Here's the twist: How the brain updates the representations of naturalistic events as our understanding of the past changes

Asieh Zadbood[1], Samuel A. Nastase[2], Janice Chen[3], Kenneth A. Norman[2], Uri Hasson[2]

1 Department of Psychology, Columbia University
2 Princeton Neuroscience Institute and Department of Psychology, Princeton University
3 Department of Psychological and Brain Sciences, John Hopkins University

Abstract

The brain actively reshapes our understanding of past events in light of new incoming information. In the current study, we ask how the brain supports this updating process during the encoding and recall of naturalistic stimuli. One group of participants watched a movie ("The Sixth Sense") with a cinematic "twist" at the end that dramatically changed the interpretation of previous events. Next, participants were asked to verbally recall the movie events, taking into account the new "twist" information. Most participants updated their recall to incorporate the twist. Two additional groups recalled the movie without having to update their memories during recall: one group never saw the twist; another group was exposed to the twist prior to the beginning of the movie, and thus the twist information was incorporated both during encoding and recall. We found that providing participants with information about the twist beforehand altered neural response patterns during movie-viewing in the default mode network (DMN). Moreover, presenting participants with the twist at the end of the movie changed the neural representation of the previously-encoded information during recall in a subset of DMN regions. Further evidence for this transformation was obtained by comparing the neural activation patterns during encoding and recall and correlating them with behavioral signatures of memory updating. Our results demonstrate that neural representations of past events encoded in the DMN are dynamically integrated with new information that reshapes our understanding in natural contexts.

Introduction

In a constantly changing world, it is critical to update prior beliefs and memories in light of new circumstances. As new information arrives, we often need to update previously encoded information in the brain retrospectively. Imagine discovering that a longtime friend has lied to you about something important. You might automatically start looking back and reinterpreting their behavior, perhaps finding different motives for their past actions. This updated understanding of the past will assist you in your future interactions with that friend. Importantly, updating representations of real-world events does not necessarily involve rewriting or erasing the content of the previous memory for the event – it can also include adding new information that alters one's overall interpretation of what happened. In this paper, we use the term 'memory updating' to refer to this process of updating representations of past events based on new information. To effectively support 'memory updating,' the episodic memory system must be capable of modifying stored representations in light of new incoming information. Under this framework, memories are dynamic entities that can be reorganized or reconstructed even after encoding takes place (Bartlett & Burt, 1933; Conway & Pleydell-Pearce, 2000; Hassabis & Maguire, 2007; Schacter et al., 1998; Schacter, 2012).

Research in the last few decades suggests that memories are malleable to modification when they are reactivated (Przybyslawski et al. 1997), and relevant new information is presented (Besnard et al., 2012; Hupbach et al., 2015; Nader & Einarsson, 2010; Sinclair and Barense 2019). Behavioral paradigms using a retroactive interference design have been widely used to study post-encoding changes in human memory (e.g., Lee et al. 2017; Hupbach et al., 2015; Samide & Ritchey, 2020; Scully et al., 2017). Only a subset of studies, however, have investigated changes in the *content* of memory, as opposed to the weakening or strengthening of old memories (Dongaonkar et al., 2013; Hupbach et al., 2007). At the neural level, changes in the functional connectivity of mPFC and amygdala circuitry have been associated with post-retrieval fear extinction (Feng et al., 2016; Schiller et al., 2013). These experimental studies have clinical significance and provide valuable insight into the behavioral and neural substrates of memory updating in humans. However, it is unclear how findings obtained using tightly-controlled paradigms and isolated stimuli generalize to memory updating in everyday life (Nastase et al., 2020). In the present work, we introduce a naturalistic interference-based design that resembles our real-world experiences where new information obtained post-encoding is not compatible with previously encoded events. Using an audiovisual movie and verbal recall, we aim to utilize recent advances in naturalistic neuroimaging to study how memories are reshaped to incorporate new incoming information.

The default mode network is an important hub for integrating old memories with new incoming information for constructing situation models (Hassabis & Maguire, 2009, Yeshurun et al., 2021). These constructive processes are highly relevant to real-world memory updating, which involves selecting and combining the relevant parts of old and new memories. Recent work has shown that neural patterns during encoding and recall of naturalistic stimuli (movies) are reliably similar across participants in this network (Chen et al. 2017; Oedekoven et al., 2017; Zadbood et al., 2017; see Bird 2020 for a review of recent naturalistic studies on memory).

In the current work, using a novel naturalistic paradigm intended to simulate a real-life situation of adaptive memory updating, we asked how new information changes the neural representations in the DMN during the recall of prior knowledge. To answer this question, we used a popular Hollywood-style film titled “The Sixth Sense” (M. Night Shyamalan, 1999), which contains a dramatic twist in the final scene. [Spoiler alert!] The movie depicts the story of a clinical psychologist treating a child who claims to see ghosts. In the final scene, it is revealed that the doctor was in fact, a ghost himself throughout the movie. Therefore, there are two coherent interpretations of the movie: the *Doctor* (or *naive*) interpretation (labeled D in Fig. 1), which is typically held by viewers up until they encounter the “twist ending”; and the *Ghost* (or *spoiled*) interpretation (labeled G in Figure 1), which is held by viewers after they learn about the twist. In this setting, memory updating is operationalized as the transition from the *Doctor* (D) interpretation to the *Ghost* (G) interpretation.

Our study design hinges on the hypothesis that participants who received the twist and are aware that the doctor is a ghost might have distinct neural representations of the events from those who encoded the movie while ignorant of the twist. Importantly, we predicted that encountering the twist after encoding the movie would initiate a retrospective update in the interpretation of the encoded movie and that this update would be reflected in both verbal recall and patterns of brain activity during remembering. In contrast, the neural representations of the events in the movie will remain unchanged during recall in subjects who do not need to update their memories during recall (i.e., in subjects in the no-twist condition who are only aware of the D interpretation, or subjects in the spoiler condition who knew all along about the G interpretation).

In a large set of regions in the DMN, we found that context changed how the movie was encoded into memory. In other words, the neural representations for each event in the movie were different for viewers who believed the doctor was alive versus viewers who believed the doctor was a ghost. Furthermore, in several DMN regions, we found that neural representations were updated during recall for viewers who learned that the doctor was a ghost after watching the movie. Together these results suggest that areas in the default mode network are actively updating the neural representations as they integrate incoming information with prior knowledge.

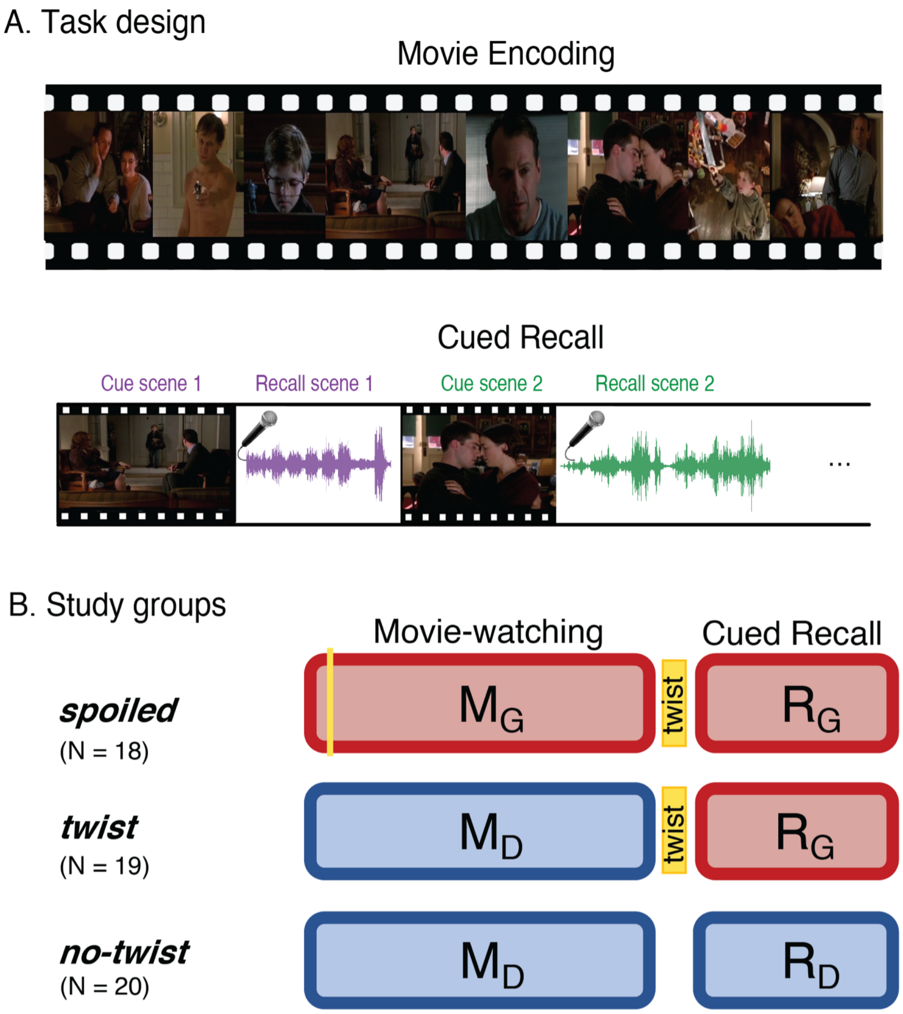


Figure 1: Experimental design. A) Participants watched edited versions of the movie and performed a scene-by-scene cued verbal recall task in the scanner. B) Experimental groups. Red boxes refer to the *Ghost* interpretation, and blue boxes refer to the *Doctor* interpretation. The "twist" group (middle row) is the main experimental group that encodes the movie with *Doctor* interpretation (left blue box) but recalls it with *Ghost* interpretation (right red box)—essentially following the narrative as intended by the filmmaker. The two additional groups keep the same interpretation across the encoding and recall: the "spoiled" group receives a spoiler at the beginning, thus encoding the movie and performing the recall task with the red *Ghost* interpretation, whereas the "no-twist" group never receives the twist and therefore encodes the movie and performs the recall task under the blue *Doctor* interpretation.

Results

Three distinct experimental groups watched concise versions of a popular Hollywood-style film titled "The Sixth Sense" (M. Night Shyamalan, 1999) in the fMRI scanner (Figure 1, right column). Following the movie viewing all three groups were asked to freely recall the movie in the scanner (Figure 1, left column). Participants in the main group (the "twist" condition, Figure 1B, middle row), watched the movie with the twist scene *at the end*. Therefore, they watched the movie naive to the true nature of the doctor (*Movie-Doctor* or M_D). During their recall, however, they were aware of the twist information and could use it to update their memory (*Recall-Ghost* or R_G). In order to identify interpretation-specific neural patterns, we needed two comparison conditions: the *Movie-Ghost* (M_G) condition during viewing, and the *Recall-Doctor* (R_D) condition during recall. Therefore, we introduced two other groups to the study: participants in one group (the "spoiled" condition; Figure 1B, top row) were exposed to the twist at the beginning of the movie. This group watched and recalled the movie knowing that the doctor was a ghost (M_G and R_G). The other group (the "no-twist" condition; Figure 1B, bottom row) never received the twist information throughout encoding and remained naive to the true nature of the doctor in both their encoding and recall (M_D and R_D). This design allowed us to compare the behavioral and neural patterns of response in participants across the two interpretations.

We compared the patterns of neural responses in the "twist" group with the patterns in the "spoiled" and "no-twist" groups during encoding and recall. We predicted that the "twist" group would be more similar to the "no-twist" group during encoding (both having the *Doctor* interpretation) but more similar to the "spoiled" group during recall (both having the *Ghost* interpretation). Moreover, we asked whether the memory updating would make the recall of the "twist" group more similar to the encoding of the "spoiled" group (see the "prediction legends" in Figures 2 and 3). We used intersubject pattern similarity analysis (intersubject pattern correlation: pISC, see Methods) to analyze the neural data. The analysis was performed on scene-specific neural patterns obtained by averaging data across time within each scene (Chen et al. 2017, Zadbood et al. 2017). This analytic approach is motivated by prior work showing that slowly-evolving activity patterns in DMN represent event-level information (see Introduction). As stated above, we focused on DMN regions as *a priori* regions of interest based on this prior work (main results were reproduced in an ROI-based whole-brain analysis; Supplementary Figure 1).

Memory update in recall behavior

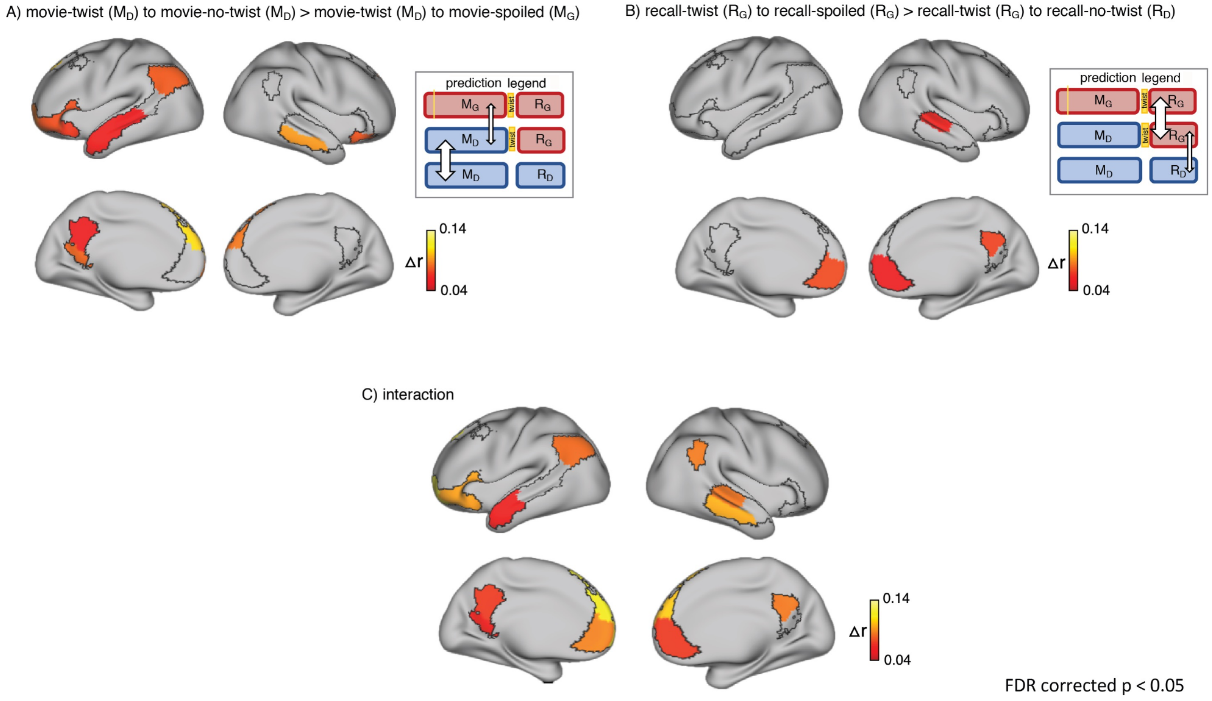
After watching the movie, participants performed a cued-recall task in which they watched a few seconds of the beginning of selected movie scenes and were asked to describe what happened next in that scene. The recall task was identical across the three experimental conditions. Participants were highly accurate in recognizing the corresponding scenes from the movie cues (94% accuracy in the "twist" group, 93% in the "spoiled" group, and 97% in the "no-twist" group). Only the scenes that were correctly recalled were included in the neural analyses. The content of recall was evaluated using two separate measures assigned by human raters. *Memory score* assessed the quality and detail of memory. *Twist score* assessed whether the twist information was incorporated into the recall and ranged from 1 (the recall purely reflected the *Doctor* interpretation) to 5 (the recall purely reflected the *Ghost* interpretation). Memory score and twist score were expected to capture different aspects of the recall behavior; e.g. a detailed recall of the original scene about the doctor treating the child (high memory score) may not include information about the doctor being a ghost (low twist score). Indeed, there was no significant correlation between memory scores and twist scores across participants ($r = 0.07$, $p = 0.56$). If participants were unaware of the twist or did not incorporate it into their recall at all, we would expect the average twist score of the critical scenes to be approximately equal to 1 ("purely reflects the *Doctor* interpretation"). In the main experimental group ("twist" group), 14 out of 19 participants scored above 2 (median score = 3.25) on the twist score, indicating that they incorporated the new interpretation into their recall. Importantly, the "twist" group (twist score: $M = 3.16$, $SD = 1.03$) exhibited a significantly higher twist score ($t(37) = 6.37$, $p < 0.001$) than the "no-twist" group (twist score: $M = 1.65$, $SD = 0.22$). Note that these two groups had no knowledge of the twist when they encoded the movie. Therefore, this result confirms that participants in the "twist group" updated their memories of the movie to incorporate the twist. No significant difference ($t(35) = 1.46$, $p = 0.15$) was observed between the twist score of the "twist" group and the "spoiled" group (twist score: $M = 2.72$, $SD = 0.74$). This finding suggests that the "twist" group recalled the movie more similarly to the group that knew the twist while watching the movie.

A surprising observation during the analysis of the behavioral recall in the "twist" condition was that most participants talked about *both* interpretations of the movie scenes in many of the recalled scenes (this pattern was observed in the recall of the "spoiled" group as well). Thus, it appeared that participants kept both interpretations in mind during the recall, instead of overwriting the *Doctor* representation with the *Ghost* representation. These recalls were typically structured as, "Initially I thought that… but now I know that…" Interestingly, some instances of this recall behavior were also observed in the "spoiled" group, who had watched the movie knowing the doctor is a ghost (e.g. "You could think that… but I knew that…"). This suggests that the neural representations supporting recall in the "twist" and "spoiled" groups included *both* the original (*Doctor*) and updated (*Ghost*) interpretations, which could make differentiating these representations in the neural analysis more challenging (see Discussion).

Neural representation of the twist information during movie-viewing

First, we set out to test how contextual knowledge about the twist modifies the neural patterns in the DMN during the encoding of the movie into memory. We compared the spatially distributed neural activity patterns elicited during movie-viewing (encoding) in the "twist" group (M_D) to the activity patterns obtained during encoding in the "no-twist" group (M_D) and the "spoiled" group (M_G). We hypothesized that within the regions of the brain that are sensitive to different interpretations, the pattern similarity between the “twist” group (M_D) and the “no-twist” group (M_D) should be higher than the similarity between the “twist” group (M_D) and the “spoiled” group (M_G) (Figure 2A, prediction legends).

Indeed, there was significantly greater intersubject pattern correlation in parts of the DMN between the “twist” and “no-twist” experimental groups (who had a similar M_D interpretation of the movie during encoding) than across experimental groups with opposing interpretations (M_D versus M_G). These areas included the dorsal and lateral PFC, left precuneus, left retrosplenial cortex, left angular gyrus, middle temporal cortex, left superior temporal cortex, and left temporal pole (Figure 2A). These results fit with previous findings demonstrating that the timecourse of brain responses in DMN regions reflects different perspectives when listening to a spoken narrative (Yeshurun et al., 2017). Our results extend these findings by showing that different interpretations are discriminable in *spatial* response patterns measured while viewing audiovisual movie stimuli.


A) movie-twist (M_D) to movie-no-twist (M_D) > movie-twist (M_D) to movie-spoiled (M_G)

B) recall-twist (R_G) to recall-spoiled (R_G) > recall-twist (R_G) to recall-no-twist (R_D)

C) interaction

FDR corrected p < 0.05

Figure 2: Brain regions coding for story interpretation at encoding and recall. “Prediction legends” depict the predicted pattern of correlations between groups based on our hypotheses. A) Areas with significantly greater intersubject pattern correlation between groups who encoded the movie with the same interpretation (*Doctor*). B) Areas with significantly greater intersubject pattern correlation between groups who recalled the movie with the same interpretation (*Ghost*). C) Areas with a significant interaction effect, indicating a change in interpretation between encoding and recall (see “Pattern similarity analysis” in Methods). Statistical significance was assessed using a nonparametric randomization test, FDR corrected p < .05.

Neural representation of the twist information during cued recall

Results from the encoding phase suggest that regions in DMN exhibit different patterns of neural response to *Ghost* vs. *Doctor* interpretations. In the next step, we sought to measure memory updating, which we define as a shift during recall from the neural patterns associated with the *Doctor* interpretation to incorporate information associated with *Ghost* interpretation. As described earlier, the analysis of recall behavior suggests that participants in the "twist" condition utilized the twist information to update their recall of the movie. Hence, we ask whether the neural patterns observed during recall would reflect these changes. We predicted that the "no-twist" group and the "spoiled" group would keep the same interpretation of the movie during encoding and recall (M_D to R_D in the "no-twist" group and M_G to R_G in the "spoiled" group). However, in the "twist" group, we expected to observe an update during recall to accommodate the twist information (M_D to R_G). Therefore, we hypothesized that, during recall, the neural patterns for the "twist" group might shift from being more similar to the "no-twist" group as observed during encoding to be more similar to the neural patterns in the "spoiled" group during recall (Figure 2B – prediction legends).

Indeed, as subjects recalled the movie in the scanner, there was significantly greater intersubject pattern correlation in parts of the DMN between the "twist" and "spoiled" experimental groups (who believed that the doctor is a ghost: R_G) than across the "twist" and "no-twist" groups (who had opposing interpretations: R_G versus R_D). These areas included the ventromedial prefrontal cortex (vmPFC), right precuneus, and right superior temporal cortex (Figure 2B). In addition, we ran an interaction analysis to further emphasize the reversal of neural similarity during encoding and recall (see Methods). This analysis highlights a large set of DMN regions, including medial, dorsal, and lateral PFC, precuneus, left retrosplenial cortex, angular gyrus, right superior and middle temporal cortex, and left temporal pole, where neural patterns in the "twist" group were relatively more similar to the *Ghost* (vs. *Doctor*) interpretation at recall than at encoding (Figure 2C).

Relationship between the neural representations during encoding and recall

To provide further neural evidence for the shift from *Doctor* interpretation during encoding to *Ghost* interpretation during recall in the "twist" group, we directly compared the brain responses elicited during encoding and recall. Chen and colleagues (2017) have demonstrated that, across free recall of a movie, neural patterns are reinstated in DMN. In addition, these scene-specific neural patterns changed between encoding and recall in a systematic manner across individuals (Chen et al 2017). We hypothesized that updating one's interpretation to incorporate twist information might alter the neural representations during recall, such that they become more similar to the neural patterns elicited during encoding of the spoiled movie.

We tested this hypothesis in two ways. First, we predicted that (Figure 3A, prediction legend) the neural pattern similarity between recall in the "twist" group and encoding in the "spoiled" group (R_G to M_G) would be higher than the pattern similarity between recall in the "no-twist" group and encoding of the "spoiled" group (R_D to M_G). Our analysis confirmed this prediction in the left angular gyrus, left dorsomedial PFC, and right middle temporal cortex (Figure 3A).

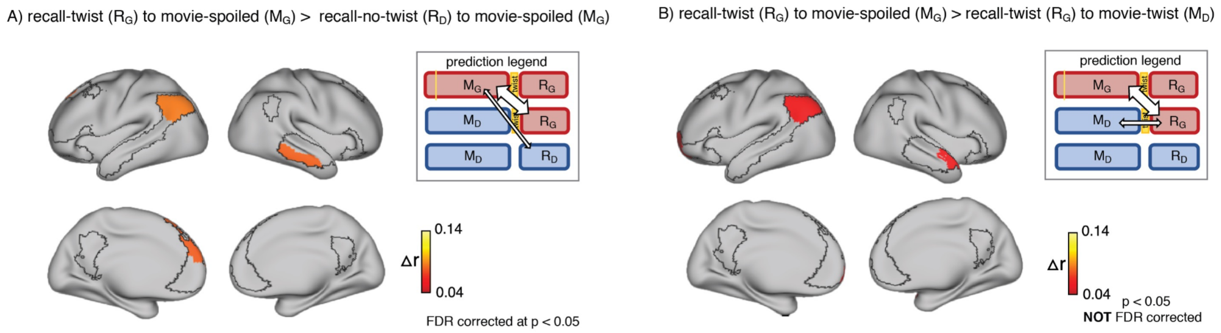

A) recall-twist (R_G) to movie-spoiled (M_G) > recall-no-twist (R_D) to movie-spoiled (M_G)
prediction legend
M_G
twist
R_G
M_D
R_G
M_D
R_D
0.14
Δr
0.04
FDR corrected at p < 0.05
B) recall-twist (R_G) to movie-spoiled (M_G) > recall-twist (R_G) to movie-twist (M_D)
prediction legend
M_G
twist
R_G
M_D
R_G
M_D
R_D
0.14
Δr
0.04
p < 0.05
NOT FDR corrected

Figure 3: Encoding-retrieval similarity analyses to test our memory updating predictions. "Prediction legends" depict the predicted pattern of correlations between groups based on our hypotheses. A) Areas where intersubject pattern correlations were significantly greater when comparing updated recall (R_G) to spoiled encoding (M_G) than when comparing naive recalls (R_D) to spoiled encoding (M_G). B) Areas where intersubject pattern correlations between updated recall (R_G) and spoiled encoding (M_G) were greater than between updated recall (R_G) and naive encoding (M_D); note that these results were not significant after correction for multiple tests.

Second, if participants in the "twist" group were to *fully* update their interpretation at recall from *Doctor* to *Ghost*, we would expect activity patterns during recall in the "twist" group to be more similar to encoding in the "spoiled" group (R_G to M_G) compared to encoding in their own ("twist") group (R_G to M_D) (Figure 3B, prediction legends). When we looked for regions showing this effect, we found weak effects in the predicted direction in the left angular gyrus, left frontal pole, and right anterior temporal ROIs (note that all of these comparisons were performed across participants; see Methods for details); however, these effects did not survive correction for multiple comparisons at an FDR-corrected $p < 0.05$ (Figure 3B). The most straightforward interpretation of these weak effects is, in general, "twist" participants did *not* fully update their interpretations; that is, there may have been some lingering memory of the *Doctor* interpretation in the "twist" group in some participants even after they were exposed to *Ghost* interpretation and updated their memory.

To test this hypothesis, we ran an exploratory analysis where we correlated neural pattern change (i.e., the degree to which the neural pattern at recall matched the *Doctor* or *Ghost* encoding pattern) with behavioral twist scores (i.e., how much each subject discussed the twist during recall) across participants in the "twist" group, in each DMN ROI (Supplementary Figure 2). If weak neural pattern change effects are due to incomplete memory updating, we would expect to see a positive correlation between these measures. We observed a positive correlation between the neural and behavioral indices of memory update in posterior regions of the DMN, including precuneus and angular gyrus. The right precuneus ROI exhibited a notable relationship (r = 0.62); however, this did not survive FDR correction across ROIs.

Discussion

Using a novel naturalistic paradigm that prompted participants to update their previously-encoded memories, we studied how new information can retrospectively change the event representations in the default mode network. At encoding, a widespread network of frontal, parietal, and temporal regions exhibited significantly higher pattern similarity between groups in which participants had the same interpretation of the movie (naive to the twist; see Figure 2A). This result demonstrates how a belief about the identity of the doctor (which can broadly be construed as the context or the state of mind of the observer) can shape the encoding processes of new information (the same movie) into memory. But information is not only shaped by context during encoding, as stored representations must also be amenable to change as the context changes at a later stage. Indeed, our unique paradigm allows us to see how the patterns of stored representations change, as we learn about the twist in the movie. In particular, the neural patterns during recall changed in the twist condition to better match the neural patterns in the spoiled condition observed during recall in the ventromedial PFC, right precuneus, and temporal cortex (see Figure 2B). Furthermore, numerous areas throughout the DMN showed a significant interaction whereby neural patterns in the “twist” group became relatively more similar to patterns from the “spoiled” *Ghost* group (compared to the “no-twist” *Doctor* group) at recall (compared to encoding; Figure 2C).

We also found evidence for memory updating by directly comparing patterns from encoding and retrieval. In the left angular gyrus, left dorsomedial PFC, and right middle temporal cortex, viewing the twist at the end of the movie (vs. not viewing the twist) resulted in neural patterns at recall becoming more similar to the “spoiled” *Ghost* encoding patterns (Figure 3A). In some regions, this updating effect led to “twist” recall patterns being numerically more similar to the “spoiled” encoding patterns than to encoding patterns from the “twist” condition, but this effect did not survive multiple comparisons correction (Figure 3B). We suggested that the weakness of this effect may be attributable to some participants not fully discarding the *Doctor* interpretation when they update their interpretation; in line with this, an exploratory analysis showed that—in some DMN ROIs—the degree of neural change was nominally correlated (across participants) with behavioral “twist scores” capturing how strongly a participant’s recall was influenced by the twist (Supplementary Figure 2; these exploratory correlations did not survive multiple comparisons correction). Taken together, our results provide further evidence for the involvement of DMN regions in integrating new information with prior knowledge to form distinct high-level event representations. In particular, we suggest a subset of core DMN regions are implicated in representing changes in event interpretations during memory updating.

The default mode network, traditionally known to support internally oriented processes, is now considered a major hub for actively processing incoming external information and integrating it with prior knowledge in the social world (Yeshurun et al., 2021). Our experimental design targets naturalistic event representations unfolding over seconds to minutes. There have been many studies to date corroborating the discovery of a cortical hierarchy of increasing temporal receptive windows where high-level event representations are encoded at the top of the hierarchy—in the DMN (Hasson et al., 2008; Lerner et al., 2011, Hasson et al., 2015; Baldassano et al., 2017; etc). This network is involved in episodic encoding and retrieval (Rugg & Vilberg, 2013) and constructive memory-related tasks such as imagining fictitious scenes and future events (Addis et al., 2007; Hassabis et al., 2007; Hassabis & Maguire, 2007; Rugg & Vilberg, 2013; Schacter & Addis, 2007; Schacter et al., 2007). Our design relies on an event-level correspondence between the encoding (viewing) and verbal recall of movie scenes. Previous research has localized modality-independent representations of movie scenes (Zadbood et al 2017) and their similarity during encoding and recall (Chen et al 2017) to the DMN. These characteristics make this network a good candidate to contribute to memory updating—a constructive process in which new information is integrated into past event memories in service of better guiding behavior. Our findings support this idea by showing the shift in neural representations during updated recall in a subset of regions in this network.

At encoding, a widespread set of areas including dorsal and lateral PFC, left precuneus, left retrosplenial cortex, and left angular gyrus had differentiable neural patterns across the two interpretations of the movie. These results are consistent with previous work that showed the time course of brain responses in DMN distinguishes between groups when participants are prompted to take two different perspectives before listening to an audio story (Yeshurun et al., 2017). We extend these results to an audiovisual movie, and provide evidence that interpretative perspective is also encoded in spatially distributed neural response patterns for narrative events, averaged across minutes-long scenes. Interestingly, the difference in neural responses measured by Yeshurun and colleagues was not significant between the two perspectives of the story in the ventral portion of mPFC. Similarly, vmPFC ROIs did not exhibit a significant difference between the *Doctor* and *Ghost* representations during the encoding phase in our experiment. Previous research has implicated mPFC in processing schematic information and integration of new information into prior knowledge (Gilboa & Marlatte, 2017; Schlichting & Preston, 2017; van Kesteren et al., 2012). Using naturalistic clips as schematic events, it has been shown that response patterns in mPFC are particularly dependent on intact and predictable schemas (Baldassano et al., 2018). Together, these results suggest that our manipulation (*Doctor* and *Ghost* interpretations) may not have substantially altered the schemas that participants were using during movie-viewing (e.g., during a restaurant scene, participants will need to use their “restaurant” schema to interpret it, regardless of whether the doctor is alive or a ghost)—although we interpret these null results with caution.

Even though groups had different knowledge/perspectives during encoding, we found higher pattern similarity across groups if they had similar twist knowledge during recall in vmPFC, right precuneus, and parts of temporal cortex. Previous findings suggest mPFC is involved in not just encoding but retrieval of memories in relation to prior knowledge (Brod et al., 2015; van Kesteren et al., 2010) and retrieval of overlapping representations to support integration and organization of related memories (Tompary & Davachi, 2017). Our observations during recall fit with these findings and suggest that shifting toward a more similar perspective during recall leads to higher neural similarity in mPFC. However, during encoding, we did not observe a significant pattern correlation between groups that held the same interpretation of the movie. Furthermore, vmPFC was significant in our interaction analysis (Figure 2C), indicating that the similarity structure of vmPFC patterns across conditions was significantly different at encoding versus retrieval. Together, these results suggest vmPFC is differently implicated in encoding and recall of story-specific representations during processing of naturalistic events.

During recall, many participants recounted both the old and new interpretations (*Ghost* and *Doctor*) of movie scenes. This behavior indicated that they maintained both representations in parallel (possibly competing), rather than overwriting the old representation with new information. The simultaneous presence of these representations poses an interesting theoretical question for future studies: When does updating the memory cause us to lose traces of the old interpretation, and when do the old and new interpretations end up co-existing in memory? Previous studies have shown that old and new memory traces are simultaneously reactivated in the brain, leading to competition (e.g., Kuhl et al., 2012), and this competition can trigger learning processes that resolve the competition; e.g., by weakening one of the memories or by restructuring the memories so they can coexist (Ritvo et al., 2019). Understanding how competition between interpretations plays out over time is an important topic for future work; existing research on memory revaluation suggests that updating may be a temporally-extended process driven by successive replays of the new information, rather than taking place all at once (see, e.g., Momennejad et al., 2018). In clinical settings, methods inspired by reconsolidation and memory updating are extensively used to treat maladaptive memories (Phelps & Hofmann, 2019). In these clinical contexts, it will be especially important to understand the factors that influence the “end state” of this competition between interpretations (in terms of our study: who ends up fully adopting the *Ghost* interpretation and who ends up with lingering traces of the *Doctor* interpretation).

Setting aside these points about whether the original (*Doctor*) interpretation lingers on, our findings clearly show that the twist causes the *Ghost* interpretation to take root in participants’ brains. Overall, these results highlight the importance of DMN regions in updating naturalistic memories and suggest new approaches to studying real-world memory modification in both experimental and clinical treatment settings.

Materials and Methods

Stimuli

The stimuli consisted of three edited versions of “The Sixth Sense” (M. Night Shyamalan, 1999) movie. The movie depicts the story of a child psychologist treating a young boy who claims he can see and speak with dead people. In the film's ending scene, however, it is revealed that the psychologist died prior to the events of the movie and has actually been one of the ghosts the boy was seeing all along. Three different edited versions of the movie were created for the experiment. The first version was a ~60-min shortened movie including the final scene with the big reveal followed by a text on the screen describing the twist to ensure all participants in the "twist" group fully understood the twist information. The second version was identical to the first version, but a spoiler was presented as text on screen early in the movie (the "spoiled" group). In the third version, the final scene was cut out and the movie ended at a point where it appeared that the doctor successfully completed the treatment and therefore did not raise any suspicion about the twist in participants who watched this version ("no-twist" group). Eighteen scenes were selected to be included in the cued recall task (see the section on timestamping and scene selection below). For each of these scenes, a short clip from the beginning of that scene (lasting from 5 to 36 seconds. Mean = 12.9 sec) was used as a retrieval cue for the scene during the recall task.

Participants

Sixty-six right-handed, native English speakers (ages 18–24, average = 20, 21 males) were scanned in the experiment. None of the participants had previously seen The Sixth Sense in full or in part, which was confirmed through an online questionnaire before the session. However, because the movie is well-known and frequently referenced in popular culture, participants with some knowledge about the twist (e.g. knowing that this is a movie about ghosts and the main character is actually dead) were admitted to the "spoiled" group (see *Experimental design*) in order to facilitate data collection. In the post-scan questionnaire, two participants reported guessing the twist while watching the movie and their data were excluded. One participant did not understand the twist after watching the final scene and receiving the text explanation, so their data were omitted as well. Six participants were excluded due to large head motion (spikes of framewise displacement > 4 mm). The data of the remaining fifty-seven participants were used in the analyses. All participants provided written informed consent prior to the experiment and received information about the conditions of the experiment and their rights. The experiment protocol and the consent forms were approved by the Institutional Review Board of Princeton University.

Experimental design

Participants were pseudo-randomly divided into three groups: the "twist" group (N = 19) watched a 60-min audio-visual edition of The Sixth Sense movie, including the twist at the end while undergoing fMRI scanning. The "spoiled" group (N = 18) watched a spoiled version of the movie (see Stimuli). The "no-twist" group (N = 20) watched a 55 min version of the movie with no twist scene (Figure 3B). Participants were instructed to watch the movie naturally and attentively, as there will be a task related to the movie content after watching. However, no specific information about the upcoming recall task was revealed. After the movie, participants performed a verbal cued recall task. During the cued recall task, participants watched short clips from 18 scenes of the movie. After each clip, they were asked to freely describe the events of that particular scene and to provide the most accurate interpretation of the scene given all the information they have gathered throughout watching. The instructions were identical for all three groups. The movie cue and recall were separated by 14 seconds, which ended as a countdown on the screen. The recall task was self-paced and participants pressed a button to continue to the next scene after each recall. After scanning, participants filled out a questionnaire about their experience in the scanner, including information about the movie and recall tasks and whether they guessed the twist in the middle of the movie (and if yes in which scene). All participants rated the movie as engaging. Participants in the “no-twist” group were debriefed about the real ending of the movie before leaving the facility.

Scanning procedure

The scanning session began with an anatomical scan. Participants watched the movie and read the instructions through a mirror mounted to the head coil which reflected a rear screen. The main screen was located at the back of the magnet bore and the movie was projected on the screen via an LCD projector. MR-safe, in-ear headphones were used for the movie audio. Eye-tracking was set up to monitor participants during the scans in real-time and ensure they stayed awake and attentive during the experiment. The movie and recall stimuli were presented using the Psychophysics Toolbox in MATLAB (Mathworks), which enabled coordinating the onset of the stimuli (movie and recall cues) with data acquisition. The volume level of the movie was adjusted separately for each participant using a sample clip to assure a clear and comfortable audio signal. Recall speech was recorded during the fMRI scan using a customized MR-compatible recording system (FOMRI II; Optoacoustics Ltd). The MR recording system used two orthogonally-oriented optical microphones. The reference microphone captures the background noise, and the source microphone captures both background noise and the speaker's speech (signal). A dual-adaptive filter subtracted the reference input from the source channel using a least mean square approach. To achieve an optimal subtraction, the reference signal was adaptively filtered so the filter gains are learned continuously from the residual signal and the reference input. To prevent divergence of the filter when speech was present, a voice activity detector was integrated into the algorithm. A speech enhancement spectral filtering algorithm further preprocessed the speech output to achieve a real-time speech enhancement. Audio recordings were further cleaned using noise removal software (Adobe Audition). The output recall recordings were fully comprehensible. A response box was used to collect the participants’ manual button-presses during the recall task. Participants were instructed to press a button when they finished the recall of a scene to proceed with the task. In five participants, the recall scans were stopped due to problems in pressing the buttons (or just by mistake) and were resumed after they received feedback and further instructions. In these cases, the recalls were resumed starting with the next scene. In three participants the recall scan was stopped after the first scene and in one participant before the last two scenes. In one participant the scan stopped and resumed in the middle of the recall task.

MRI acquisition

MRI data were collected on a 3T full-body scanner (Siemens Prisma) with a 64-channel head coil. Functional images were acquired using an interleaved multiband EPI sequence (TR= 1500 ms, TE 33 ms, flip angle 80 degrees , whole-brain coverage, 2 mm slice thickness, FOV 192 mm^2, SMS = 4). Anatomical images were acquired using a T1-weighted magnetization-prepared rapid-acquisition gradient echo (MPRAGE) pulse sequence (1 mm^3 resolution). Anatomical images were acquired in a 6-min scan before the functional scans with no stimulus on the screen. Field maps were collected for B0 correction at the end of the recall run.

Preprocessing

Preprocessing was performed using fMRIPrep, version stable 1.0.11(Esteban et al., 2019, RRID:SCR_016216), a Nipype (Gorgolewski et al., 2011, RRID:SCR_002502) based tool. Each T1w (T1-weighted) volume was corrected for INU (intensity non-uniformity) using N4BiasFieldCorrection v2.1.0 (Tustison et al., 2010) and skull-stripped using antsBrainExtraction.sh v2.1.0 (using the OASIS template). Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al., 2009, RRID:SCR_008796) was performed through nonlinear registration with the antsRegistration tool of ANTs v2.1.0 (Avants et al., 2008, RRID:SCR_004757), using brain-extracted versions of both T1w volume and template.

Functional data were motion corrected using mcflirt (FSL v5.0.9, Jenkinson et al., 2002). "Fieldmap-less" distortion correction was performed by co-registering the functional image to the same-subject T1w image with intensity inverted (Wang et al., 2017) constrained with an average fieldmap template (Treiber et al., 2016), implemented with antsRegistration (ANTs). This was followed by co-registration to the corresponding T1w using boundary-based registration (Greve & Fischl, 2009) with six degrees of freedom, using flirt (FSL). Motion correcting transformations, field distortion correcting warp, BOLD-to-T1w transformation and T1w-to-template (MNI) warp were concatenated and applied in a single step using antsApplyTransforms (ANTs v2.1.0) using Lanczos interpolation.Frame-wise displacement (Power et al., 2014) was calculated for each functional run using the implementation of Nipype.

Then, the datasets were adaptively smoothed using AFNI's 3dBlurToFWHM to reach 7 mm global smoothness (Cox, 1996). Note that the 7 mm reported smoothness is the *global smoothness*, which is the “final” smoothness of the images given their original, intrinsic smoothness and the applied smoothing. In other words, we did not apply an additional 7 mm smoothing kernel to the data; rather, we iteratively smoothed the data until a 7 mm global smoothness was attained (using AFNI's 3dBlurToFWHM). If the initial smoothness of the raw data was roughly 2 mm, this would be similar to applying a 5 mm smoothing kernel. This amount of smoothing is comparable to previous papers using similar intersubject pattern similarity methods to compare event-level representations during encoding and recall (Chen et al., 2017; Zadbood et al., 2017). AFNI's 3dTproject was used to regress out confound variables comprising head motion (6 motion parameters and their temporal derivatives), second-order polynomial detrending variables, and high-pass filtering (140 second cutoff). De-spiking and subsequent analyses were conducted using custom MATLAB scripts (see Code Accessibility). The movie data were acquired in a single run and the time series were z-scored across the entire run prior to further analysis. Inspection of the recall data revealed a dramatic difference in mean signal intensity between the audiovisual movie cues and the verbal recall sections during the cued-recall task. To account for this, we used the least-squares-separate (LSS) method (Mumford et al., 2012) implemented by AFNI's 3dLSS to model the recall data. In this method each verbal recall section was modeled independently of both the other recall scenes and the preceding movie cue. Regression coefficients (beta values) obtained by this method (one beta value per scene) were used in the main analyses. In four participants where the recall scan was split due to button-press issues, the smaller section of the recall only included 1–2 scenes. These scans were too short to be modeled using LSS and the data for these scenes were ignored. All analyses were performed in volume space. The results were projected onto the surface for visualization using Connectome Workbench (https://www.humanconnectome.org/software/connectome-workbench).

Atlas and ROI definition

Whole brain ROI analysis was performed on a set of 100 ROIs grouped into seven networks based on functional connectivity during rest (Schaefer et al., 2018). Thirty-four of these ROIs labeled as "DMN" were used in the main analysis.

Timestamping and scene selection

The movie was time-stamped by an independent rater naive to the purpose and design of the experiment to identify the main scenes of the movie. Eighteen scenes with clear scene boundaries were selected to be used in the cued-recall task. Very short snippets from the beginning of these scenes were used as cues in the recall task. A group of four raters watched the movie and rated the selected 18 scenes in terms of how much the twist information might change the interpretation of these scenes. They were instructed to rate each scene on a scale of 1 to 5 (1 = Interpretation does not change at all, 2 = Interpretation is mildly changed, 3 = Interpretation is moderately changed, 4 = Interpretation is strongly changed, 5 = Interpretation is very strongly changed). Six scenes scored 4 or higher ("Interpretation is strongly changed")—these *critical scenes* were selected for the main neural analyses. In the independent analysis of the recall behavior data, this same set of 6 scenes scored highest in *twist score* (described in the next section) which indicates a match between expected and perceived effect of twist information on recall behavior. Scene number one, in which the doctor and child meet for the first time was scored ~3 (Interpretation is moderately changed) but showed a high twist score in the behavioral recall analysis. This scene was the first time participants recalled the doctor after the main reveal (watching the twist) and given its high twist score, the recall and possibly the corresponding neural patterns appeared to be more strongly affected by the twist information. Therefore, we added this scene as a seventh *critical scene* to be used in the main neural analyses.

Behavioral analysis

The recall data were transcribed from speech to text and subject numbers (and group information) were removed. The same four raters who watched the movie and rated the "twist influence" in the previous section read the recall data scene by scene. They rated each scene for all subjects, while the order of scenes across subjects was shuffled and there was no information indicating to which experimental group the scene belonged. They were asked to report a score for each scene based on the "ghostness" or "doctorness" of the depiction of the main character in that scene. The scores were from 1 to 5 (1 = Purely reflects the *Doctor* interpretation, 2 = More strongly reflects the *Doctor* interpretation, 3 = Balanced between *Doctor* and *Ghost* interpretation, 4 = More strongly reflects the *Ghost* interpretation, 5 = Purely reflects the *Ghost* interpretation). Raters showed strong agreement on their scoring (pairwise correlations between raters' scores ranged from $r = 0.84$, $p = 6.6 \times 10^{-18}$ to $r = 0.97$, $p = 7.5 \times 10^{-42}$). Scores for each scene were averaged across 4 raters and were used as the twist score in the main analyses. Two separate raters scored the recall data based on the details and accuracy of recall irrespective of the twist information. Scores provided by these raters were averaged and used as the "memory score."

Intersubject pattern correlation (pISC) analysis

The multivariate analysis of the data was performed by measuring the similarity between the spatial patterns of brain response in each ROI. To obtain this measure, first the time series of brain responses to the movie in each subject/ROI was averaged within each of the seven critical scenes. This method has been used to study scene-specific patterns of brain activity in previous studies (Chen et al., 2017; Zadbood et al., 2017). Averaging the time series within each scene resulted in seven spatially distributed patterns of brain activity in each ROI. For the recall phase, the beta values extracted via LSS modeling were used, similarly providing 7 activity patterns in each ROI. All pattern similarity analyses were performed between subjects to capitalize on the between-group design of the experiment (Nastase et al., 2019). For the encoding phase, the patterns of brain activity in each subject from the "twist" group were correlated (Pearson correlation) with the average of activity patterns for the "spoiled" group in corresponding scenes and averaged across scenes. The same procedure was performed to compare the "twist" and "no-twist" groups which resulted in two correlation values assigned to each subject in the "twist" group. All correlation values were Fisher transformed prior to further analysis (Fisher, 1915). In each ROI, the difference between these two comparisons was calculated and averaged across participants (difference r values depicted on each map). To determine statistical significance, we compared these two sets of values using a non-parametric paired t-test by shuffling the sign of difference values across subjects 1000 times and calculating a p-value for the observed difference based on this null distribution (one-tailed). P values were corrected for multiple comparisons across DMN ROIs by controlling the false discovery rate (FDR) at $p < .05$ (Benjamini & Hochberg, 1995). The same procedure was performed in the recall and encoding-recall analysis except for two differences in the encoding-recall analysis: during the analysis to compare "twist" and "no-twist" recall with "spoiled" encoding (Figure 3A), an independent sample non-parametric t test was performed by shuffling the group labels 1000 times and calculating the difference between the two permuted groups at each iteration to create the null distribution. To compare the “twist” recall with the “twist” encoding (Figure 3B), each subject’s recall was compared to the average of the rest of the group’s encoding to ensure all comparisons were made across subjects. To match the number of subjects in the encoding groups, one subject was randomly dropped from the encoding set in each iteration when comparing “twist” recall to “spoiled” encoding.

The interaction analysis assessed whether neural patterns in the "twist" group were relatively more similar to the "spoiled" (vs. "no-twist") group at recall (vs. encoding), and was computed as follows:

interaction index (r) = (*movie-no-twist* vs. *movie-twist*) – (*movie-spoiled* vs. *movie-twist*) – [(*recall-no-twist* vs. *recall-twist*) – (*recall-spoiled* vs. *recall-twist*)]

To ensure that our results were not biased due to any systematic differences in the noise level of neural activity patterns between the groups (*spoiled*, *twist*, *no-twist*), we calculated the pISC within each group by correlating each subject’s pattern with the average pattern from the rest of the subjects in that group. We performed this procedure for the movie and recall conditions separately in each of the 15 ROIs that showed any significant effect in any of the reported analyses. We then submitted all the correlation values across subjects to an ANOVA including all groups, conditions, and ROIs. As expected, we did not find any main effect of group or an interaction of group with condition or ROI.

In the analysis to identify the relationship between the neural and behavioral signature of memory update (Supplementary Figure 2), the neural data were obtained by computing (r*ecall-twist* vs.*movie-spoiled*) – (r*ecall-twist* vs.*movie-twist*), as mentioned above and described in the results section. However, the difference values were not averaged and were correlated with the twist score across participants.

Code and data accessibility:

Code available at: https://github.com/azadbood/sixthsense

Data are accessible upon request and will be made publicly available soon as well.

Acknowledgments

This work was supported by the National Institutes of Health (NIMH R01 MH12357 awarded to UH and KAN). We thank Savannah Born for help in analyzing the behavioral data.

Authors contributions

A.Z., J.C., K.A.N., and U.H. designed the experiment. A.Z. collected and analyzed the data and wrote the manuscript. S.A.N. assisted in data collection and analyses and edited the manuscript. U.H., K.A.N., and J.C. supervised the project and edited the manuscript.

References

Addis, D. R., Wong, A. T., & Schacter, D. L. (2007). Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia*, *45*(7), 1363–1377. https://doi.org/10.1016/j.neuropsychologia.2006.10.016

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41. https://doi.org/10.1016/j.media.2007.06.004

Baldassano, C., Hasson, U., & Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *Journal of Neuroscience*, *38*(45), 9689–9699. https://doi.org/10.1523/JNEUROSCI.0251-18.2018

Bartlett, F. C., & Burt, C. (1933). Remembering: A study in experimental and social psychology. *British Journal of Educational Psychology*, *3*(2), 187–192. https://doi.org/10.1111/j.2044-8279.1933.tb02913.x

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Besnard, A., Caboche, J., & Laroche, S. (2012). Reconsolidation of memory: A decade of debate. *Progress in Neurobiology*, *99*(1), 61–80. https://doi.org/10.1016/j.pneurobio.2012.07.002

Bird, C. M. (2020). How do we remember events? *Current Opinion in Behavioral Sciences*,*32*, 120–125. https://doi.org/10.1016/j.cobeha.2020.01.020

Brod, G., Lindenberger, U., Werkle-Bergner, M., & Shing, Y. L. (2015). Differences in the neural signature of remembering schema-congruent and schema-incongruent events. *NeuroImage*, *117*, 358–366. https://doi.org/10.1016/j.neuroimage.2015.05.086

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, *20*(1), 115–125. https://doi.org/10.1038/nn.4450

Conway, Martin. A., & Pleydell-Pearce, Christopher. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review*, *107*(2), 261–288.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, an International Journal*, *29*(3), 162–173. https://doi.org/10.1006/cbmr.1996.0014

Dongaonkar, B., Hupbach, A., Gomez, R., & Nadel, L. (2013). Effects of psychosocial stress on episodic memory updating. *Psychopharmacology*, *226*(4), 769–779. https://doi.org/10.1007/s00213-013-2998-8

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Feng, P., Zheng, Y., & Feng, T. (2016). Resting-state functional connectivity between amygdala and the ventromedial prefrontal cortex following fear reminder predicts fear extinction. *Social Cognitive and Affective Neuroscience*, *11*(6), 991–1001. https://doi.org/10.1093/scan/nsw031

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*(4), 507–521. https://doi.org/10.2307/2331838

Fonov, V., Evans, A., McKinstry, R., Almli, C., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47*, S102. https://doi.org/10.1016/S1053-8119(09)70884-5

Gilboa, A., & Marlatte, H. (2017). Neurobiology of schemas and schema-mediated memory. *Trends in Cognitive Sciences*, *21*(8), 618–631. https://doi.org/10.1016/j.tics.2017.04.013

Gorgolewski, K., Burns, C., Madison, C., Clark, D., Halchenko, Y., Waskom, M., & Ghosh, S. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in Python. *Frontiers in Neuroinformatics*, *5*, 13. https://doi.org/10.3389/fninf.2011.00013

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in Cognitive Sciences*, *11*(7), 299–306. https://doi.org/10.1016/j.tics.2007.05.001

Hassabis, D., & Maguire, E. A. (2009). The construction system of the brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1263–1271. https://doi.org/10.1098/rstb.2008.0296

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *The Journal of Neuroscience*, *28*(10), 2539–2550. https://doi.org/10.1523/JNEUROSCI.5487-07.2008

Hupbach, A., Gomez, R., Hardt, O., & Nadel, L. (2007). Reconsolidation of episodic memories: A subtle reminder triggers integration of new information. *Learning & Memory*, *14*(1–2), 47–53. https://doi.org/10.1101/lm.365707

Hupbach, A., Gomez, R., & Nadel, L. (2015). Memory reconsolidation. In *The Wiley Handbook on the Cognitive Neuroscience of Memory* (pp. 244–264). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118332634.ch12

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain Images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1006/nimg.2002.1132

Kesteren, M. T. R. van, Rijpkema, M., Ruiter, D. J., & Fernández, G. (2010). Retrieval of associative information congruent with prior knowledge is related to increased medial prefrontal activity and connectivity. *Journal of Neuroscience*, *30*(47), 15888–15894. https://doi.org/10.1523/JNEUROSCI.2674-10.2010

Kuhl, B. A., Bainbridge, W. A., & Chun, M. M. (2012). Neural reactivation reveals mechanisms for updating memory. *Journal of Neuroscience*, *32*(10), 3453–3461. https://doi.org/10.1523/JNEUROSCI.5846-11.2012

Lee, J. L. C., Nader, K., & Schiller, D. (2017). An update on memory reconsolidation updating. *Trends in Cognitive Sciences*, *21*(7), 531–545. https://doi.org/10.1016/j.tics.2017.04.006

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *The Journal of Neuroscience*, *31*(8), 2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *ELife*, *7*, e32548. https://doi.org/10.7554/eLife.32548

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage*, *59*(3), 2636–2643. https://doi.org/10.1016/j.neuroimage.2011.08.076

Nader, K., & Einarsson, E. Ö. (2010). Memory reconsolidation: An update. *Annals of the New York Academy of Sciences*, *1191*(1), 27–41. https://doi.org/10.1111/j.1749-6632.2010.05443.x

Nastase, S. A., Gazzola, V., Hasson, U., & Keysers, C. (2019). Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience*, *14*(6), 667–685. https://doi.org/10.1093/scan/nsz037

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, *222*, 117254. https://doi.org/10.1016/j.neuroimage.2020.117254

Oedekoven, C. S. H., Keidel, J. L., Berens, S. C., & Bird, C. M. (2017). Reinstatement of memory representations for lifelike events over the course of a week. *Scientific Reports*, *7*(1), 14305. https://doi.org/10.1038/s41598-017-13938-4

Phelps, E. A., & Hofmann, S. G. (2019). Memory editing from science fiction to clinical practice. *Nature*, *572*(7767), 43–50. https://doi.org/10.1038/s41586-019-1433-7

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, *84*, 320–341. https://doi.org/10.1016/j.neuroimage.2013.08.048

Przybyslawski, J., & Sara, S. J. (1997). Reconsolidation of memory after its reactivation. *Behavioural Brain Research*, *84*(1–2), 241–246.

Ritvo, V. J. H., Turk-Browne, N. B., & Norman, K. A. (2019). Nonmonotonic plasticity: How memory retrieval drives learning. *Trends in Cognitive Sciences*, *23*(9), 726–742. https://doi.org/10.1016/j.tics.2019.06.007

Rugg, M. D., & Vilberg, K. L. (2013). Brain networks underlying episodic memory retrieval. *Current Opinion in Neurobiology*, *23*(2), 255–260. https://doi.org/10.1016/j.conb.2012.11.005

Samide, R., & Ritchey, M. (2020). Reframing the past: role of memory processes in emotion regulation. *Cognitive Therapy and Research*. https://doi.org/10.1007/s10608-020-10166-5

Schacter, D. L. (2012). Constructive memory: Past and future. *Dialogues in Clinical Neuroscience*, *14*(1), 7–18.

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 773–786. https://doi.org/10.1098/rstb.2007.2087

Schacter, D. L., Norman, K. A., & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology*, *49*, 289–318. https://doi.org/10.1146/annurev.psych.49.1.289

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex (New York, N.Y.: 1991)*, *28*(9), 3095–3114. https://doi.org/10.1093/cercor/bhx179

Schiller, D., Kanen, J. W., LeDoux, J. E., Monfils, M.-H., & Phelps, E. A. (2013). Extinction during reconsolidation of threat memory diminishes prefrontal cortex involvement. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(50), 20040–20045. https://doi.org/10.1073/pnas.1320322110

Schlichting, M. L., & Preston, A. R. (2017). The Hippocampus and memory integration: Building knowledge to navigate future decisions. In D. E. Hannula & M. C. Duff (Eds.), *The Hippocampus from Cells to Systems: Structure, Connectivity, and Functional Contributions to Memory and Flexible Cognition* (pp. 405–437). Springer International Publishing. https://doi.org/10.1007/978-3-319-50406-3_13

Scully, I. D., Napper, L. E., & Hupbach, A. (2017). Does reactivation trigger episodic memory change? A meta-analysis. *Neurobiology of Learning and Memory*, *142*, 99–107. https://doi.org/10.1016/j.nlm.2016.12.012

Sinclair, A. H., & Barense, M. D. (2019). Prediction error and memory reactivation: How incomplete reminders drive reconsolidation. *Trends in Neurosciences*, *42*(10), 727–739. https://doi.org/10.1016/j.tins.2019.08.007

Tompary, A., & Davachi, L. (2017). Consolidation promotes the emergence of representational overlap in the hippocampus and medial prefrontal cortex. *Neuron*, *96*(1), 228-241.e5. https://doi.org/10.1016/j.neuron.2017.09.005

Treiber, J. M., White, N. S., Steed, T. C., Bartsch, H., Holland, D., Farid, N., McDonald, C. R., Carter, B. S., Dale, A. M., & Chen, C. C. (2016). Characterization and correction of geometric distortions in 814 diffusion weighted images. *PLOS ONE*, *11*(3), e0152472. https://doi.org/10.1371/journal.pone.0152472

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320. https://doi.org/10.1109/TMI.2010.2046908

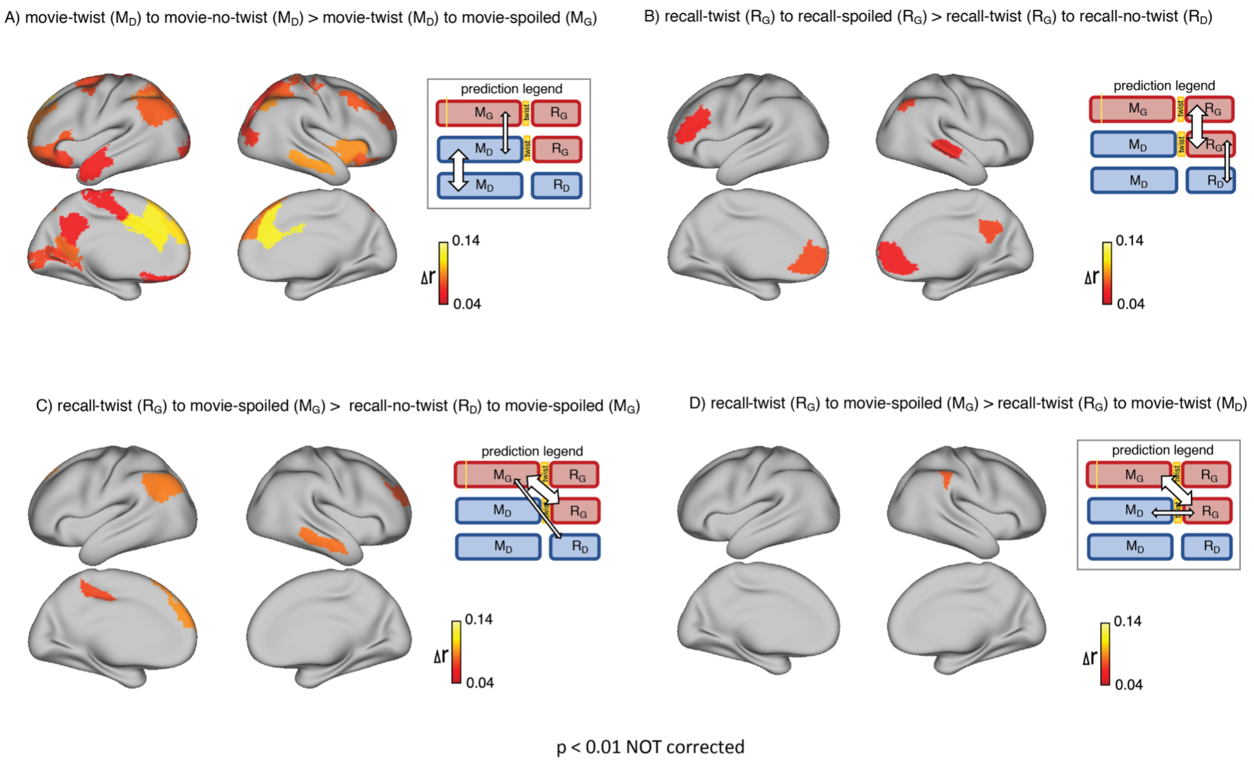
van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, *35*(4), 211–219. https://doi.org/10.1016/j.tins.2012.02.001

Wang, S., Peterson, D. J., Gatenby, J. C., Li, W., Grabowski, T. J., & Madhyastha, T. M. (2017). Evaluation of field map and nonlinear registration methods for correction of susceptibility artifacts in diffusion MRI. *Frontiers in Neuroinformatics*, *11*, 17. https://doi.org/10.3389/fninf.2017.00017

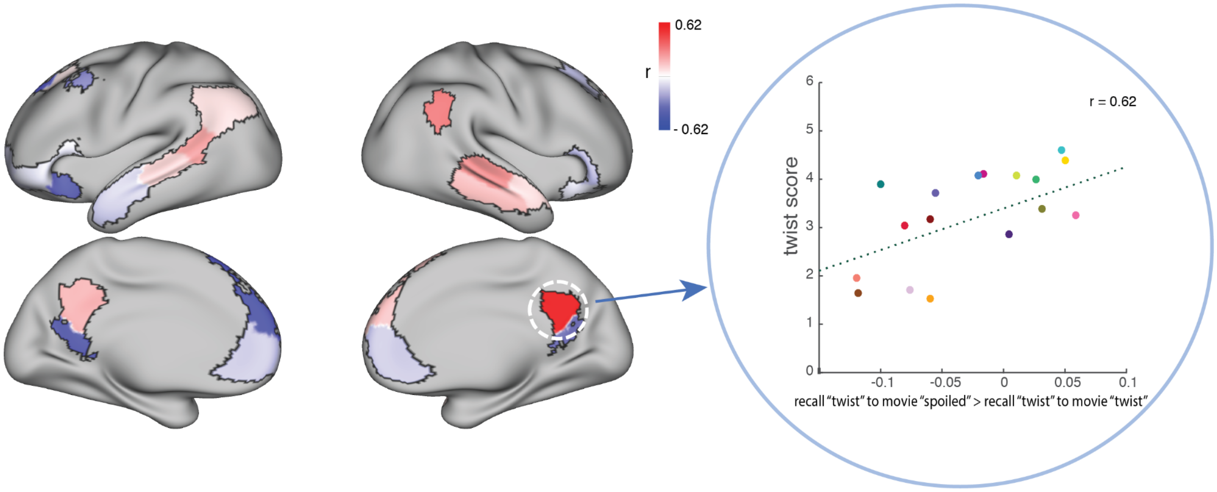
Yeshurun, Y., Nguyen, M., & Hasson, U. (2021). The default mode network: Where the idiosyncratic self meets the shared social world. *Nature Reviews Neuroscience*, *22*(3), 181–192. https://doi.org/10.1038/s41583-020-00420-w

Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C. J., & Hasson, U. (2017). Same story, different story: The neural representation of interpretive frameworks. *Psychological Science*, *28*(3), 307–319. https://doi.org/10.1177/0956797616682029

Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., & Hasson, U. (n.d.). How we transmit memories to other brains: Constructing shared neural representations via communication. *Cerebral Cortex*, 1–13. https://doi.org/10.1093/cercor/bhx202


A) movie-twist (M_D) to movie-no-twist (M_D) > movie-twist (M_D) to movie-spoiled (M_G)
B) recall-twist (R_G) to recall-spoiled (R_G) > recall-twist (R_G) to recall-no-twist (R_D)
C) recall-twist (R_G) to movie-spoiled (M_G) > recall-no-twist (R_D) to movie-spoiled (M_G)
D) recall-twist (R_G) to movie-spoiled (M_G) > recall-twist (R_G) to movie-twist (M_D)
prediction legend
M_G
R_G
M_D
R_D
twist
Δr
0.14
0.04
p < 0.01 NOT corrected

Supplementary Figure 1: This figure depicts the same set of results as in Figure 2 (upper row) and Figure 3 (lower row) in the whole brain (not restricted to DMN). The maps show ROIs with $p < 0.01$ calculated by nonparametric randomization test without correction (areas missing on these maps compared to the original maps had p values greater than 0.01). A) Areas with significantly greater intersubject pattern correlation between groups who encoded the movie with the same interpretation (*Doctor*). B) Areas with significantly greater intersubject pattern correlation between groups who recalled the movie with the same interpretation (*Ghost*). C) Areas where intersubject pattern correlations are significantly greater when comparing updated recall (R_G) to spoiled encoding (M_G) than when comparing naive recalls (R_D) to spoiled encoding (M_G). D) Areas where intersubject pattern correlations between updated recall (R_G) and spoiled encoding (M_G) are greater than between updated recall (R_G) and naive encoding (M_D)



Supplementary Figure 2: The relationship between the behavioral (twist score) and neural (recall "twist" to movie "spoiled" > recall "twist" to movie "twist") measures of memory update in each DMN ROI. The panel on the right depicts the correlation in the precuneus. Each dot is a participant in the "twist" group (N = 19). Note that the example at right was selected post-hoc for high correlation and is not significant after correction for multiple tests.