Biological Sciences - Neuroscience
Social Sciences - Psychological and Cognitive Sciences

Title: Semantic representations in the temporal pole predict false memories

Martin J. Chadwick[1,2*], Raeesa S. Anjum[2], Dharshan Kumaran[1], Daniel L. Schacter[3,4*], Hugo J. Spiers[2†], Demis Hassabis[1†]

Affiliations:

[1]Google DeepMind, London, UK

[2]Institute of Behavioural Neuroscience, Department of Experimental Psychology, Division of Psychology and Language Sciences, University College London, London, UK

[3]Department of Psychology, Harvard University, Cambridge, MA, USA

[4]Center for Brain Science, Harvard University, Cambridge, MA, USA

[*]To whom correspondence may be addressed. Email: mjchadwick@google.com or dls@wjh.harvard.edu

[†]Joint senior authors

Keywords:

False Memory; Semantic; Temporal Pole; fMRI; Pattern Similarity

Abstract

Recent advances in neuroscience have given us unprecedented insight into the neural mechanisms of false memory, showing that artificial memories can be inserted into the memory cells of the hippocampus in a way that is indistinguishable from true memories. However, this alone is not enough to explain how false memories can arise naturally in the course of our daily lives. Cognitive psychology has demonstrated that many instances of false memory, both in the lab and the real world, can be attributed to semantic interference. While previous studies have found that a diverse set of regions show some involvement in semantic false memory, none have revealed the nature of the semantic representations underpinning the phenomenon. Here we use functional MRI (fMRI) with representational similarity analysis to search for a neural code consistent with semantic false memory. We find clear evidence that false memories emerge from a similarity-based neural code in the temporal pole, a region that has been called the "semantic hub" of the brain. We further show that each individual has a partially unique semantic code within the temporal pole, and this unique code can predict idiosyncratic patterns of memory errors. Finally, we show that the same neural code can also predict variation in true memory performance, consistent with an adaptive perspective on false memory. Together, our findings reveal the underlying structure of neural representations of semantic knowledge, and how this semantic structure can both enhance and distort our memories.

Significance Statement

False memories can arise in daily life through a mixture of factors including misinformation and prior conceptual knowledge. This can have serious consequences in settings such as legal eyewitness testimony, which depend on the accuracy of memory. We investigated the brain basis of false memory with functional MRI, and found that patterns of activity in the temporal pole region of the brain can predict false memories. Furthermore, we show that each individual has unique patterns of brain activation that can predict their own idiosyncratic set of false memory errors. Together, these results suggest that the temporal pole may be responsible for the conceptual component of illusory memories.

\body

Each of us has a vast store of semantic knowledge that we apply to incoming sensory data in order to extract meaning from the world around us. Semantic representations are capable of capturing important structural features of the world at many different levels of abstraction, which allows for rapid and flexible responses to a diverse array of environmental challenges. This pre-existing knowledge structure guides ongoing cognition which usually aids performance, but under some circumstances can lead us into error (1–3). A striking example is the widely studied DRM (Deese, Roediger, and McDermott) false memory illusion (4, 5). In a typical DRM task, subjects are asked to memorize a set of words such as “snow”, “winter”, “ice”, and “warm”. After a delay, subjects will typically falsely remember having seeing the semantically related word “cold”. It is widely agreed that this memory

illusion is driven by the semantic relatedness between words contained in the encoding list (e.g. “snow”) and falsely remembered words that were not actually presented (e.g. “cold”). As such, it is thought that each list item automatically, but weakly, activates the semantically related concept (Fig. 1A). This activation leads to memory confusion either through a cumulative priming of the related lure (5, 6), or the encoding of the semantic overlap as a “gist” memory (3), resulting in a false memory unless the error is detected by some internal monitoring process (7). As such, the DRM effect provides a powerful method for investigating both the nature of false memories as well as the structure of semantic knowledge and its effects on cognition.

Despite the well-characterized cognitive mechanisms involved in the DRM effect (7), its neural basis is currently not well understood. Previous neuroimaging and patient studies have provided robust evidence that a core network of regions in the medial and lateral temporal lobe, as well as frontal and parietal regions (8–16) is involved when encoding or retrieving semantic false memories. However, a mechanistic understanding of how these regions generate false memories is lacking. In particular, although it is known that the semantic relatedness between the different words drives the illusion (3, 6, 7), little is known about the neural basis of this semantic relatedness. Computational models of semantic cognition propose that concepts are represented by a similarity-based code in an amodal “semantic hub”, situated in the apex of the ventral processing stream in the temporal pole (17, 18). While other regions such as the temporo-parietal cortex (19) have also been linked

to the representation of abstract conceptual knowledge, the temporal pole is most consistently implicated in both patient and neuroimaging studies (17, 20, 21).

These computational models therefore make clear predictions about the expected neural basis of semantic false memory. Namely, the temporal pole (TP) semantic hub should contain a similarity-based code such that the neural representations of DRM words reflect the known semantic relatedness between those words. Furthermore, the likelihood that a given word list will generate a false memory should be directly related to the degree of neural overlap. This prediction has not previously been investigated, despite the clear implications for understanding both false memory and the structure of semantic knowledge. Here we used functional MRI (fMRI) to measure the neural overlap between DRM lists and related lures, allowing us to directly test this prediction.

Results

We used a representational similarity analysis (RSA) approach, which uses the neural pattern similarity between pairs of stimuli to infer the representational similarity (22). This method is therefore well-suited for assessing neural overlap between semantic representations (23–25), as the degree of overlap should be directly reflected in the representational similarity. We used this approach to measure the degree of neural overlap between each set of DRM words and their related lure word (Fig. 1). Crucially, each DRM list is known to have a different probability of inducing a false memory, with some much greater than others (26,

27). If our prediction is correct, then we should find a brain region displaying a direct correspondence between the degree of neural overlap and false memory likelihood across the different DRM word lists. To measure the neural representations of the DRM word lists, eighteen participants viewed 40 separate four-word DRM lists, along with the 40 associated lure words (Table S1), while we collected fMRI data. While viewing the words, subjects performed an incidental categorization task (manmade or natural) in order to ensure that all words were processed at the semantic level. We used the canonical false recognition scores reported in (27) as our measure of false memory likelihood, and applied a searchlight analysis (28) across the whole brain to establish whether any brain region displayed the predicted positive correlation between neural overlap and false memory likelihood.

This analysis revealed a significant cluster in the left temporal pole (TP), with no other significant information anywhere else in the brain. This result provides evidence that this specific region is responsible for encoding the semantic relatedness between thematically related words (Fig. 2). Furthermore, this result shows that the precise level of neural overlap in the TP predicts the probability that a false memory will be constructed for a given DRM list. This result is therefore fully consistent with the computational accounts of semantic cognition, and demonstrates that a similarity-based code in the TP is capable of generating false memories. Strikingly, our measure of false memory likelihood is a canonical measure taken from an independent set of subjects (26, 27), yet we can nevertheless

successfully predict this information based purely on the neural data of our group of subjects. This result clearly demonstrates a robust level of agreement across different individuals in the neural representation of the concepts contained within the DRM lists. Thus, it appears that the TP is responsible for representing a shared conceptual space, which is a vital component of successful communication. For completeness, we also looked for regions displaying a negative correlation between neural overlap and false memory likelihood, although it is not clear that any such correlation is theoretically meaningful. This analysis revealed a single significant cluster in the right superior frontal gyrus (peak MNI coordinates: 24, 20, 50; Pseudo-t = 4.71; cluster extent = 163 voxels).

In order to ensure that TP neural data are really capturing meaningful semantic representations, we ran an additional set of control analyses based on the functionally defined TP region of interest (ROI). We examined four issues. First, if the neural data are capturing semantic relatedness between the lure and list items, we should find that each lure is more similar to its own list than any other list, regardless of any differences in false memory strength. To assess this hypothesis, we directly compared the neural similarity within and across the 40 DRM sets. As expected, this analysis revealed a significant within-set increase in similarity ($Z=3.11$, $p<0.001$). Second, to ensure that the neural effects were not driven by extraneous factors such as the word frequency or the visual similarity of the lure and list words, we investigated whether either of these factors correlated with the TP neural overlap. Neither of these variables significantly predicted the neural data

(Word Frequency: Z=0.02, p=0.98; Visual Similarity: Z=1.72, p=0.085), suggesting that they are not significant drivers of neural similarity in this region. Third, the task performed in the scanner while subjects viewed the words was a semantic category judgment task (man-made or natural). While the categorical nature of the encoding task was incidental to the effect of interest, it is nevertheless possible that neural representations related to the semantic categories are present in the TP. To investigate this possibility we derived a subject-specific measure of task category similarity between the lure and list of each DRM set based on the pattern of responses to each word (see Methods). We found no evidence for a correlation between this variable and the neural data (Z=0.46, p=0.65), suggesting that task-driven categorical representations are not present in the TP. Finally, we investigated whether the correlation between neural overlap and canonical false memory strength was still present after controlling for the three additional variables (word frequency, visual similarity, and categorical representation). Using a cross-validated ROI approach to avoid issues of statistical circularity (29), we found clear evidence for a significant correlation between neural overlap and canonical false memory even after partialling out the control variables (Z=2.03, p=0.021). Thus, our result cannot be explained by extraneous factors such as word frequency or visual similarity. We further explored each of these three control variables using a searchlight analysis across the whole brain, but none of these analyses revealed any significant results. Given that our measure of neural overlap is in each case based on the average pattern expressed over four list items, this will greatly reduce the power of any analysis that is not explicitly based on some shared representation,

such as the semantic gist. Thus, it is not surprising that these additional analyses did not find any significant results.

While our initial analysis focused on shared semantic representations, it is also likely that each of us forms some idiosyncratic semantic associations through our own individual experience. Such quirks of experience could lead to measurable differences in neural overlap in the TP, and consequently to unique patterns of false memory errors. In order to investigate this possibility, our subjects participated in a DRM false memory recognition task in a separate session that took place several weeks prior to the scanning session. The same 40 DRM word lists were used in both the behavioral and scanning sessions, which allowed us to directly compare each individual's neural data to their behavioral data. As expected, the subjects displayed the typical false memory effect, and committed a large number of high-confidence false alarms to the critical lure stimuli (Fig. 3). As a further quality control check, we explored the consistency of our group's behavioral data compared to the canonical false memory data (26, 27). The group false memory likelihood correlated positively with both the canonical data ($r(39)=0.53$, $p<0.001$) and with the neural data ($Z=1.68$, $p=0.047$), demonstrating that this subject group's data conforms to the canonical data as expected.

However, the key question was whether there might be a subject-specific mapping between the TP neural overlap and the pattern of false memory errors, over and above any shared semantic representations common to all subjects. To assess this

issue we used an individuation analysis (30), comparing the within-subject neural-behavioral correlations (unique semantic information) to the between-subject neural-behavioral correlations (shared semantic information), in each case controlling for the canonical false memory strength to remove additional shared semantic information. This analysis revealed a significantly higher within- than between-subject correlation ($Z=2.63$, $p=0.0042$). This result was still significant after additionally partialling out the influence of the three control variables discussed above ($Z=2.55$, $p=0.0054$). This result provides clear evidence that each individual has a partially unique set of semantic representations within the TP that have a direct impact on memory distortions (Fig. 4). Importantly, such a result cannot be explained by incidental differences in TP physiology or anatomy alone (30), as these more basic properties would not predict each subject's false memory behavior.

Importantly, the fMRI and behavioral data for each subject were collected in separate sessions separated by many weeks, which demonstrates that the structure of neural overlap must be stable over at least this length of time, and plausibly for much longer than this. This long delay also minimized any possible influence of the initial behavioral session on the neural representations expressed during scanning. To further ensure that there was no such influence, we leveraged the wide range of inter-session delay lengths across subjects (min=21 days, max=239 days) that emerged as a consequence of differences in subject availability. If there were an effect of the behavioral session due to memory for the items experienced in this

session, we would expect this effect to degrade over time. We would therefore expect a negative correlation between the length of delay and the strength of neural-behavioral mapping. In fact we find a non-significant positive correlation instead ($r=0.26$, $p=0.30$), which clearly shows that memory is not enhancing the neural overlap data.

The main focus of this study was to investigate the neural basis of false memory. However, an adaptive perspective on false memory (1) would suggest that our semantic knowledge should aid cognition under most circumstances (31), rather than purely acting as a source of memory distortion. This hypothesis would therefore suggest that we ought to also find a positive correlation between neural overlap and true memory performance for the list items that were actually presented during encoding. As predicted, we found a significant mapping between TP neural overlap and true memory strength ($Z=2.33$, $p=0.0099$), which remained significant after partialling out the three control variables ($Z=2.29$, $p=0.011$). We also investigated the possibility that subject-specific TP neural coding might predict true memory performance, using an individuation analysis (30). This analysis demonstrated a significant individuation effect in the TP for true memories ($Z=2.16$, $p=0.016$) as well as false, and this result remained significant after partialling out the three control variables ($Z=2.11$, $p=0.017$). These results provide clear evidence that the semantic similarity code within this region can be beneficial for memory systems, as well as potentially leading to memory distortions.

Discussion

Our study demonstrates that the left temporal pole (TP) contains partially overlapping neural representations of related concepts, and that the extent of this neural overlap directly reflects the degree of semantic similarity between the concepts. Furthermore, the neural overlap between sets of related words predicts the likelihood of making a false memory error. Together, these findings provide support for neural network models of semantic cognition that posit that the TP utilizes a similarity-based coding scheme to sustain amodal distributed representations of individual concepts and their relationships within an abstract semantic space (17, 18). The similarity-based coding of concepts has significant computational advantages in allowing efficient generalization of existing knowledge to novel situations, while still allowing for the "grounding" of each concept in a full set of domain-specific cortical regions mediated by the hub-like connectivity of the TP (17, 18) – consistent with previous studies showing that domain-specific semantic features are indeed distributed across a wide range of cortical regions (32). Nevertheless, our results suggest that the type of coding scheme underpinning the representation of concepts within the TP has a potential cost, specifically the emergent property of false memories.

Although the focus of our experiment was on the structure of semantic representations, it is likely that interactions between regions in the medial temporal lobe (MTL) and the TP are critical to the generation of false memories. While our data cannot speak directly to this issue, there are two clear lines of evidence that are

suggestive. First, results from both rodent (33) and human research (34) demonstrate that false memories can be established as memory "engrams" within the hippocampus through artificial manipulation or reconsolidation processes. Second, recent studies have shown that recognition memory performance can be driven by a dynamic similarity-based computation within the MTL at retrieval (35, 36). It is likely that both sets of processes interact with the semantic representations in the TP in order to produce semantic false memories (3, 6, 37) .

While we found clear evidence for a semantic code within the left TP, other neighboring regions have also been found to contain semantic-like representations – particularly the anterior ventral temporal cortex (23, 25). We suggest that these differing locations are likely due to the level of semantic abstraction, as the set of studies with results located in anterior ventral temporal cortex all used highly concrete stimuli (either concrete words, or direct use of pictures). By contrast, our set of words included many abstract concepts such as "justice" and "desire" (see Table S1 for full list). This suggestion is supported by a meta-analysis that contrasted regions involved in abstract versus concrete words, and that found clear evidence that the temporal pole was more active in response to abstract words, while ventral temporal regions showed a preference for concrete words (20).

Finally, our results show that each individual's unique TP representations predict idiosyncratic patterns of false memory errors. Given that we rely on shared semantic representation in order to communicate with one another, this individual

variation is perhaps surprising. However, it does converge with other recent reports of individual differences in semantic (30) and episodic representation (38), and suggests that divergent personal experience is sufficient to create individually unique representations in higher-level semantic regions. This striking finding suggests that it will be important to further characterize both the shared and individually unique aspects of semantic cognition to better understand the nature of conceptual knowledge.

Methods

Participants

18 participants (11 female) took part in this study. All were right-handed native English speakers, and had normal or corrected-to-normal vision. The study was approved by the University College London research ethics committee, and all participants gave written consent to take part in the study.

Stimuli

For both testing sessions, the stimuli were drawn from forty standard DRM lists (26, 27). Due to time-constraints in the fMRI scanning session, we used only four list items from each DRM list (see Fig. 3B for a comparison with previous results based on the full set of 15 list items). Where possible these were four list items with the highest associative strength with the lure word. However, this full set of forty lists contains some words that were repeated across lists, and some words that we considered to be culturally specific in semantic relatedness to the related concept,

such as “United States” in the “army” list. Any unsuitable list words were therefore excluded, and alternative, lower list associates were included instead. The full set of stimuli used is displayed in the Table S1. In total, the stimulus set included 40 DRM related concept lures, and 160 associated DRM list items. For the behavioral DRM task, we also used an additional 160 unrelated novel words, which were matched to the DRM lists in average concreteness and frequency.

Overall task structure

All participants took part in two separate experimental sessions separated by several weeks (mean 65 days, min=21, max=239). The first session was a behavioral session involving a standard DRM recognition paradigm, providing subject-specific false memory data. The second session was a functional MRI session, where subjects viewed words taken from the DRM lists during an incidental task. Both sessions are explained in more detail below. We elected to run the behavioral session before the scanning session to ensure that repeated exposure to the DRM words during scanning did not impact the behavioral false memory effects. This procedure therefore provided us with a “pure” measure of individual false memories. However, we acknowledge that this design could still have the reverse problem, in that there could be carry-over effects from the behavioral to the fMRI session. The long delay between sessions was built in to minimize any such issues, and further control analyses were conducted to further rule this out as a problem (see Results).

Behavioral DRM task

The first testing session was purely behavioral, and involved a standard DRM recognition paradigm. During an encoding phase, subjects were presented with 40 sets of four-word DRM lists. They were instructed to memorize as many of the presented words as possible. For each list, the four words were presented consecutively for 500ms each, with a 3s interval between each list. The order of the 40 lists was randomized across subjects. Subjects were then required to perform an incidental visual discrimination task for 15 minutes, in order to minimize explicit rehearsal of the list words. Following this distraction period, subjects were given a recognition memory test for the previously presented words. All 160 DRM list words were presented, along with the 40 related concept lure words, and 160 unrelated novel words. This set of 360 words was presented one at a time, in a randomized order. Subjects were required to decide whether they thought the word was old or new, along with a confidence judgment (sure or unsure). The task was self-paced, with no time limit.

Behavioral DRM analysis

The recognition memory test data were analyzed to determine whether the subjects displayed the expected false memory effect. To do this analysis, for each subject we calculated the proportion of words categories as 'Old' for each of the three conditions (Old items, New items, and Lure items). This provided us with a measure of the Hit rate, False Alarm rate, and for the related concept lure items, False Memory rate. To assess whether the expected false memory effects were present, we conducted a series of planned pairwise comparisons between conditions using

two-tailed Wilcoxon signed rank tests. First, we tested for a basic recognition memory effect, by comparing the Hit rate to the False Alarm rate. Next, we investigated whether our subjects displayed the expected false memory effect, by comparing the False Memory rate with the False Alarm rate. Following this, we compared False Memory and False Alarm rate using just high confidence trials in order to determine whether the task induced robust false memories. Finally, we investigated whether the pattern of false memory errors across the 40 DRM lists correlated with the false memory rate reported by (27), despite the fact that we used only the first four list associates from each DRM list, as opposed to the full set of 15. All behavioral results are reported in Fig. 3.

fMRI task

Several weeks later, the same subjects came in for a second testing session in the fMRI scanner. During each of four functional runs, the 40 DRM related concepts, and 160 DRM list words were presented one at a time, for 3s each. Thus, in total, each word was presented 4 times, in order to allow a more stable estimate of the neural pattern. The order of presentation was randomized, with the additional constraint that words from the same DRM set were never presented consecutively. Each subject took part in four functional runs in total. The behavioral task involved a semantic category decision for each presented word. Specifically, subjects had to decide whether they thought each word was more related to the category of 'manmade' or 'natural', and indicate this by pressing the relevant button. As we were simply interested in measuring the neural patterns expressed for each DRM

concept, the task itself was incidental. However, we reasoned that a semantic categorization task would require the subjects to fully process the semantic meaning of each word. The fact that each word would be repeated four times during scanning introduces a possible source of noise due to novelty effects on the first presentation. We therefore allowed the subjects one practice block prior to entering the scanner in order to reduce any novelty signals in the subsequent scanning session.

MRI scan details

It is well known that fMRI of the temporal poles can be problematic due to susceptibility artifacts (signal dropout) in this region, which can substantially reduce BOLD sensitivity (39). We therefore elected to use a 1.5 tesla MRI scanner, which suffers from less pronounced dropout in this region, and therefore can actually have greater BOLD sensitivity than higher field-strength scanners (40). The precise sequences used were further optimized to reduce signal dropout in ventral anterior regions, including the temporal pole. All MRI data was collected using a Siemens Avanto 1.5 tesla MRI scanner with a 32-channel head coil at the Birkbeck-UCL Centre for Neuroimaging (BUCNI) in London. The functional data were acquired using a gradient-echo EPI sequence in an ascending sequence, with a slice thickness of 2mm and a 1mm gap, TR=85ms, TE=50ms, slice tilt=-30°, field of view 192mm, and matrix size 64×64. The whole brain was acquired with 40 slices, leading to a volume acquisition time of 3.4s. The precise slice tilt was chosen as a compromise between sensitivity, coverage, and speed (40, 41). Four functional runs were

collected for each participant. Following functional imaging, an anatomical image was acquired for each participant (T1-weighted FLASH, TR = 12ms, TE = 5.6ms, $1mm^3$ resolution).

fMRI pre-processing

The first six functional volumes were discarded to allow for T1 equilibration. The remaining data were slice-time corrected, and spatially realigned. Each participant's structural image was co-registered to the first functional image. The structural images were segmented, and the deformations estimated during this step were applied to both the structural and functional images in order to normalize them into MNI space. All preprocessing steps were conducted using SPM12. Default parameters were chosen for each step.

Pattern estimation

We were interested in investigating neural overlap between the neural representations of each set of DRM list words and their related concept. Our hypothesis was that each DRM list word should have a neural representation that overlaps with that related concept. In order to assess this hypothesis, we estimated two patterns for each DRM set – one pattern for the related concept itself, and another for all four of the list words combined. This latter pattern captured the neural pattern that was common across all four list items, which should therefore capture the representational overlap. If our hypothesis is correct, then this pattern should correlate with the related concept pattern. To estimate this set of patterns,

we used the GLMdenoise toolbox (42), which implements a denoising step in addition to estimating the beta weights for each regressor. Each pattern was estimated using an event-related regressor indicating the onset of the related concept/set of list words across the four functional sessions. This procedure resulted in a set of 80 beta weight images (40 DRM list patterns and 40 related lure concept patterns). These were converted to t-statistics by dividing the parameter estimate by the estimate of the standard error, thereby normalizing the responses of each voxel (43). The resulting t-statistic images were left unsmoothed to preserve any fine-grained spatial information (44).

Searchlight analysis

We used a searchlight representational similarity analysis (22, 28) to search for brain regions containing the predicted neural code. Representational similarity analysis (RSA) uses the neural pattern similarity between pairs of stimuli to infer the representational similarity. This approach is therefore highly appropriate for assessing neural overlap between semantic representations, as the degree of overlap should be directly reflected in the representational similarity. We first assessed the neural overlap between each DRM related concept and its related set of list words, by measuring the Pearson correlation between the pair of voxel patterns. In each case, we normalized the similarity data by subtracting the mean Pearson correlation between the DRM related concept and each unrelated DRM list pattern. This procedure removed any general effects of similarity that were not driven by semantic relatedness, and resulted in a vector of 40 neural overlap scores for the 40

DRM lists. Our prediction is that the degree of neural overlap within the temporal pole should reflect semantic relatedness, and therefore predict false memory likelihood across the 40 DRM lists. We used the canonical false recognition scores reported in (26, 27) as our measure of false memory likelihood for the 40 DRM lists. We used a searchlight approach (28) to search across the whole brain for regions containing a neural code consistent with our predictions. This approach involves stepping through each voxel in the brain, and in each case running a representational similarity analysis on the cluster of voxels surrounding that central voxel (for all analyses, we used a spherical searchlight with 10mm radius). We used a variation of this approach, where the value at each voxel was the average value of all searchlight analyses that included that voxel. This information-averaging approach more accurately reflects the multivariate nature of the analysis, and results in a smoother image (45). For computational efficiency, we restricted our analysis to a whole-brain gray matter mask, created by averaging the normalized, segmented gray matter images, and applying a threshold of 0.5. This searchlight approach was applied to the analysis described above, using a Fisher-transformed Pearson correlation to assess the mapping between neural overlap and false memory likelihood in each searchlight. This was repeated for all subjects, and statistical significance at each voxel was assessed at the group level using a nonparametric permutation approach (46). This procedure provides a means of applying strict family-wise error correction for multiple comparisons without any parametric assumptions. For this analysis, 10,000 permutations were applied with 10mm variance smoothing, and a standard cluster threshold of Pseudo-t > 3 was

used to assess statistical significance (46). Only regions that are significant at $p<0.05$ with family-wise error correction are reported.

Temporal pole region of interest

In order to further explore the representations contained within the temporal pole, a region-of-interest (ROI) was created based on the initial searchlight results. The ROI included all voxels within the TP that passed the searchlight cluster threshold of $t>3$. The resulting ROI consisted of 92 voxels.

Correlation between canonical false memory and neural overlap

In Fig. 2B we report the correlation between the group average TP neural overlap for each DRM list, and the canonical false memory likelihood (26, 27), in order to illustrate the strength of the relationship. To avoid an artificial inflation of the effect size estimate due to non-independence in the choice of ROI, we used a leave-one-subject out cross-validation approach (29). On a given cross-validation fold, we took the searchlight maps for 17/18 subjects, averaged the maps, and selected the top 200 voxels. We then used these voxels as an ROI to measure the neural overlap in the 40 DRM lists for the remaining subject (note that in this case we did not normalize the neural overlap score by subtracting the between-list correlation, as we consider the raw correlation values to be descriptively more informative). This procedure was repeated 18 times, each time leaving out a different subject. This analysis resulted in neural overlap data for all 18 subjects based on an independently selected ROI, thereby avoiding statistical 'double-dipping' (29). We

averaged the neural overlap across the subjects to create a single summary neural overlap score for each of the 40 DRM lists. This score was then correlated with the canonical false memory scores, as reported in Fig. 2B. An additional correlation was conducted after removing two potential outliers, identified using a bootstrapped Mahalanobis distance, and a threshold of Ds > 6. Note that while this cross-validation approach is guaranteed to provide an unbiased correlation, it is not guaranteed that the voxels used will be based on the same TP region as reported in the searchlight results. We therefore investigated the number of voxels falling with the temporal pole region (defined using the Harvard-Oxford Atlas) for each fold of the cross-validation. Every single fold included at least some voxels within this region, ranging from 5 to 42, with an average of 25.5. Thus a good proportion of the neural information going into this analysis was indeed based on the TP.

False memory individuation analysis

In order to test for the presence of unique TP neural information that predicts subject-specific false memories, we used an individuation analysis (30). The logic here is that if an individual has a unique set of TP neural representations that meaningfully influences cognition, then that individual's neural overlap data should predict their own pattern of false memory errors better than any other subject's false memory errors. To assess this possibility, we created a false memory vector for each individual subject based on their specific pattern of false memory errors in the behavioral DRM session. Given that we were specifically interested in genuine false memories rather than mistakes driven by uncertainty, we defined a false memory as

a high confidence 'old' response to a related concept lure. The false memory vector was in each case a binary vector of 40 values for the 40 DRM lists, with a 1 indicating the DRM lists that results in a false memory, and 0 elsewhere. For each subject, we calculated the Spearman correlation between their TP neural overlap and false memory data (within-subject correlation). We then calculated the Spearman correlation between that subject's neural overlap and each other subject's false memory data, and averaged across these correlation values to provide a summary between-subject correlation. This procedure resulted in a within-subject and between-subject correlation for every subject. To assess whether there was significantly greater within-subject predictive information in the neural data, we compared the within- and between-subject correlations with a Wilcoxon sign rank test. A one-tailed test was used due to our one-sided hypothesis that the correlation should be greater within- than between-subject.

True memory individuation analysis

In order to be consistent with the false memory analysis, we defined a 'true memory' as a high confidence 'old' response to a previously presented DRM list item. The true memory vector was created by calculating for each DRM list, the proportion of words that were judged to be 'old' with high confidence. This resulted in a true memory vector of length 40 for each individual subject. The true memory individuation analysis was otherwise identical to the false memory individuation analysis described above.

Control analyses

To ensure that the results were driven by neural similarity that was specific to the false memory strength and not additional extraneous factors, we conducted three control analyses. First, we established that word frequency (47) was not contributing to the neural data. Given that the neural overlap was based on the similarity between each lure and the respective list items, the absolute word frequency of the critical lure could not by itself explain this measure. Instead, we calculated the average difference in word frequency between each DRM lure and the four DRM list items. Second, we investigated whether visual similarity between the words could be contributing to the effects. We used the Levenshtein edit distance to assess the similarity between each pair of words, as this has been shown to be a good predictor of various lexical effects (48), and is therefore appropriate for assessing low-level word similarity. For each DRM set the visual similarity was defined as the average edit distance between the lure and list items. Finally, we explored whether the incidental task performed in the scanner could be driving the neural similarity results. To investigate this possibility, for each of the 40 DRM lists we quantified the number of list words where the subject had indicated the same category as the related lure concept for that list. The stronger the degree of correspondence, the stronger any task category representation should be for that particular DRM list. Each of these three control variables was correlated with the neural overlap data to determine whether each significantly contributed to the neural data. Additionally, all three variables were controlled for in each analysis reported in the results section.

Acknowledgements and Financial Interests.

H.J.S. is funded by the James S. McDonnell Foundation and The Wellcome Trust. D.L.S. is supported by National Institute of Mental Health Grant MH060941. None of the authors has any competing financial interests.

Author Contributions:

MJC, DK, DLS, HJS, DH all contributed to the experimental design. The data were collected and analyzed by MJC and RSA. The manuscript was written by MJC with major contributions from DK and DH, and with additional input from HJS and DLS.

References

1. Schacter DL, Guerin SA, St Jacques PL (2011) Memory distortion: An adaptive perspective. *Trends Cogn Sci* 15(10):467–74.
2. Qin S, Hermans EJ, van Marle HJF, Fernández G (2012) Understanding low reliability of memories for neutral information encoded under stress: Alterations in memory-related activation in the hippocampus and midbrain. *J Neurosci* 32(12):4032–41.
3. Reyna VF, Corbin JC, Weldon RB, Brainerd CJ (2016) How fuzzy-trace theory predicts true and false memories for words, sentences, and narratives. *J Appl Res Mem Cogn* 5(1):1–9.
4. Deese J (1959) On the prediction of occurrence of particular verbal intrusions in immediate recall. *J Exp Psychol* 58(1):17–22.

5. Roediger HL, McDermott KB (1995) Creating false memories: Remembering words not presented in lists. *J Exp Psychol Learn Mem Cogn* 21(4):803–814.

6. Roediger III HL, Balota DA, Watson JM (2001) Spreading activation and arousal of false memories. *The Nature of Remembering: Essays in Honor of Robert G. Crowder*, eds Roediger III HL, Nairne JS, Neath I, Surprenant AM (APA, Washington, DC, US), pp 95–115.

7. Gallo DA (2010) False memories and fantastic beliefs: 15 years of the DRM illusion. *Mem Cognit* 38(7):833–48.

8. Schacter DL, Verfaellie M, Pradere D (1996) The neuropsychology of memory illusions: False recall and recognition in amnesic patients. *J Mem Lang* 35(2):319–334.

9. Cabeza R, Rao SM, Wagner AD, Mayer AR, Schacter DL (2001) Can medial temporal lobe regions distinguish true from false? An event-related functional MRI study of veridical and illusory recognition memory. *Proc Natl Acad Sci U S A* 98(8):4805–10.

10. McDermott KB, Watson JM, Ojemann JG (2005) Presurgical language mapping. *Curr Dir Psychol Sci* 14(6):291–295.

11. Kim H, Cabeza R (2007) Differential contributions of prefrontal, medial temporal, and sensory-perceptual regions to true and false memory formation. *Cereb Cortex* 17(9):2143–50.

12. Boggio PS, et al. (2009) Temporal lobe cortical electrical stimulation during the encoding and retrieval phase reduces false memories. *PLoS One* 4(3):e4959.

13. Gallate J, Chi R, Ellwood S, Snyder A (2009) Reducing false memories by magnetic pulse stimulation. *Neurosci Lett* 449(3):151–4.

14. Drowos DB, Berryhill M, André JM, Olson IR (2010) True memory, false memory, and subjective recollection deficits after focal parietal lobe lesions. *Neuropsychology* 24(4):465–475.

15. Atkins AS, Reuter-Lorenz PA (2011) Neural mechanisms of semantic interference and false recognition in short-term memory. *Neuroimage* 56(3):1726–34.

16. Warren DE, Jones SH, Duff MC, Tranel D (2014) False recall is reduced by damage to the ventromedial prefrontal cortex: Implications for understanding the neural correlates of schematic memory. *J Neurosci* 34(22):7677–82.

17. Patterson K, Nestor PJ, Rogers TT (2007) Where do you know what you know? The representation of semantic knowledge in the human brain. *Nat Rev Neurosci* 8(12):976–87.

18. McClelland JL, Rogers TT (2003) The parallel distributed processing approach to semantic cognition. *Nat Rev Neurosci* 4(4):310–22.

19. Skipper-Kallal LM, Mirman D, Olson IR (2015) Converging evidence from fMRI and aphasia that the left temporoparietal cortex has an essential role in representing abstract semantic knowledge. *Cortex* 69:104–120.

20. Wang J, Conder JA, Blitzer DN, Shinkareva S V. (2010) Neural representation of abstract and concrete concepts: A meta-analysis of neuroimaging studies. *Hum Brain Mapp* 31(10):1459–1468.

21. Clarke A, Tyler LK (2015) Understanding What We See: How We Derive

Meaning From Vision. *Trends Cogn Sci* 19(11):677–687.

22. Kriegeskorte N, Kievit RA (2013) Representational geometry: Integrating cognition, computation, and the brain. *Trends Cogn Sci* 17(8):401–12.
23. Bruffaerts R, et al. (2013) Similarity of fMRI activity patterns in left perirhinal cortex reflects semantic similarity between words. *J Neurosci* 33(47):18597–607.
24. Devereux BJ, Clarke A, Marouchos A, Tyler LK (2013) Representational Similarity Analysis Reveals Commonalities and Differences in the Semantic Processing of Words and Objects. *J Neurosci* 33(48):18906–18916.
25. Carlson TA, Simmons RA, Kriegeskorte N, Slevc LR (2014) The emergence of semantic meaning in the ventral temporal pathway. *J Cogn Neurosci* 26(1):120–31.
26. Stadler MA, Roediger HL, McDermott KB (1999) Norms for word lists that create false memories. *Mem Cognit* 27(3):494–500.
27. Roediger HL, Watson JM, McDermott KB, Gallo DA (2001) Factors that determine false recall: A multiple regression analysis. *Psychon Bull Rev* 8(3):385–407.
28. Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. *Proc Natl Acad Sci U S A* 103(10):3863–8.
29. Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci* 12(5):535–40.
30. Charest I, Kievit RA, Schmitz TW, Deca D, Kriegeskorte N (2014) Unique

semantic space in the brain of each beholder predicts perceived similarity. *Proc Natl Acad Sci U S A* 111(40):14565–70.

31. van Kesteren MTR, Ruiter DJ, Fernández G, Henson RN (2012) How schema and novelty augment memory formation. *Trends Neurosci* 35(4):211–9.
32. Mitchell TM, et al. (2008) Predicting human brain activity associated with the meanings of nouns. *Science* 320(5880):1191–5.
33. Ramirez S, et al. (2013) Creating a false memory in the hippocampus. *Science* 341(6144):387–391.
34. Edelson M, Sharot T, Dolan RJ, Dudai Y (2011) Following the crowd: Brain substrates of long-term memory conformity. *Science* 333(6038):108–11.
35. LaRocque KF, et al. (2013) Global similarity and pattern separation in the human medial temporal lobe predict subsequent memory. *J Neurosci* 33(13):5466–74.
36. Davis T, Xue G, Love BC, Preston AR, Poldrack R a (2014) Global neural pattern similarity as a common basis for categorization and recognition memory. *J Neurosci* 34(22):7472–84.
37. Arndt J, Hirshman E (1998) True and false recognition in MINERVA2: Explanations from a global matching perspective. *J Mem Lang* 39:371–391.
38. Chadwick MJ, Bonnici HM, Maguire EA (2014) CA3 size predicts the precision of memory recall. *Proc Natl Acad Sci U S A* 111(29):10720–5.
39. Devlin JT, et al. (2000) Susceptibility-induced loss of signal: Comparing PET and fMRI on a semantic task. *Neuroimage* 11(6 Pt 1):589–600.
40. Weiskopf N, Hutton C, Josephs O, Deichmann R (2006) Optimal EPI

parameters for reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at 3 T and 1.5 T. *Neuroimage* 33(2):493–504.

41. Weiskopf N, Hutton C, Josephs O, Turner R, Deichmann R (2007) Optimized EPI for fMRI studies of the orbitofrontal cortex: Compensation of susceptibility-induced gradients in the readout direction. *MAGMA* 20(1):39–49.

42. Kay KN, Rokem A, Winawer J, Dougherty RF, Wandell BA (2013) GLMdenoise: A fast, automated technique for denoising task-based fMRI data. *Front Neurosci* 7:247.

43. Misaki M, Kim Y, Bandettini P a, Kriegeskorte N (2010) Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *Neuroimage* 53(1):103–18.

44. Chadwick MJ, Bonnici HM, Maguire EA (2012) Decoding information in the human hippocampus: A user's guide. *Neuropsychologia* 50(13):3107–21.

45. Björnsdotter M, Rylander K, Wessberg J (2011) A Monte Carlo method for locally multivariate brain mapping. *Neuroimage* 56(2):508–16.

46. Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Hum Brain Mapp* 15(1):1–25.

47. Brysbaert M, New B (2009) Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behav Res Methods* 41(4):977–90.

48. Yarkoni T, Balota D, Yap M (2008) Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychon Bull Rev* 15(5):971–979.

49. Xia M, Wang J, He Y (2013) BrainNet Viewer: a network visualization tool for human brain connectomics. *PLoS One* 8(7):e68910.

50. Morey R (2008) Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutor Quant Methods Psychol* 4:61 – 64.

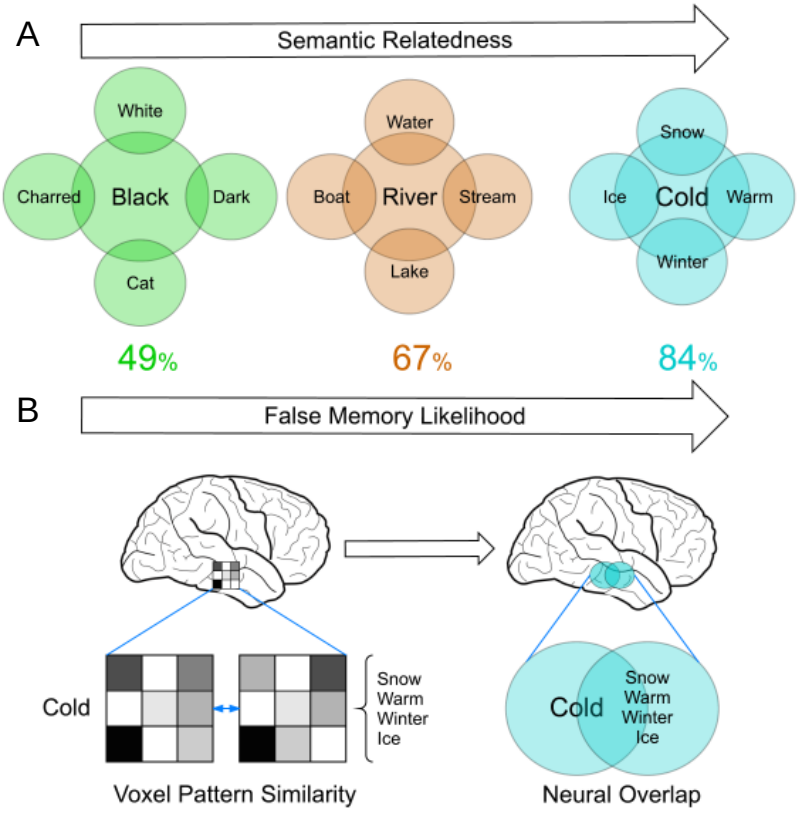

Fig. 1. Neural predictions arising from the DRM false memory illusion. (*A*) For three DRM lists, we illustrate the semantic relatedness between presented list items on the periphery, and the unseen related concept at the center. Beneath each, we show the likelihood that this word list will produce a false memory for the related lure concept. As semantic relatedness increases, so does the false memory likelihood. (*B*) We hypothesize that the neural representation of DRM list items should overlap with the neural representation of the related lure concept. The extent of overlap should directly reflect the semantic relatedness and false memory likelihood of each DRM lure concept. To assess neural overlap for a DRM list we measured the fMRI voxel pattern similarity between the lure concept and the four related list items.

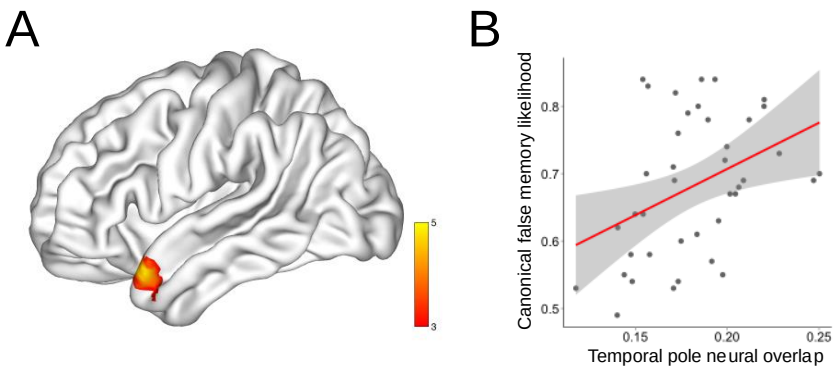

Fig. 2. Neural overlap correlates with canonical false memory likelihood in the left temporal pole (TP). (*A*) A whole-brain searchlight analysis revealed a significant cluster in the left TP (peak MNI coordinates: -51, 17, -25; Pseudo-t = 5.33; cluster extent = 92 voxels), with no other region displaying any significant information. Results are displayed on a cortical surface map using BrainNet Viewer (49). (*B*) To visualize the relationship between neural overlap and false memory, we plot the group average neural overlap for each of the 40 DRM lists against canonical false memory likelihood, using a cross-validation procedure over subjects to avoid artificial inflation of the effect size. There is a clear positive correlation between the two (r(39)=0.40, p=0.012), showing that the degree of semantic relatedness in the neural data predicts variation in false memory strength across the DRM lists. This correlation remains (r(39)=0.45, p=0.005), after removing two potential outliers.

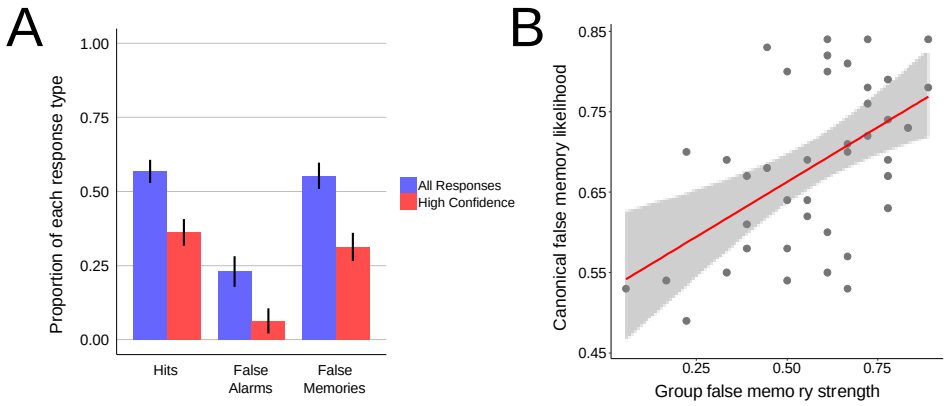

Fig. 3. DRM recognition memory results. (*A*) Group mean hit rate, false alarm rate and false memory rate are displayed for all responses regardless of confidence (blue), and for high confidence responses (red). Error bars represent 95% confidence intervals, adjusted for within-subject data (50). False Memory rates were significantly greater than False Alarm rates ($Z=3.72$, $p<0.001$). This is clear evidence for the expected false memory illusion, which is robust even for high confidence responses ($Z=3.72$, $p<0.001$). (*B*) The group level false memory likelihood across the 40 DRM lists correlated positively ($r(39)=0.48$, $p=0.0016$) with canonical false recognition rates, even after removing one potential outlier ($r(39)=0.48$, $p=0.0018$). All plots are based on a sample size of 18.

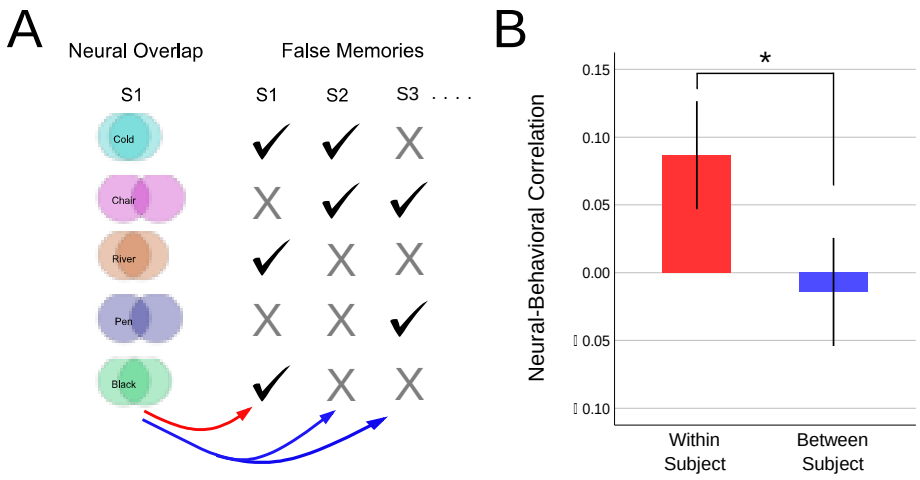

Fig. 4. The temporal pole (TP) contains subject-specific neural information that predicts false memories. (*A*) For each subject, we calculated the correlation between their TP neural overlap, and their pattern of false memories. As a baseline, we calculated the average between-subject correlation. This procedure is illustrated schematically. By comparing the within-subject (red arrow) and average between-subject (blue arrows) correlations we can determine whether there is any subject-specific mapping between the neural and behavioral data. (*B*) The group average within- and between-subject correlations are displayed for the false memory data. Error bars display 95% confidence intervals on a one-way t-test, corrected for within-subjects statistical testing (50). The plots are based on a sample size of 18.