a result of reinterpretation.

Greater behavioral shifts in the pre-twist segment.

During both listens, participants reported their real-time impression (negative to positive) of the shopkeeper ("Lucy") as part of a continuous rater task. Within the behavioral data, impressions of Lucy generally moved from positive to negative across the story, reflecting the evolving understandings of the broader situation and indicating a transition from viewing Lucy as a store clerk dealing with a difficult customer to perceiving her as a robot with Steve struggling to survive.

As hypothesized, behavior shifted more between listens in the pre-twist segment compared to the post-twist segment, indicating a greater change in how participants perceived the situation ("behavioral shift"; paired t-test, $t(34)=5.31$, $p < 0.001$; **Fig. 1C**). Behavioral shifts were calculated as one minus the intra-subject correlation between each participant's behavioral timeseries from the continuous rater task for Listen 1 and Listen 2.

Greater neural shifts in the pre-twist segment.

We operationalized neural representations as the multivoxel pattern of activity in each region at each timepoint. At each matched timepoint in Listen and Listen 2, we computed the within-subject correlation of these patterns ("pattern intra-subject correlation" (pattern intra-SC) [27,28]) and calculated "neural shifts" as one minus this correlation (henceforth "intra-subject pattern distance"). As hypothesized (**Fig. 1A**), the intra-subject pattern distance was higher in the pre-twist segment than in the post-twist segment across the cortex, indicating greater neural shifts in response to the updated information (main effect of segment: estimate = 0.01, $p < 0.001$; whole-brain linear mixed effects model (LMEM) with region and participant as random effects). The regions that showed the strongest differences, suggesting a strong role in maintaining and updating global situation models, included the left hippocampus, the angular gyrus, temporal parietal junction (TPJ), dorsomedial prefrontal cortex (dmPFC), and the bilateral posterior medial cortex (PMC)/precuneus (one LMEM per region with participant as a random effect; **Fig. 1D**). These regions align with findings from across-subject studies on contextual modulation and representations of situations and schemas [40–43] and studies of interpretational shifts during auditory narrative processing [6,44].

Notably, we did not see significant neural shifts between listens in primary auditory cortex. This was expected given that the low-level sensory properties of the stimulus are identical across listens and also helps to mitigate concerns that participants may have simply been paying less attention during the second listen. Some effects, albeit weaker than those in multimodal association regions, were also seen in early and middle visual regions (e.g., V1, MT). These effects are likely due, at least in part, to differences in how participants were looking at the screen to report or consider reporting the changes in the continuous rating task; slider movement overall varied more in each listen in the pre-twist segment compared to the post-twist segment ($t(34) > 5.63$, $p < 0.001$ for both listens).

We ran a series of control analyses to ensure the robustness of our findings. We first aimed to rule out the possibility that differences in brain activity were driven by participants' movements on the continuous rater task. To address this, we regressed the movement of the slider from each participant's neural timeseries in each listen, repeated the analyses on the residuals of this regression, and found that the results were largely unchanged (**Supplementary Fig. 1A**).

Next, we sought to dissociate the effects of our stimulus' specialized situation model structure from the effects of simply re-listening to the same information. First, one may expect that given that participants have already heard this narrative once, they may become less interested on the second listen. However, both our hypothesis and our observed results work against expected attention or "boredom" effects: if participants were simply mind-wandering more as time went on during the second listen, we would have expected to see greater shifts post-twist compared to pre-twist due to decreased engagement and more off-task (as opposed to stimulus-driven) activity. A second possibility is that regardless of any situation-model updating, participants simply become more synchronized to themselves over time when relistening to the same narrative, which could also explain our pre- versus post-twist differences. To help rule out this explanation, we turned to an independent dataset [33] where the same participants listened to an auditory story (from *The Moth*) multiple times, of which we used the first two listens (**Supplementary Fig. 1B**). Critically, this story did *not* contain a twist or any other feature that would induce a global situation model update akin to our stimulus. Encouragingly, we found (at a liberal, uncorrected threshold of $p < 0.05$) that only three regions showed a linear effect of time on pattern intra-SC: dorsolateral PFC and the bilateral auditory cortex. The latter region (auditory cortex) actually showed a decrease over time, potentially indicating reduced attention. Finally, in our dataset, we tested for any linear effects of time *within* segments (pre/post-twist) that could inflate our findings. By splitting each of the pre- and post-twist segments into an early and late period, we found that only a few inconsistent regions showed differences in pattern intra-SC between the early and late periods of each segment (again, at a liberal, uncorrected threshold of $p < 0.05$; **Supplementary Fig. 1C**). This combination of analyses further strengthens the likelihood that our observed pre- versus post-twist differences were, in fact, driven by the situation-model updates induced by the twist in our stimulus, rather than simpler phenomena inherent to listening to the same stimulus a second time more generally.

Regions show greater neural shifts when individuals report behavioral shifts.

Our first two analyses showed that, as hypothesized, both neural and behavioral shifts are greater in the pre-twist than the post-twist segment, likely reflecting global situation model updates that changed how this segment was interpreted overall. In a follow-up analysis, we sought to detect a more fine-grained, parametric relationship between these two types of shifts. In other words, moment-to-moment, do greater neural shifts track with greater behavioral shifts? Towards this goal, we first binarized individual participants' behavioral timeseries into timepoints where a shift was present or absent, corresponding to differences in their behavioral rating between their first and second listen ("behavioral shift"). We then compared neural shifts (intra-subject pattern distances) between these two sets of timepoints (see Methods section *Linking moments of neural and behavioral shift*).

Participants generally differed in how faithfully they complied with this behavioral task (see *Methods* for more information), limiting our power for this analysis. Although no regions

withstood FDR correction, those that showed the strongest effects were the bilateral precuneus/PMC, right hemisphere TPJ, and the right medial PFC ($p < 0.05$), dovetailing with past work that implicates these regions in active contextual updating (**Fig. 1D**). Furthermore, regions that showed an effect in this analysis also showed stronger effects in the global situation model segment analysis ($r = 0.27$, $p < 0.01$; correlation of estimates across regions between analyses). This suggests that, as hypothesized, the regions showing greater situation model updating (pre- versus post-twist contrast) were also involved in tracking the changed perceptions of Lucy throughout the stimulus (compare **Fig. 1D** and **1E**). Importantly, early visual regions did not show effects despite the task-induced eye movements towards the slider, indicating that we are likely capturing higher-level cognitive mechanisms (e.g., model updating) that operate at a more abstract level than simple visual or motor behavior.

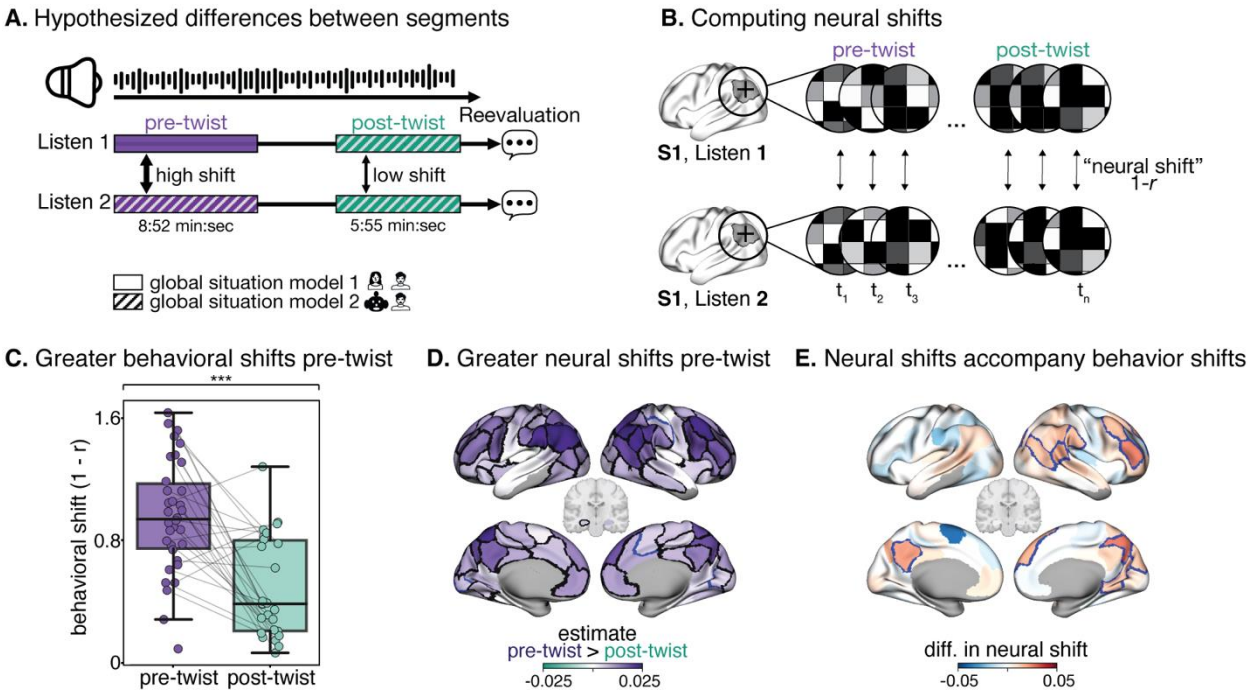

Figure 1. Neural and behavioral shifts reflect global situation model updating. **A. Hypothesized differences between segments.** The same individuals listened to an auditory narrative two times. The narrative was divided into three segments: 1. pre-twist, 2. twist and 3. post-twist. Novel, recontextualizing information is learned during the 'twist' segment, inducing a shift in interpretation. This new interpretation (global situation model 2) is carried into the post-twist segment on the first listen and into the entirety of the second listen. Thus, greater neural shifts are expected in the pre-twist as compared to the post-twist segment. After each listen, participants were tasked with reporting the specific moments (episodes) that they reevaluated in light of the twist (see Fig. 2). **B. Computing neural shifts.** For each participant, neural shifts between listens were computed per region per timepoint as one minus the correlation between the multivoxel spatial patterns of activity in Listen 1 and Listen 2 (pattern intra-SC). **C. Greater behavioral shifts in the pre-twist segment.** Within-subject shifts between listens in behavioral ratings (character rating) were greater in the pre-twist compared to the post-twist segment (paired t-test, *** indicating $p < 0.001$). **D. Greater neural shifts in the pre-twist segment.** The median pre-twist and post-twist neural shift value was taken for each participant and compared using a linear mixed effects model per region. Estimates plotted reflect the difference between the pre- and post-twist segments (set up as pre-twist > post-twist). Regions contoured in black show an FDR-corrected significant effect at $q_{FDR} < 0.05$ for all matched-length sample comparisons between segments (see *Methods*). Regions contoured in blue show an effect at $p < 0.05$ (uncorrected) for all matched-length sample comparisons. **E. Greater neural shifts accompany behavioral shifts.** For each participant, we binarized timepoints into those with a behavioral shift (absolute difference in ratings between Listen 1 and Listen 2 > 0) and those without a behavioral shift (absolute difference in ratings equal to 0), then compared neural shifts between these two groups of timepoints. The observed median difference in neural shifts (behavioral shift "present" moments minus behavioral shift "absent" moments) across participants is plotted. Blue contours indicate regions showing significant relationships between neural and behavioral shifts (thresholded at $p < 0.05$, uncorrected, determined via block permutations, n = 10,000).

Representations of episodes.

Having detected evidence for representations of a coarse-level global situation model in certain brain regions, we next investigated if and where the brain represents interpretations of smaller units of a narrative, namely specific episodes [45]. Here we defined *episodes* as punctate events with a clear beginning, middle, and end, that drove the plot forward.

After each listen, we prompted participants to identify specific moments that they reevaluated or reinterpreted in light of the twist (see Methods section *Experimental Procedures* for more information). All episodes occurred within the pre-twist segment, aligning with the neural and behavioral results that suggested greater interpretation updating during this segment (**Fig. 1**).

We hypothesized that, over and above the generally greater neural shifts in the pre-twist relative to post-twist segment, neural shifts would be even more exaggerated *specifically* during the episodes that participants reported reevaluating between listens. To test this hypothesis, we selected five episodes that were reevaluated by the majority of participants and chose five control episodes of matched length that were also in the pre-twist segment that most participants did not report reevaluating (shown in **Fig. 2A**). Then, for each participant, we modeled each individual episode using an event-related general linear model (GLM) and used the extracted episode-wise betas to compute a neural shift (intra-subject pattern distance between listens, **Fig. 2B**; see Methods section *Computing shifts in the reevaluated episodes* for more information).

By comparing neural shifts between reevaluated and control episodes, we found evidence that interpretations of episodes are represented along the bilateral superior temporal lobes, in the left TPJ, and, at an uncorrected threshold, in the left superior frontal cortex (**Fig. 2C**; LMEM per region predicting the difference in the neural shift between reevaluated and control episodes, treating both participant and episode pair (reevaluated, matched control) as random effects). These findings replicate a related study that identified the left anterior middle temporal gyrus as among regions supporting 'aha' moments [46]. Furthermore, the observed left lateralization aligns with previous findings that functional representations of semantics and social cognition are predominantly left-lateralized [47–51].

Compared to the global situation model, interpretations of specific episodes appeared to be represented in distinct, more lateralized temporal regions (compare **Fig. 1D** to **Fig. 2C**). The one region that showed strong effects in both episode and situation model representations was the left TPJ; this finding is unsurprising considering this region's involvement in binding of external information and managing competing beliefs [52,53].

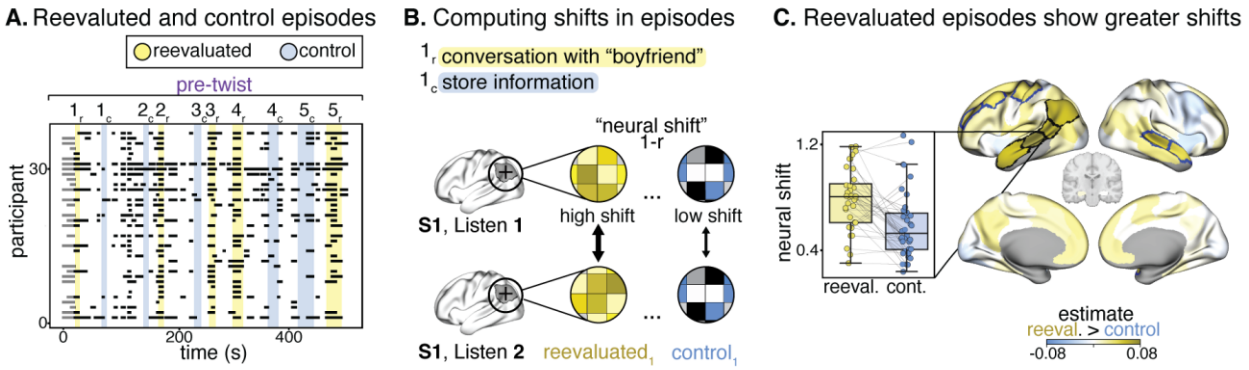

Figure 2. Episodes that are reevaluated show greater neural shifts between listens. **A. Identifying reevaluated episodes and matched controls.** Using behavioral data provided by the participants inside and outside of the scanner, we selected the top five most commonly reevaluated episodes and paired each one with a matched control episode that was nearby in the narrative and the same length, but not reported as reevaluated by most participants. All reevaluated and control episodes were within the pre-twist segment. We plot the temporal location and duration of each episode plus a raster-style depiction of participants' behavioral reports (black lines correspond to moments that participants reported reevaluating; gray lines correspond to the first TRs which were removed to avoid transience effects). The five reevaluated and control episode pairs are highlighted and labeled. Timepoints in gray were not included. **B. Computing neural shifts between the reevaluated and control episodes.** For each participant, we used an event-related GLM to model each individual episode in each listen, then computed neural shifts as one minus the correlation between the spatial pattern of beta values in Listen 1 and Listen 2 (one value per region per episode). Greater neural shifts were hypothesized for the reevaluated, as opposed to the control, episodes. **C. Reevaluated episodes show greater neural shifts within individuals.** Plotted estimates show the strength of the difference between reevaluated and null episodes within participants. (Estimates reflect output from a linear mixed effects model in which within-subject neural shifts were predicted by episode type (set up as reevaluated > control), using participant and episode pair as a random effect.) Regions contoured in black show an FDR-corrected significant effect at $q_{FDR} < 0.05$. Blue contours reflect a relationship thresholded at $p < 0.05$ (uncorrected). The across-subject distribution of median neural shifts within the superior temporal sulcus are plotted in the inset. Dots represent participants' median neural shifts across episodes within each listen.

Representations of Characters.

Having demonstrated that global situation models and models of specific episodes are represented in largely distinct brain regions, we next examined how information about characters was represented and updated across the narrative. Characters link episodic details to the global situation model by embodying the motivations and goals that influence the narrative's progression and, especially in this stimulus, undergo major reinterpretation (see Methods section *Stimulus Description*). To study character representations, we investigated how representations of Lucy (from shopkeeper to robot) are constructed and updated across the two listens. We focused specifically on Lucy because the updates to her character are larger and better motivated than those for Steve; she starts and ends the story and undergoes a much greater identity shift.

We expected that by the end of Listen 1, participants would have converged on a final interpretation of Lucy (as a robot) and that they would then "reload" this interpretation at the start of Listen 2. We operationalized these assumptions into predictions about what neural activity patterns should look like in regions tracking latent representations of the character.

To this end, we split the narrative into "Lucy" or "Steve" conversational turns based on speaking onset and offset times and modeled each turn in each listen using an event-related GLM. For each

participant, we designated a "template" neural representation of Lucy from her final speech event (conversational turn) in Listen 1, when participants had all the information necessary to fully interpret (represent) her identity. We then correlated representations of Lucy at each turn in Listen 1 and Listen 2 with this template, yielding a series of intra-subject character event-template correlations per region (**Fig. 3A**). The magnitude and change over time of these correlations were compared between listens. We considered a region as representing Lucy if it showed the following properties: (1) a steady increase over time in similarity between Lucy events and the template over Listen 1, (2) a reloaded template-like representation at the start of Listen 2, (3) a stabilization in representation (i.e., flatter slope toward the template) over the course of Listen 2, and (4) a dissociable representation of Steve (i.e., lower or negative correlations with the template that stay flat or decrease over time) in both listens (**Fig. 3B "Criteria"**). For more information on these criteria and how they were tested, see Methods section *"Computing updates in the representations of characters"*.

Results indicated that much of the brain showed effects consistent with the hypothesized directions across our four criteria. A subset of regions exhibited a significantly steeper slope over time (increasing event to template correlations) in Listen 1 relative to Listen 2 ($q_{FDR} < 0.05$; **Fig. 3C**, all contours), which we considered the most important index for a region representing latent character interpretations. This effect emerged in the bilateral mPFC and right anterior temporal pole, as well as in the precuneus and rostral posterior cingulate regions that flank the posteromedial DMN-associated areas (parietal memory network [54,55]). Some of these regions have previously been implicated in investigations of characters in narratives [18,42,56,57].

Taken together with the previous section, these results show that episode and character representations, while both contributing to the formation of the global situation models, rely on distinct brain regions. Unlike left-lateralized episode representations, character representations show stronger right-lateralized involvement (see: right angular gyrus, TPJ, anterior temporal pole).

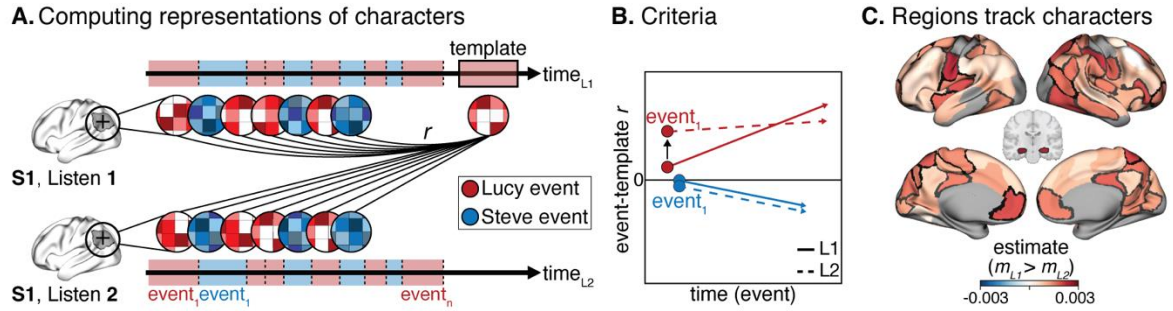

Figure 3. Character representations are updated on the second listen. **A. Computing representations of characters.** The dialogue was split into 'Lucy' and 'Steve' events based on speaking onset and offset times. We designated a per-participant 'template' representation of Lucy based on her last speech event in Listen 1 (L1). Each event was correlated with the template to test for a series of criteria (see panel B). **B. Schematic of criteria.** Correlations between Lucy events and the template were hypothesized to be positive and to increase progressively over the course of the story (positive slope). In Listen 2, they were expected to start higher and exhibit a weaker slope compared to Listen 1, reflecting the "loading" of Lucy's representation from the end of Listen 1. Correlations between Steve events and the template were hypothesized to be non-existent or, if anything, to show a negative slope over time (as representations of the characters diverged). **C. Regions track character representations.** Estimates reflect the magnitude of the effect for our main criterion, which was that the similarity between Lucy events and the final

template representation of Lucy should show a steeper slope over time in Listen 1 than Listen 2 (computed with a linear mixed effects model predicting event-template correlations from an interaction between listen and event number with a random effect of participant). Regions plotted meet all of our criteria (see B; *Methods*). All contoured regions show an FDR-corrected significant effect at $q_{FDR} < 0.05$. Gray contours indicate regions that show an effect at $q_{FDR} < 0.05$ for the main criterion. Black contours additionally indicate regions that an effect for all Criteria (1-3) at $q_{FDR} < 0.05$.

Dissociable neural substrates for representing distinct elements of narrative interpretation.

Results thus far revealed that neural representations of different narrative elements involve partially overlapping yet somewhat distinct sets of brain regions, consistent with the hierarchical organization of these elements. This can be appreciated visually by comparing the maps for global situation models, episodes, and characters (compare **Fig. 1D** to **2C** to **3C**). To quantify this dissociation, we first assessed the degree of overlap in the regions involved in each analysis by correlating effect estimates across regions. Regions representing the global situation model were distinct from those representing either episodes ($r = - 0.04$, $p > 0.05$) or characters ($r = - 0.13$; $p > 0.05$); these two subcomponents (characters and episodes) were also distinct from one another ($r = - 0.20$, $p = 0.07$; **Fig. 4A**).

To further explore these dissociations, we applied KMeans clustering to group regions based on their distribution (pattern) of estimates from each analysis, which identified four informative clusters (**Fig. 4B**). We highlight two key outcomes from our clustering results. First, while the default mode network (DMN), broadly defined, was involved in representing all three narrative elements, its different sub-regions and sub-networks have distinct and variable contributions for representing these elements. Second, the clustering solution that emerged reinforces our hypothesized hierarchical framework in that global situation models were represented to some extent in every cluster, sometimes along with either episodes or characters, but not both.

To elaborate, while Clusters 1 and 2 were involved in representing all three narrative elements, Cluster 1 was largely representative of global situation models and Cluster 2 was largely representative of character-level information. Interestingly, these two clusters both comprise parts of the DMN; the "core" regions (bilateral AG, TPJ, and precuneus/PMC) in Cluster 1 and the bilateral mPFC, bilateral hippocampus, and right temporal pole in Cluster 2. The former regions have been previously reported to facilitate broad "interpretation" updating across individuals, while the latter have been associated with maintaining representations of schemas, identities, and mental simulations [58,59]. In turn, Clusters 3 and 4 exhibited greater specificity. Both clusters were involved in representing global situation models, but each cluster was paired with a distinct narrative subcomponent: Cluster 3 with episode representations and Cluster 4 with character representations. Together, these analyses highlight how distinct sets of brain regions are differentially engaged to support the hierarchical structure of narrative elements.

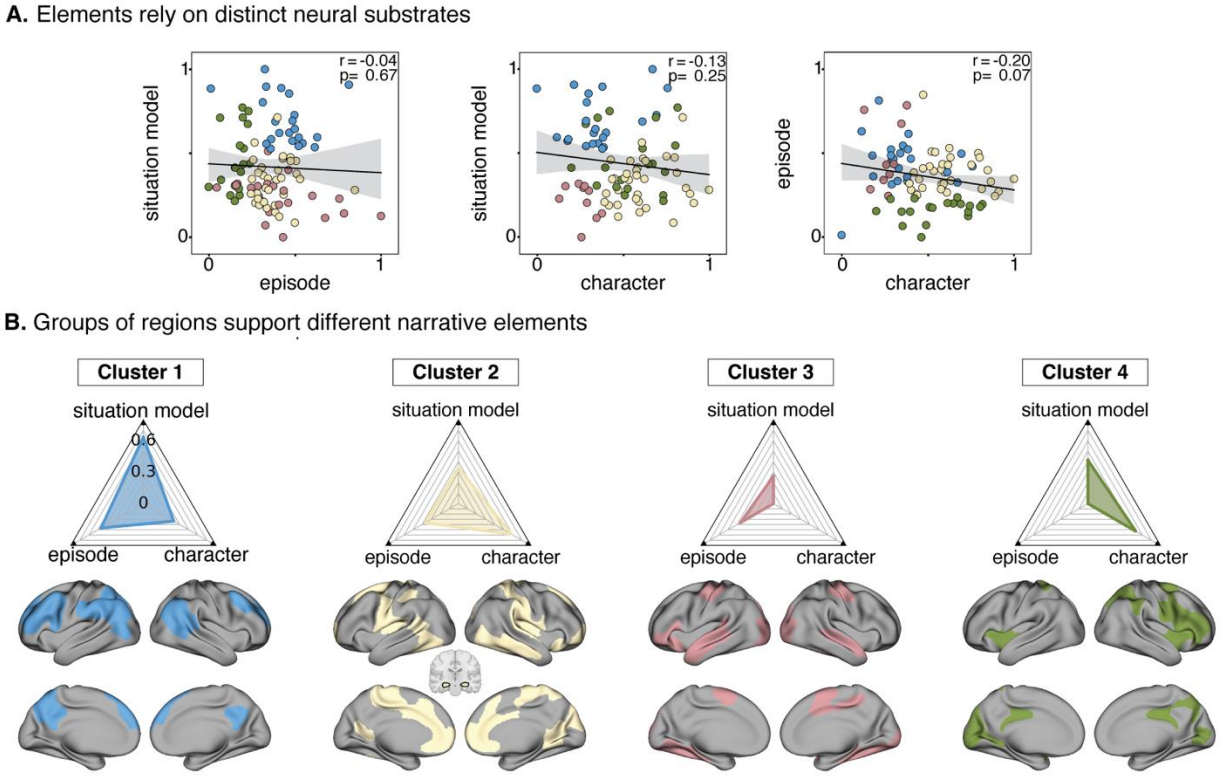

Figure 4. Representations of narrative elements are differentially represented and coupled within distinct sets of brain regions. **A. Narrative elements are represented in distinct neural substrates.** Correlations between normalized estimates in each region across analyses show that representations of global situation models, episodes and characters rely on distinct neural substrates. Each dot indicates a region. Coloring of a region is based on the assigned cluster (see B). **B. Groups of regions support different narrative elements.** We clustered regions according to their relative involvement in representing the three narrative elements: global situation model (as indexed by the pre- versus post-twist analysis), specific episodes, and characters. A solution of $k = 4$ clusters was found. Clusters 1 and 2 showed relatively high involvement in representing all three narrative elements, while Clusters 3 and 4 exhibited a paired coupling of the global situation model with each of the other two sub-components (episodes and characters, respectively).

Discussion

In this study, we investigated where and how the brain supports latent belief representations of distinct narrative elements. To do so, we deliberately selected an auditory narrative that featured a mid-story 'twist' or shift in the ground truth that fundamentally altered participants' understanding of earlier events. Participants listened to the stimulus twice over, carrying forward the global situation model formed after the twist into the second listen. This within-subject design enabled us to directly compare each participant to themselves as they updated their interpretations and understand how this interpretational shift altered the representation of the same sensory input. We decomposed large-scale representations of the narrative into three hierarchical subcomponents (global situation models, episodes, characters) and detected multivariate representations of these

components that might otherwise be obscured or confounded at the across-subject level. We found that global situation models exhibited the most widespread representations across the brain. In turn, episodes and characters relied on partially overlapping regions with each other and with global situation models, yet each also engaged distinct cortical regions, suggesting a degree of specialization in neural roles for representing and integrating different narrative elements.

While narratives are increasingly used to study the neural integration of information over time [12,60], researchers have paid limited attention to how subcomponents of narratives are instantiated within underlying neural representations. Many studies inherently assume that narratives are represented as a unified whole. However, behavioral evidence from prior work [61] suggests that different narrative elements can be updated independently, proffering the possibility that distinct neural systems may underlie the representation of specific narrative elements. We provide evidence for this idea: the default mode network (DMN) supports narrative representations broadly, but there are notable distinctions across transmodal cortex and the hippocampus in the degree to which specific regions are involved in representing these different narrative elements.

Much prior work has focused on across-subject differences in activity within the DMN during narrative processing [6,7,40,62]. For instance, Zadbood *et al.,* (2022) used an across-subjects design and a movie with a plot twist to demonstrate that representations in specific core subregions of the DMN (e.g., TPJ, mPFC, temporal poles) varied based on participants' prior knowledge of the twist *and* were updated with the new information gained via the twist. Our findings align with and extend these prior across-subject studies. By localizing within-individual representations of distinct narrative elements, we provide greater specificity to how narratives are represented within the DMN. We showed that different subnetworks and subregions of this larger network have greater contributions to some narrative elements, as compared to others. While changes in activity within the core cortical DMN regions track within-individual global situation model updating and correlate with behavioral shifts, lateral temporal regions, such as the STS and temporal pole, appear to support more focal representations for episodes and characters, respectively. Even among the core DMN regions, there are some distinctions – for example, the mPFC represents global situation models and characters, but not episodes. Taken together, our findings add to the longstanding evidence that the DMN comprises multiple, interacting subsystems with distinct functions [63–66].

These topographic distinctions may, in part, reflect differences in not only what types of information these transmodal regions are sensitive to, but also their temporal windows of information processing and integration [67]. There is extensive evidence that hierarchical processing architectures may be effective for processing stimuli that possess a hierarchical structure (see [68] for a review); regions earlier in the cortical hierarchy (namely, primary sensory and sensory association regions) are sensitive to fast fluctuations in the input stream, while higher order transmodal regions are sensitive to slower fluctuations, changing only in response to longer windows of prior stimulus context. While temporal receptive windows have been commonly studied with language (comparing words to sentences to paragraphs), these windows likely support the processing of latent, time-varying narrative features. Global situation models and character representations operate over longer timescales than episodes; situation models sit at the apex, episodes serve as the fundamental building blocks, and characters function as dynamic agents driving transitions between episodes. We find that regions with shorter receptive windows track

faster, more time-bound fundamental narrative units (e.g., left STS uniquely represents episodes) compared to regions with longer intrinsic timescales that process slower dynamics (e.g., the dorsolateral prefrontal cortex uniquely represents the global situation; see [13,69] and [70] for a review). Future work should directly manipulate the timescales at which these features operate and interact to investigate this more systematically.

Despite these dissociations in representation, the lateral posterior parietal and temporal cortex (regions including and around the TPJ) represented all three narrative elements regardless of their position in the narrative hierarchy (**Fig. 4B, Cluster 1**). Outside of its general association with the DMN, a recent proposal has termed this patch of cortex as "gestalt cortex," theorizing that it is specifically involved in supporting subjective experiences, or construals, by reconciling competing interpretations [71]. To our knowledge, we are providing the first within-subject evidence for "gestalt cortex," highlighting that representational shifts within these regions reflect internal updating of construals.

There are several limitations to this work. First, our analyses rely on a single stimulus. Although this stimulus was carefully chosen for our study design, we acknowledge that some observed effects could be driven in part by idiosyncratic properties of this particular stimulus rather than more general features of narrative interpretation [72]. We benefited from focusing solely on one, relatively long stimulus in the auditory domain, but future work may consider employing carefully crafted multisensory stimuli to broaden generalizability. Second, relatedly, our study specifically focuses on two mid-level subcomponents in narrative representation—episode and character representations—that were well-suited to our stimulus. Future research could explore more fine-grained features, such as distinctions between main and secondary characters or hierarchical (nested) episode and event structures. Third, participants were quite variable in their behaviors during the study, specifically how often they used the slider to report their character impressions as well as the number of and detail associated with episodes they reported reevaluating, which limited our ability to create individualized models of representations and how they were updated. Lastly, issues of MRI data quality interfered with our ability to investigate subcortical regions. Future work should explore subcortical involvement in these processes, including regions such as the amygdala, which has been implicated in supporting episodic memories and narrative processing.

In sum, we capture how latent interpretive frameworks are instantiated in the brain, highlighting the advantages of a within-subject approach. By holding both the participant and the sensory input constant, we robustly identified shifts in patterns of neural activity induced by new context for the same information. These shifts reflect updated interpretations and situation models, as well as more nuanced representations of episodes and characters. This approach allowed us to better pinpoint where these distinct narrative elements are neurally represented, offering a clearer understanding of the hierarchical organization of narrative processing in the brain. Together, this work provides a foundation for understanding how exogenous input and endogenous belief frameworks interact to shape subjective experience.

Funding

Funding was provided by the National Institutes of Health (R00MH120257 to E.S.F.) and (1F31MH138084-01 to C.S.S.) and by the National Science Foundation Graduate Research Fellowship to C.S.S.

Acknowledgements

The authors thank Josefa M. Equita for assistance with data collection and behavioral preprocessing; Eneko Uruñuela for implementing denoising with Tedana; Katherine Bartolino and Evan Bloch for their contribution to behavioral data cleaning and processing; Rekha Varrier for early discussions on study design and data collection; Thomas L. Botch for preprocessing the *Moth Stories* data and providing thoughtful comments on the manuscript.

Code availability

Data analysis, including links to code and other supporting materials, can be found at: https://github.com/thefinnlab/darkend_narrative_rep.

Data availability

Data from this study, including raw MRI data, will be made available on OpenNeuro upon publication.

Methods

Stimulus description.
We used the "Dark End of the Mall" episode (18:25 min:sec) from the podcast *The Truth*, which consists of a dialogue between two characters, Lucy and Steve [19]. The non-speaking time is limited to moments of a dog barking and brief moments of a "song" playing in the background. We chose this stimulus due to the feasibility of working with only two characters and, more importantly, its unique narrative structure. Specifically, the narrative contained a twist in the middle that required participants to globally update their situation model of events that preceded the twist, creating three distinct and meaningful narrative segments (pre-twist, twist, and post-twist) for within-subject comparison. We provide a brief synopsis below.

> *The story starts off with a phone conversation between Lucy, a sweet but vapid bridal shop employee, and presumably her boyfriend (whom listeners do not hear) which gets interrupted by Steve running into the shop. Listeners initially perceive Steve as a cranky dress shopper who is abrasive toward Lucy as he tries multiple attempts to convince her that she should give him some of the food hidden in the shop. Lucy gets frustrated with Steve and calls mall security and tries to kick him out of the shop. Eventually, Steve asks Lucy if he can tell her a story. It is revealed via Steve's story that Lucy is, in fact, a robot programmed to work in a 1950's style bridal shop, that they are both living in an apocalypse in 2050, and that Steve is one of the last surviving humans and has figured out that bridal shops have hidden snacks that sustain his survival. He almost convinces Lucy to help him, but ultimately fails as she kicks him out of the shop where, presumably, he meets his death. Listeners last hear a distressed Steve confronting barking dogs and Lucy again on the phone with her boyfriend, but listeners now realize via the narrative that the dogs are likely zombies and the boyfriend is fictitious.*

Participants. All data was collected at the Dartmouth Brain Imaging Center. Participants (n=36; 24F, 12M, 1 non-binary; median age = 20, range = 18 to 33) were healthy individuals, with normal or corrected-to-normal vision and hearing and no recent psychiatric or neurological diagnoses or MRI contraindications. They were recruited from the local areas of New Hampshire and Vermont, including the Dartmouth College student body. The Committee for the Protection of Human Subjects of Dartmouth College approved the study, and all participants provided written consent.

Experimental procedures. All participants listened to the same auditory stimulus twice. At the beginning of the study, they were told that they *may* hear the auditory narrative a second time, but that they also may have the opportunity to hear a second story. No participant actually heard another story. While in the scanner, we used Sensimetrics Model S14 insert earphones to present the sound, and participants were given a trackpad (Cedrus Lumina) to continuously indicate their impressions of Lucy from very negative to very positive throughout each listen. They were given minimal visual input: the screen displayed throughout both listens showed a static photograph of a bridal shop (to promote imagery and engagement with the story) and, underneath that image, the continuous scale used to rate Lucy impressions.

Continuous Rater Task.
Participants were tasked with rating the character of Lucy by answering the question: "*Overall, how much do you like Lucy?*" During the presentation of the stimulus, participants used the trackpad to update their rating while the stimulus played along a scale from -3 to 3. We opted to do this task in real time in the scanner as opposed to in an independent dataset of non-fMRI participants because pilot participants showed considerable variability in their ratings. Furthermore, to emphasize the within-subject design of our study, we did not want to use other participants' data as a proxy for fMRI participants' ratings of the character.

Before the second listen, participants were instructed as follows: "*For your 2nd story, you have been assigned to listen to the same story again and complete the same prompt. For this 2nd listen of the same story, consider how your impression has changed. Because you have already listened to the story, we expect that your impressions of Lucy are different than your 1st listen. Given what you know about this story, what is your impression of Lucy now? Please use the slider to continuously rate your impression.*"

Tasks after each listen.
Each scanner run consisted of the entire narrative; after each listen, participants were asked to do a series of character rating questions and memory tests (maximum of 10 seconds each) and engage in a "reevaluation task" (**Fig. 1A**). For the character rating questions, participants were asked to report "Overall, how much do you like [Lucy/Steve]" as independent questions. For memory questions, participants were asked "*1. What is Lucy, 2. What does Lucy hear running throughout the story, 3. What is the name of the shop where this story takes place?"* after Listen 1 and "*1. What caused the destruction of humankind?, 2. What dish does Lucy recommend Steve buy at the food court?, 3. Who does Lucy think she is talking to at the beginning of the story?"* These questions were intended to be relatively challenging and therefore to serve as attention checks. Participants performed well on these questions (median score 100%; mean score = 93%). After completing these questions, participants then completed the reevaluation task after each listen. After Listen 1, they were instructed "*Using the microphone, please describe the moments at the beginning of the story that you reconsidered after hearing the end."* After Listen 2, they were instructed, "*please describe the moments of the story that changed for you after hearing the story once before.*" They had 60 seconds to answer using free speech.

Post-scan tasks.
Outside the scanner, participants were presented with the transcript and asked to "*highlight the 1-3 sentences that mark the moment in the story when the twist occurred."* They were also given all of the sentences in the pre-twist segment and tasked to indicate the ones that they reevaluated. Instructions stated "*highlight the sentences that mark the moments in the story that you reinterpreted when listening to it a second time*."

fMRI data processing.
MRI acquisition. All data were collected at Dartmouth College in a 3.0 Tesla Siemens MAGNETOM Prisma whole-body MRI system (Siemens Medical Solutions, Erlangen, Germany) equipped with a 64-channel head coil.

T1 image.
For registration purposes, a high-resolution T1-weighted magnetization-prepared rapid acquisition gradient echo (MPRAGE) imaging sequence was acquired (TR = 2,300 ms, echo time (TE) = 2.32 ms, inversion time = 933 ms, flip angle = 8°, field of view = 256 × 256 mm, slices = 255, voxel size = 3 × 3 × 3 mm isotropic). T1 images were segmented, and surfaces were generated using FreeSurfer [20].

fMRI acquisition.
fMRI data were acquired using a multi-echo T2*-weighted sequence. The sequence parameters were: TR = 1,000 ms, TEs = [14.2, 34.84, 55.48], GRAPPA factor = 4, flip angle = 60°, matrix size = 90 × 72, slices = 52, multiband factor = 4, voxel size = 3mm isotropic. To account for field stabilization and hemodynamic delay, an additional two TRs were added to the front of the stimulus and 10 TRs were added to the end.

Preprocessing.
Multi-echo data preprocessing was implemented in AFNI [21] using afni_proc.py for alignment, transformation, and optimization steps. Each participant's data was processed to align the anatomical (T1) image and functional images, with motion correction based on the second echo and alignment parameters applied to all echoes. Functional data underwent despiking (3dDespike) for outlier attenuation, followed by the concatenation and extraction of functional time series for each echo. The three echoes were then optimally combined and denoised using multi-echo ICA via *tedana* [22–24]. Signals were then normalized to percent signal change and spatially blurred (3dBlurInMask), with motion regressors applied to reduce artifacts in final volumes. Following preprocessing, to account for transitory changes at the start of the stimulus [4], we removed the first 18 TRs from the start of the stimulus for all of our subsequent analyses (also excluded in **Fig. 1A**, grayed out in **Fig. 2A**).

Defining regions of interest.
The Schaefer parcellation [25] was used to designate 100 cortical regions; five of these regions—around the ventral part of the brain—were removed because more than 50% of participants were missing more than 40% of the data in these regions. The Harvard-Oxford Atlas was used to identify the hippocampus in both the left and right hemispheres [26]. We were unable to include other subcortical regions, including the amygdala, due to data loss (almost 50% of participants (17/36) were missing signal in more than 40% of voxels). Parcel sizes ranged from 113 to 759 voxels. All results shown here were robust to parcellation granularity in that effects persisted when using a 400-region parcellation [25].

Computing the 'twist'.
We defined the "twist" in the story as moment(s) when participants transition from one interpretation/situation model to another—specifically, from believing the setting is a bridal shop to realizing it is a post-apocalyptic world. To capture this shift, participants were asked to identify the twist in a post-scan survey (see *Experimental Procedures* for more information on instructions). Participant responses varied considerably, with some selecting multiple points in the story. To address this variability, we adopted a conservative approach to identifying the twist, defining its start as the point before the earliest event chosen by the majority of participants, and its end as the point after the latest event chosen by the majority. This approach allowed us to split

the stimulus into pre-twist (length = 532 TRs) and post-twist segments (length = 355 TRs), plus a segment in the middle corresponding to the twist itself (length = 200 TRs). Pre- and post-twist segments were matched for length when appropriate (see Section Methods: *Computing neural shifts to assess global situation model representations*).

Computing behavioral shifts to assess situation model representations.
We compared "behavioral shifts" between the pre-twist and post-twist segments using the timeseries from each participant's continuous rating of Lucy acquired in both listens. See *Experimental Procedures* for specific instructions on how this continuous rater task was conducted. We quantified the dissimilarity ("behavioral shift") as one minus an intra-subject correlation (intra-SC) between the behavioral timeseries from Listen 1 and Listen 2 for each segment. We compared these within-subject intra-SC values between segments using a paired t-test.

Computing neural shifts to assess global situation model representations.
Our first goal was to quantify changes in the within-subject representation of the narrative ("neural shifts") between listens and to compare the magnitude of these changes between the pre-twist and post-twist segments. For each participant and region, we correlated the multivoxel spatial pattern at each timepoint between listens, yielding a pattern intra-subject correlation (intraSC) [27,28] for each timepoint. This was converted into a "neural shift" at each timepoint by subtracting the pattern intraSC from one (distance).

To test for a difference between segments, we computed the median neural shift value within each segment for each participant and conducted a linear mixed effects model (LMEM; using lme4 in R; [29]) where median neural shift was predicted by the segment (pre- or post-twist) it belonged to, using participant as a random effect. Note that taking the median neural shift from each segment, as opposed to using shifts from all timepoints, helps accounts for autocorrelation in the functional data. This model was run per region. The estimates from each of these LMEMs were plotted (**Fig. 1D**).

To ensure observed neural shifts were not driven by differences in length of the two segments (532 versus 355 TRs), we trimmed the pre-twist segment to match the length of the post-twist segment. Specifically, we generated all possible 355-TR subsets of the pre-twist segment by sequentially trimming the pre-twist data from the start, creating 178 distinct samples (532 - 355 + 1). For each sample, we ran an LMEM for each region to compare the pre-twist and post-twist segments. The p-values from these models were then corrected for multiple comparisons using false discovery rate (FDR) based on the number of regions in our analyses (97 total: 95 cortical and two hippocampi) using an alpha of 0.05. To be as conservative as possible, we only considered regions to be significant if they were $q_{FDR} < 0.05$ in all 178 matched-length samples (**Fig. 1D**).

Linking moments of neural and behavioral shift.
For each participant, we aimed to identify which neural regions track behavioral shifts in interpretation. To this end, we binarized timepoints into moments where behavioral shifts were present (i.e., absolute difference of behavioral rating between Listen 1 and Listen 2 >0) or absent (i.e., absolute difference = 0). We took this binary approach, rather than directly correlating the behavioral continuous response timeseries with the neural timecourse as has been done in other

studies [30,31], for two reasons. First, this approach better isolates specific moments where shifts in character impressions occur, allowing us to directly link these discrete behavioral shifts to changes in neural activity. Second, participants exhibited variability across listens in their use of the sliders both in the range of values used and in the frequency of movements (Wilcoxon signed-rank test, $p < 0.05$). Participants also differed amongst themselves, though not statistically significantly (std. within Listen 1: 20.5 ± 24; Listen 2: 16 ± 32.4; Kruskal-Wallis test, $p > 0.05$). This variability introduced potential confounds, limiting the validity of direct correlations between the neural and behavioral timeseries. For example, two participants did not move the sliders at all in the second listen (yielding an n=34 for this analysis altogether) and several moved them very infrequently, violating the basic assumptions for such correlations. By adopting a binary approach, we circumvented these issues and instead focused on the presence or absence of meaningful differences.

For each participant, we then compared the median neural shift between the timepoints when a behavioral shift was present or absent, using this (present minus absent) as our observed difference. To account for the hemodynamic delay, we shifted the behavioral timeseries by 4 TRs (4 seconds) relative to the neural data. To evaluate the statistical significance of these observed differences, we generated a null distribution by randomly shuffling blocks of time (of length 10 TRs [32]) of the behavioral data 10,000 times for each participant, effectively breaking any relationship between the neural data and behavioral labels. For each permutation, we recalculated the differences between the shuffled 'shift present' and 'shift absent' timepoints, to generate a distribution of differences that would be expected under the null hypothesis (i.e., H_0: no true relationship between the neural and behavioral data). The p-value was calculated as the proportion of null differences greater than or equal to the observed difference.

Control analyses: ruling out possible confounding effects of time on within-subject similarity. To further ensure that the observed results, which were consistent with our hypothesized directionality (higher similarity in the post-twist relative to pre-twist segment), were due primarily to shifts in interpretation rather than other explanations, we investigated the alternative hypothesis that individuals simply become more similar to themselves over time when processing the same long-timescale narrative. Critically, we tested this hypothesis both in an independent dataset as well as in our own dataset, as described below.

Computing within-subject similarity over time in an independent dataset.

We used fMRI data from an existing dataset [33] in which participants (N = 8) listened to the same *The Moth* story ("*Where There's Smoke*") multiple times. Importantly, this story lacked a twist or any other feature that might induce interpretational differences, making it suitable as a control. We used data from the first two times participants listened to the story (run 1 from session 2 and run 2 from session 3), performed functional alignment using hyperalignment [34] with a leave-one-session-out cross-validation procedure, and again parcellated the data using the Schaefer parcellation. Taking the same approach as in our main analyses, we then computed the pattern intra-SC at each timepoint. To assess whether within-subject similarity changed with time, for each region, we fit a linear model for each participant predicting pattern intra-SC as a function of timepoint (TR). We then evaluated statistical significance using a one-sample t-test (two-sided) for each region on the

resulting beta values across participants. We applied a liberal, uncorrected threshold of $p < 0.05$ to see which regions, if any, showed increased or decreased similarity over time.

Computing changes over time within segments of our narrative.

To further rule out possible confounding effects of time (rather than interpretation) on the similarity of within-subject neural representations, in our dataset, we tested for any linear effects of time *within* segments. Specifically, we further divided our pre-twist and post-twist segments into early and late periods, resulting in four distinct periods: pre-twist early, pre-twist late, post-twist early, and post-twist late. Within each of these four periods, we took the same approach: for each participant, we computed pattern intra-SC values between listens at each timepoint within each region. We then fit a linear model for each participant predicting pattern intra-SC as a function of timepoint (TR) in a given period. Next, we aggregated beta values from these models across participants and performed two two-sample t-tests within each region. Specifically, we compared beta values across participants for pre-twist early vs. pre-twist late and post-twist early vs. post-twist late. These tests assessed whether similarity was higher in the earlier versus later segments. We applied a liberal, uncorrected threshold of $p < 0.05$ to identify which regions, if any, showed an effect of increased (or decreased) similarity over time.

Computing shifts in the reevaluated episodes.

As discussed in further detail in *Experimental Procedures,* after each listen, participants verbally reported episodes – distinct events with a clear beginning, middle, and end that advanced the storyline – that they reevaluated. Then, outside the scanner, they manually highlighted the text to indicate these episodes. Our goal was to identify where and how these episodes are represented in the brain.

Given that participants varied in the number of episodes they chose, we selected five episodes consistently noted by the majority (at least 25/36; ~70%) of participants in both their in-scanner verbal reports and post-scanner written highlighting task. These episodes varied in duration (8, 9, 11, 18, and 27 TRs) and were manually checked by the experimenter to ensure that they included the entirety of an episode, i.e., if a participant chose only one of the two sentences that comprised an episode, we considered the entirety of the episode if the majority of other participants reported reevaluating all of it. Importantly, all identified episodes occurred within the pre-twist segment (**Fig. 2A**).

We then selected a set of "control episodes" to serve as a comparison point for the reevaluated episodes. To this end, we identified episodes within the pre-twist segment that the vast majority of participants (no more than 11/36; less than 30%) did not report reevaluating. We intentionally chose these episodes to be matched in length and nearby in time to the reevaluated episodes to account for any neural drift in the signal and to ensure that both the "control" and "reevaluated" episodes were within the same pre-twist segment (which had more overall reinterpretation, see **Fig. 1**). A brief description of these episodes is provided below.

> *The reevaluated episodes include the following: 1r. a conversation that Lucy has with an imaginary boyfriend; 2r. when Steve calls her a robot (reinterpreted from 'corporate*

drone' to actual robot); 3r. when Lucy calls Steve skinny which participants begin to realize is because he has been in survival mode for years; 4r. the 'emergency song' which is not a 'hit song' of the summer, but rather an emergency signal in the apocalypse; 5r. when Lucy tells Steve that the reason she cannot give him food and water is policy (because she is programmed to prevent this). The corresponding 'control' moments are when 1c. Lucy welcomes Steve to the store; 2c. when Lucy is impressed with Steve's knowledge of the store's policies; 3c. when he asks if his trying on a dress is against their policy; 4c. when Lucy chastises Steve; 5c. and when he calls her kind.

To compute the timings of each episode, we first used WhisperX [35] to force-align the stimulus transcript with the auditory narrative. This process yielded an onset and offset timing for each word in seconds. We defined each episode as lasting from the onset of the first word to the offset of the last word. It is important to note that the start of the first reevaluated episode was excluded because it overlapped with the portion of the stimulus excluded to account for transitory delays (i.e., therefore, we only included the remaining portion of the episode).

Next, to directly assess whether the processed of reevaluated episodes showed greater differences between listens compared to control episodes, we applied a general linear model (GLM) analysis. Using a GLM, for each participant, we modeled all episodes (10 total; reevaluated and control) in each listen using individual regressors for each episode (implemented as an individual-modulated event-related analysis using AFNI's *3dDeconvolve* function). This allowed us to obtain voxel-wise beta values for each episode. We then calculated "neural shifts" for each episode as one minus the correlation (correlation distance) between the spatial pattern of voxelwise beta values in each region between Listen 1 and Listen 2 (**Fig. 2B**). Lastly, per region, we fit a LMEM to test the hypothesis that neural shifts would be greater for the reevaluated compared to the control episodes. This was set up using a main effect of episode type (set as a contrast of reevaluated > control), using a random effect of participant and episode pair. Here, episode pair refers to each pair of reevaluated and control episodes that were close in time and matched in length. P-values from the models were corrected for multiple comparisons using FDR with an alpha of 0.05, based on the number of regions analyzed (97; **Fig. 2C**).

Computing updates in the representations of characters.

In our final analysis, we aimed to track where and how characters are represented. To this end, we identified brain regions where, despite receiving the same sensory input on Listen 2, the representations of the character Lucy were updated given the interpretation gained throughout Listen 1. We translated this into a series of four criteria that a region had to meet to be considered as representing character interpretations. Three of these criteria were in regard to Lucy and one was in regard to Steve; this final Steve criterion served as a control to ensure that the representation of Steve was distinct from Lucy. We motivate the focus on Lucy in *Results.* All criteria are described in detail later in this section (see "*Criteria for character representation updating"*).

To isolate the representations of each character, we segmented the stimulus into blocks either Lucy or Steve was speaking (conversational "turns"), defining these as "Lucy events" and "Steve events." We identified the onset and offset of these events using the word-level alignment times provided by WhisperX [35] and manually verified each event. We excluded events shorter than 5

TRs (such as Steve saying his name), resulting in 41 events for Lucy and 33 events for Steve (median event length: 9 TRs; range 5-27 TRs).

For each participant, we ran a GLM for each listen with individual regressors for each speaking event (implemented as an individual-modulated event-related analysis using AFNI's *3dDeconvolve* function, similar to the episodes analysis described above). Then, within each region, we used the voxel-wise beta values from the final event of Listen 1 to define a "template" representation for Lucy. Specifically, in this event, participants should have a finalized understanding of who she is under their revised interpretation following the twist. A description of the template event can be found below.

> *This template event follows Lucy ignoring Steve's pleas and continuing to have an imaginary conversation with a boyfriend about a fictional dinner and wondering about what Steve might be up to. In this event, she references the "dogs not barking" anymore and suggests that perhaps Steve may have fed them; listeners are left to realize that the "dogs" have stopped "barking" because Steve has died.*

Thus, for each participant, we correlated the multivoxel patterns of beta values for each non-template event in both Listen 1 and Listen 2 (either Lucy or Steve) to the participant's own Lucy-template event (**Fig. 3A**). Lastly, we leveraged these event-template pattern intra-SC values to identify character representations (tested using the following criteria).

Criteria for character representation updating.

We expected that representations of Lucy naturally evolved for participants throughout Listen 1 (shopkeeper to robot), and that the updated (robot) representation would be "loaded" back into memory at the start of Listen 2. We evaluated which brain regions exhibited this representational transition—and could therefore be considered to instantiate latent interpretations of characters—as defined by the following criteria.

Criterion 1— Representations of Lucy become more like the template throughout Listen 1.

To test if representations of Lucy become more like the template, we fit a LMEM for each region, predicting the event-template pattern intra-SC in Listen 1 from the event number (with higher numbers corresponding to later events) and treating participant as a random effect. We hypothesized a positive linear trend across character events, with later events showing stronger correlations with the template as representations converge toward the final template, reflecting participants' learning about Lucy across the first listen. For this criterion to be met, the statistic had to be positive.

Criterion 2—Representations of Lucy during her first event are 'updated' in Listen 2.

To test if participants "load in" their updated representation of Lucy when starting Listen 2, for each region and individual, we compared the correlation to the template for the first Lucy character event between the two listens. By comparing the same character event (matched sensory input) across listens to the same template, we inferred that stronger correlations with the template in Listen 2 reflected a shift toward a more updated representation of Lucy. For each region, we performed a paired t-test comparing the distribution of correlation values across participants

between Listens 1 and 2. For this criterion to be met, the statistic had to be positive (Listen 2 > Listen 1).

Criterion 3—Representations of Lucy stabilize in Listen 2.
To test if representations of Lucy "stabilize," we compared how event-template correlations evolved over time between the two listens. We fit a LMEM for each region, predicting the event-template correlation based on an interaction of listen (Listen 1 or Listen 2) and character event number, treating participant as a random effect. As noted in Criterion 1, we hypothesized that there would be a positive linear fit of event number—that is, later character events would be more correlated to the template as the representation built up over the course of the narrative. Here, additionally, we tested that the positive slope would be steeper in Listen 1 relative to Listen 2, given that in Listen 2 the character representation requires less updating and starts out closer to the template because it has been "preloaded" into memory. For this criterion to be met, the statistic of the interaction between Listen and event number had to be positive (slope in Listen 1 > slope in Listen 2) and we furthered delineated regions that were significant at $q_{FDR} < 0.05$. This was our most important criteria– see *Combining these criteria.*

Criterion 4—Control: Representations of Steve are distinct from representations of Lucy in Listen 1 and Listen 2.
As a control, we compared representations of Steve to Lucy's final template. Specifically, we computed event-template correlations using Steve events from Listen 1 and Listen 2. We then fit two independent LMEMs, with participant as a random effect, predicting this correlation from the event number. Given that we expected representations of Steve to be distinct from Lucy, we did not expect a correlation with the template. Therefore, for this criterion to be met, this relationship needed to be negative or at least not significantly positive at an uncorrected threshold of $p < 0.05$. This would indicate that the representations of the characters are becoming increasingly distinct (see **Fig. 3B**, blue lines).

Combining criteria.
We had two steps to combining these criteria. First, to be conservative, we show only regions that fit all of the expected criteria (no effect in Criterion 4 and positive in Criteria 1-3) in our character representation map (**Fig. 3C**). Second, we considered Criterion 3 – *Representations of Lucy stabilize in Listen 2* – as the most critical, given its focus on within-subject, across-listen updating and reloading of character representations. Consequently, we used estimates from this analysis for plotting and for the following analyses (see *Clustering results across analyses*). We also used the corresponding *p*-values to correct for multiple comparisons using FDR with an alpha of 0.05 across the 97 regions analyzed (gray contours in **Fig. 3C**). We additionally highlighted regions that were significant at $q_{FDR} < 0.05$ in all Lucy criteria (1-3; black contours in **Fig. 3C**).

Clustering results across analyses.
In our final analysis, we aimed to assess the extent to which representations of the three narrative elements—global situation models (**Fig. 1D**), episodes (**Fig. 2C**), and characters (**Fig. 3C**)—rely on overlapping versus distinct brain regions. The data from each analysis was normalized using min-max scaling and then concatenated across analyses. The scaled values allowed for consistent comparison across analyses and any excluded regions were set to zero. Excluded regions included those where effects were in the opposite of the expected direction in a given analysis, including 3

regions (3/97; 3%) from the global situation model analysis, 24 regions (~24.7%) from the episode analysis, and 12 regions (~12.4%) from the character representation analysis, where to be more conservative we only included regions that fit the expected direction in all four of our criteria. None of the 97 overall regions were excluded from all three of our analyses.

We then took two approaches to comparing region-wise involvement across the three narrative elements. First, we correlated the scaled estimates across regions between all possible pairs of the three analyses. Second, we used KMeans clustering [36] to perform pattern vector-based clustering, grouping regions based on the similarity of their pattern of estimates derived from each analysis. This approach captured the underlying structure of the relationships across regions and assigned a single cluster label to each region. To determine the optimal number of clusters (k), we calculated the silhouette score for values of k ranging from 2 to 5 and selected $k = 4$ based on the maximum score ($s = 0.30$).

Visualization.
Motivated by a recent recommendation [37], we present largely unthresholded, whole-brain maps for all of our main figures and add contours to indicate regions meeting criteria for statistical significance (described in detail in the caption of each figure). Although our discussion mostly focuses on only those regions meeting statistical significance, we display results across the brain to provide insight into the directionality of effects and facilitate comparisons with past and future work.

To do so, we create a translucent map weighting the data by an opacity (alpha) mask using a threshold of 20% of the maximum range of the data. For each region, we calculate an alpha value (ranging from 0 to 1) to determine its transparency level. If a value exceeds a threshold of 20%, the corresponding region is assigned an alpha value of 1 (fully opaque) and is normally plotted. For values between 0 and our threshold, the alpha value is scaled proportionally from 0 to 1, with increasing transparency for weaker effects. In the global situation model and episode analyses, regions meeting an uncorrected threshold of $p < 0.05$ are outlined in a blue contour, while those meeting a corrected threshold of $q_{FDR} < 0.05$ are outlined in black. In the character analysis, all contoured regions meet the corrected threshold of $q_{FDR} < 0.05$, with color distinctions indicating whether a region satisfies just the main criterion at $q_{FDR} < 0.05$ (in grey) or multiple criteria (in black). The SurfPlot package [38,39] was used for visualization.

Citations:

1. Klink, P. C., van Wezel, R. J. A. & van Ee, R. United we sense, divided we fail: context-driven perception of ambiguous visual stimuli. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 932–941 (2012).
2. Tamborini, R. *et al.* Using Attribution Theory To Explain The Affective Dispositions Of Tireless Moral Monitors Toward Narrative Characters. *J. Commun.* **68**, 842–871 (2018).
3. Lahnakoski, J. M. *et al.* Synchronous brain activity across individuals underlies shared psychological perspectives. *NeuroImage* **100**, 316–324 (2014).
4. Nguyen, M., Vanderwal, T. & Hasson, U. Shared understanding of narratives is correlated with shared neural responses. *NeuroImage* **184**, 161–170 (2019).
5. Sava-Segal, C., Richards, C., Leung, M. & Finn, E. S. Individual differences in neural event segmentation of continuous experiences. *Cereb. Cortex* bhad106 (2023) doi:10.1093/cercor/bhad106.
6. Yeshurun, Y. *et al.* Same Story, Different Story: The Neural Representation of Interpretive Frameworks. *Psychol. Sci.* **28**, 307–319 (2017).
7. Zadbood, A., Nastase, S., Chen, J., Norman, K. A. & Hasson, U. Neural representations of naturalistic events are updated as our understanding of the past changes. *eLife* **11**, e79045 (2022).
8. Ames, D. L., Honey, C. J., Chow, M. A., Todorov, A. & Hasson, U. Contextual Alignment of Cognitive and Neural Dynamics. *J. Cogn. Neurosci.* **27**, 655–664 (2015).
9. Brookshire, G. *et al.* Expertise Modulates Neural Stimulus-Tracking. *eNeuro* **8**, ENEURO.0065-21.2021 (2021).
10. Finn, E. S., Corlett, P. R., Chen, G., Bandettini, P. A. & Constable, R. T. Trait paranoia shapes inter-subject synchrony in brain activity during an ambiguous social narrative. *Nat. Commun.* **9**, 2043 (2018).
11. Lyu, Y., Su, Z., Neumann, D., Meidenbauer, K. L. & Leong, Y. C. Hostile attribution bias shapes neural synchrony in the left ventromedial prefrontal cortex during ambiguous social narratives. *J. Neurosci.* e1252232024 (2024) doi:10.1523/JNEUROSCI.1252-23.2024.
12. Chang, C. H. C., Nastase, S. A. & Hasson, U. Information flow across the cortical timescale hierarchy during narrative construction. *Proc. Natl. Acad. Sci.* **119**, e2209307119 (2022).
13. Grall, C., Tamborini, R., Weber, R. & Schmälzle, R. Stories Collectively Engage Listeners' Brains: Enhanced Intersubject Correlations during Reception of Personal Narratives. *J. Commun.* **71**, 332–355 (2021).
14. Johnson-Laird, P. N. Mental models and probabilistic thinking. *Cognition* **50**, 189–209 (1994).
15. Zwaan, R. A. & Radvansky, G. A. Situation models in language comprehension and memory. *Psychol. Bull.* **123**, 162–185 (1998).
16. Kurby, C. A. & Zacks, J. M. Situation models in naturalistic comprehension. in *Cognitive Neuroscience of Natural Language Use* (ed. Willems, R. M.) 59–76 (Cambridge University Press, Cambridge, 2015). doi:10.1017/CBO9781107323667.004.
17. Speer, N. K., Zacks, J. M. & Reynolds, J. R. Human Brain Activity Time-Locked to Narrative Event Boundaries. *Psychol. Sci.* **18**, 449–455 (2007).
18. Broom, T. W., Chavez, R. S. & Wagner, D. D. Becoming the King in the North: Identification with fictional characters is associated with greater self–other neural overlap. *Soc. Cogn. Affect. Neurosci.* **16**, 541–551 (2021).
19. The Dark End of the Mall. *The Truth* http://www.thetruthpodcast.com/story/2017/1/11/the-dark-end-of-the-mall (2017).
20. Fischl, B. FreeSurfer. *NeuroImage* **62**, 774–781 (2012).
21. Cox, R. W. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res. Int. J.* **29**, 162–173 (1996).
22. DuPre, E. *et al.* TE-dependent analysis of multi-echo fMRI with *tedana*. *J. Open Source Softw.* **6**, 3669 (2021).
23. Kundu, P., Inati, S. J., Evans, J. W., Luh, W.-M. & Bandettini, P. A. Differentiating BOLD and non-BOLD signals in fMRI time series using multi-echo EPI. *NeuroImage* **60**, 1759–1770 (2012).

24. Kundu, P. *et al.* Integrated strategy for improving functional connectivity mapping using multiecho fMRI. *Proc. Natl. Acad. Sci.* **110**, 16187–16192 (2013).
25. Schaefer, A. *et al.* Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cereb. Cortex N. Y. N 1991* **28**, 3095–3114 (2018).
26. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31**, 968–980 (2006).
27. Nastase, S. A., Gazzola, V., Hasson, U. & Keysers, C. Measuring shared responses across subjects using intersubject correlation. *Soc. Cogn. Affect. Neurosci.* **14**, 667–685 (2019).
28. Zhang, A. & Farivar, R. Intersubject Spatial Pattern Correlations During Movie Viewing Are Stimulus-Driven and Nonuniform Across the Cortex. *Cereb. Cortex Commun.* **1**, tgaa076 (2020).
29. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
30. Schmälzle, R. & Grall, C. The Coupled Brains of Captivated Audiences: An Investigation of the Collective Brain Dynamics of an Audience Watching a Suspenseful Film. *J. Media Psychol.* **32**, 187–199 (2020).
31. Song, H., Finn, E. S. & Rosenberg, M. D. Neural signatures of attentional engagement during narratives and its consequences for event memory. *Proc. Natl. Acad. Sci.* **118**, (2021).
32. Deniz, F., Tseng, C., Wehbe, L., Dupré La Tour, T. & Gallant, J. L. Semantic Representations during Language Comprehension Are Affected by Context. *J. Neurosci.* **43**, 3144–3158 (2023).
33. LeBel, A. *et al.* A natural language fMRI dataset for voxelwise encoding models. *Sci. Data* **10**, 555 (2023).
34. Haxby, J. V., Guntupalli, J. S., Nastase, S. A. & Feilong, M. Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife* **9**, e56601 (2020).
35. Bain, M., Huh, J., Han, T. & Zisserman, A. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. Preprint at https://doi.org/10.48550/arXiv.2303.00747 (2023).
36. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28**, 129–137 (1982).
37. Taylor, P. A. *et al.* Highlight results, don't hide them: Enhance interpretation, reduce biases and improve reproducibility. *NeuroImage* **274**, 120138 (2023).
38. Gale, D. J., Vos de Wael., R., Benkarim, O. & Bernhardt, B. Surfplot: Publication-ready brain surface figures. Zenodo https://doi.org/10.5281/zenodo.5567926 (2021).
39. Vos de Wael, R. *et al.* BrainSpace: a toolbox for the analysis of macroscale gradients in neuroimaging and connectomics datasets. *Commun. Biol.* **3**, 1–10 (2020).
40. Baldassano, C., Hasson, U. & Norman, K. A. Representation of Real-World Event Schemas during Narrative Perception. *J. Neurosci.* **38**, 9689–9699 (2018).
41. Masís-Obando, R., Norman, K. A. & Baldassano, C. *Schema Representations in Distinct Brain Networks Support Narrative Memory during Encoding and Retrieval*. 2021.05.17.444363 https://www.biorxiv.org/content/10.1101/2021.05.17.444363v1 (2021) doi:10.1101/2021.05.17.444363.
42. Reagh, Z. M. & Ranganath, C. Flexible reuse of cortico-hippocampal representations during encoding and recall of naturalistic events. *Nat. Commun.* **14**, 1279 (2023).
43. Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A. & Hasson, U. How We Transmit Memories to Other Brains: Constructing Shared Neural Representations Via Communication. *Cereb. Cortex* **27**, 4988–5000 (2017).
44. Whitney, C. *et al.* Neural correlates of narrative shifts during auditory story comprehension. *NeuroImage* **47**, 360–366 (2009).
45. Zwaan, R. A. Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychon. Bull. Rev.* **23**, 1028–1034 (2016).
46. Tik, M. *et al.* Ultra-high-field fMRI insights on insight: Neural correlates of the Aha!-moment. *Hum. Brain Mapp.* **39**, 3241 (2018).
47. Jackson, R. L. The neural correlates of semantic control revisited. *NeuroImage* **224**, 117444 (2021).

48. Jefferies, E. The neural basis of semantic cognition: Converging evidence from neuropsychology, neuroimaging and TMS. *Cortex* **49**, 611–625 (2013).
49. Noonan, K. A., Jefferies, E., Visser, M. & Lambon Ralph, M. A. Going beyond Inferior Prefrontal Involvement in Semantic Control: Evidence for the Additional Contribution of Dorsal Angular Gyrus and Posterior Middle Temporal Cortex. *J. Cogn. Neurosci.* **25**, 1824–1850 (2013).
50. Gonzalez Alam, T. R. del J. *et al.* A tale of two gradients: differences between the left and right hemispheres predict semantic cognition. *Brain Struct. Funct.* **227**, 631–654 (2022).
51. Thye, M., Hoffman, P. & Mirman, D. The neural basis of naturalistic semantic and social cognition. *Sci. Rep.* **14**, 6796 (2024).
52. Dohmatob, E., Dumas, G. & Bzdok, D. Dark control: The default mode network as a reinforcement learning agent. *Hum. Brain Mapp.* **41**, 3318–3341 (2020).
53. Ogawa, A. & Kameda, T. Dissociable roles of left and right temporoparietal junction in strategic competitive interaction. *Soc. Cogn. Affect. Neurosci.* **14**, 1037–1048 (2019).
54. Gilmore, A. W., Nelson, S. M. & McDermott, K. B. A parietal memory network revealed by multiple MRI methods. *Trends Cogn. Sci.* **19**, 534–543 (2015).
55. Kwon, Y. *et al.* Situating the parietal memory network in the context of multiple parallel distributed networks using high-resolution functional connectivity. *bioRxiv* 2023.08.16.553585 (2023) doi:10.1101/2023.08.16.553585.
56. Karagoz, A. B., Morse, S. J. & Reagh, Z. M. Cortico-hippocampal networks carry information about characters and their relationships in an extended narrative. *Neuropsychologia* **191**, 108729 (2023).
57. Ron, Y. *et al.* Brain System for Social Categorization by Narrative Roles. *J. Neurosci.* **42**, 5246–5253 (2022).
58. 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).
59. Chiou, R., Humphreys, G. F. & Lambon Ralph, M. A. Bipartite Functional Fractionation within the Default Network Supports Disparate Forms of Internally Oriented Cognition. *Cereb. Cortex* **30**, 5484–5501 (2020).
60. Song, H., Park, B., Park, H. & Shim, W. M. Cognitive and Neural State Dynamics of Narrative Comprehension. *J. Neurosci.* **41**, 8972–8990 (2021).
61. Curiel, J. M. & Radvansky, G. A. Spatial and character situation model updating. *J. Cogn. Psychol.* **26**, 205–212 (2014).
62. Simony, E. *et al.* Dynamic reconfiguration of the default mode network during narrative comprehension. *Nat. Commun.* **7**, 12141 (2016).
63. Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R. & Buckner, R. L. Functional-Anatomic Fractionation of the Brain's Default Network. *Neuron* **65**, 550–562 (2010).
64. Braga, R. M. & Buckner, R. L. Parallel Interdigitated Distributed Networks within the Individual Estimated by Intrinsic Functional Connectivity. *Neuron* **95**, 457-471.e5 (2017).
65. Buckner, R. L., Andrews-Hanna, J. R. & Schacter, D. L. The brain's default network: Anatomy, function, and relevance to disease. in *The year in cognitive neuroscience 2008* 1–38 (Blackwell Publishing, Malden, 2008).
66. Hassabis, D., Kumaran, D. & Maguire, E. A. Using Imagination to Understand the Neural Basis of Episodic Memory. *J. Neurosci.* **27**, 14365–14374 (2007).
67. Hasson, U., Chen, J. & Honey, C. J. Hierarchical process memory: memory as an integral component of information processing. *Trends Cogn. Sci.* **19**, 304–313 (2015).
68. Himberger, K. D., Chien, H.-Y. & Honey, C. J. Principles of Temporal Processing Across the Cortical Hierarchy. *Neuroscience* **389**, 161–174 (2018).
69. Lerner, Y., Honey, C. J., Silbert, L. J. & Hasson, U. Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *J. Neurosci.* **31**, 2906–2915 (2011).
70. Golesorkhi, M. *et al.* The brain and its time: intrinsic neural timescales are key for input processing. *Commun. Biol.* **4**, 1–16 (2021).

71. Lieberman, M. D. Seeing minds, matter, and meaning: The CEEing model of pre-reflective subjective construal. *Psychol. Rev.* **129**, 830–872 (2022).
72. Grall, C. & Finn, E. S. Leveraging the power of media to drive cognition: a media-informed approach to naturalistic neuroscience. *Soc. Cogn. Affect. Neurosci.* **17**, 598–608 (2022).