Behavioral/Cognitive

More Is Less: Increased Processing of Unwanted Memories Facilitates Forgetting

Tracy H. Wang,[1] Katerina Placek,[2] and Jarrod A. Lewis-Peacock[1]

[1]Department of Psychology, Institute for Neuroscience, University of Texas at Austin, Austin, Texas 78712-0805, and [2]Department of Neurology, University of Pennsylvania, Philadelphia, Pennsylvania 19104

The intention to forget can produce long-lasting effects. This ability has been linked to suppression of both rehearsal and retrieval of unwanted memories, processes mediated by the prefrontal cortex and hippocampus. Here, we describe an alternative account in which the intention to forget is associated with increased engagement with the unwanted information. We used pattern classifiers to decode human functional magnetic resonance imaging data from a task in which male and female participants viewed a series of pictures and were instructed to remember or forget each one. Pictures followed by a forget instruction elicited higher levels of processing in the ventral temporal cortex compared with those followed by a remember instruction. This boost in processing led to more forgetting, particularly for items that showed moderate (vs weak or strong) activation. This result is consistent with the nonmonotonic plasticity hypothesis, which predicts weakening and forgetting of memories that are moderately activated.

Key words: fMRI; forgetting; intentional forgetting; memory; MVPA; nonmonotonic plasticity hypothesis

Significance Statement

The human brain cannot remember everything. Forgetting has a critical role in curating memories and discarding unwanted information. Intentional forgetting has traditionally been linked to passive processes, such as the withdrawal of sustained attention or a stoppage of memory rehearsal. It has also been linked to active suppression of memory processes during encoding and retrieval. Using functional magnetic resonance imaging and machine-learning methods, we show new evidence that intentional forgetting involves an enhancement of memory processing in the sensory cortex to achieve desired forgetting of recent visual experiences. This enhancement temporarily boosts the activation of the memory representation and renders it vulnerable to disruption via homeostatic regulation. Contrary to intuition, deliberate forgetting may involve more rather than less attention to unwanted information.

Introduction

We forget most of our experiences. This may seem bleak, but memory loss is essential to the human experience; we are bombarded with too much information each moment to preserve everything. Forgetting is an adaptive feature of memory in which unwanted or irrelevant information is discarded to improve access to other memories. Which information should be saved and which discarded? While this challenge is often solved automatically by the brain (Kim et al., 2017), people can also exert volitional control over what they forget (Bjork et al., 1998; Macleod, 1999; Anderson and Green, 2001).

Successful intentional forgetting has been linked to a variety of memory processes at encoding and decision processes at retrieval (Johnson, 1994). Some recent neuroscience research has focused on the active inhibition of unwanted memories during encoding. This research has found that the ability to intentionally forget an unwanted experience engages inhibitory processes in frontal control regions that act to suppress the undesired information (Benoit and Anderson, 2012; Anderson and Hanslmayr, 2014; Hulbert et al., 2016). Successful attempts to intentionally forget new memories are associated with increased activity in the right dorsolateral prefrontal cortex and decreased activity in hippocampus, along with increased functional coupling between these two regions. However, it is unclear how neural representations of memories in the sensory cortex are related to deliberate forgetting. Recent work has shown that, during attempts to re-

Received Aug. 7, 2018; revised Feb. 8, 2019; accepted Feb. 13, 2019.

Author contributions: T.H.W., K.P., and J.A.L.-P. designed research; T.H.W. and K.P. performed research; T.H.W. contributed unpublished reagents/analytic tools; T.H.W. and J.A.L.-P. analyzed data; T.H.W. and J.A.L.-P. wrote the paper.

Research reported in this publication was supported by the National Eye Institute of the National Institutes of Health under Award Number R01EY028746 (J.A.L.-P.) and the National Institute of Neurological Disorders and Stroke under Award Number F32NS096962 (T.H.W.). This content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The authors declare no competing financial interest.

Correspondence should be addressed to Tracy H. Wang at tracy.wang@utexas.edu.

https://doi.org/10.1523/JNEUROSCI.2033-18.2019

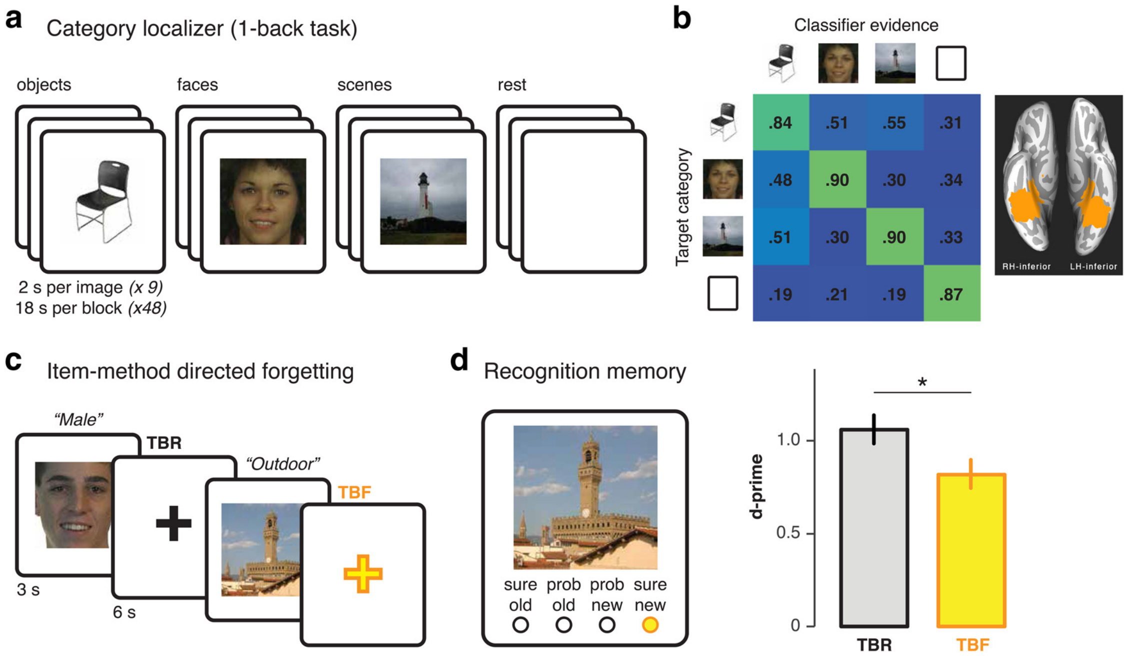

Figure 1. Task procedures, classifier sensitivity, and subsequent memory performance. ***a***, Participants performed a category localizer (1-back task) in the scanner with objects, faces, scenes, and rest. ***b***, Left, Classifier evidence scores (between 0 and 1) for each target category were obtained from cross-validation analysis of fMRI data from the localizer. Right, Visual depiction of VTC ROI used for MVPA classification visualized over standard MNI space from the ventral to dorsal perspective. ***c***, Next, participants performed item-method directed forgetting on faces and scenes in the scanner. They made a subcategory judgment on each picture, and then a cue appeared telling them either to remember (black cross) or to forget (yellow cross) that picture. ***d***, At the end of the experiment, participants were given a recognition memory test for all studied items. The d′ memory scores are based on high-confidence responses. Mean and SEM shown for each condition. **p* = 0.003. Error bars indicate the SEM, *n* = 20.

trieve memories, distributed representations in the sensory cortex of competing memories get suppressed, which contributes to the incidental forgetting of those competing memories (Wimber et al., 2015; Hulbert et al., 2016). It seems reasonable to expect, therefore, that sensory representations of to-be-forgotten (TBF) information might similarly be suppressed during deliberate attempts at forgetting.

In our experiment, we hypothesized that deliberate forgetting is facilitated by the weakening of moderately active memories represented in the sensory cortex, specifically in the ventral temporal cortex (VTC; Rissman and Wagner, 2012; D'Esposito and Postle, 2015; Christophel et al., 2017). This idea follows from the nonmonotonic plasticity hypothesis (NMPH; Norman et al., 2006, 2007; Newman and Norman, 2010; Detre et al., 2013), which proposes a U-shaped relationship between memory activation and learning such that moderate levels of memory activation lead to weakening of the memory, whereas higher levels of activation lead to strengthening. In our prior work, we showed how the NMPH can explain incidental forgetting of items in working memory (Lewis-Peacock and Norman, 2014). When participants did not decisively switch their attentional focus between two items in working memory, the previous item would linger in a state of moderate activation [as measured by pattern classifiers applied to functional magnetic resonance imaging (fMRI) data]. According to the NMPH, these items were susceptible to weakening, and indeed we found that they were associated with worse subsequent memory, compared with trials with relatively less or more activation. Here, we sought to test whether an instruction to deliberately forget an item would leave it in a state of moderate activation, thus making it susceptible to weakening and subsequent forgetting. To test this prediction, we used an item-method directed-forgetting paradigm (Bjork et al., 1998; MacLeod, 1999) in which participants were presented with pictures of faces and scenes with each picture followed by an instruction to remember or forget that picture (Fig. 1*c*). All pictures, regardless of memory instruction, were later presented along with novel pictures in a recognition memory test at the end of the experiment. We hypothesized that participants would change the amount of attention directed to a TBF item, and hence to alter its state of memory activation relative to to-be-remembered (TBR) items (Lewis-Peacock and Postle, 2012; Lewis-Peacock et al., 2012; LaRocque et al., 2014). To quantify and track memory activation, we applied pattern classifiers to human fMRI data in the VTC to measure processing of face and scene items throughout each trial. These neural measurements were contrasted between TBR and TBF trials to assess the impact of attempting to deliberately forget. Finally, trial-by-trial neural measurements were linked to subsequent memory outcomes to evaluate the relationship between memory activation in the sensory cortex and forgetting success.

Materials and Methods

Subjects

Twenty-four healthy subjects between the ages of 18 and 35 were recruited from the University of Texas at Austin student body as well as from the surrounding community in accordance with guidelines of the University of Texas Institutional Review Board. Subjects were compensated at a rate of US$20/h. Informed consent was obtained from all subjects. All subjects were right-handed and had normal or corrected-to-normal vision. Exclusionary criteria included psychiatric disorder, substance abuse, and use of psychotropic medication. During the data-collection phase, two subjects were excluded for sleeping in the scanner and one additional subject was excluded for claustrophobia. Yet another subject was excluded due to a data-storage malfunction. A total of 20 subjects (10 female; mean, 23.6 years old) are included in reported analyses, unless otherwise indicated. An fMRI response box malfunction affected behavioral data recorded for four subjects. As a consequence, in the localizer task, two subjects were not included in the analysis of response latency and accuracy while two others included only accuracy information. For the encoding phase, two subjects were not included in the analysis of encoding task accuracy and response latency, while a third subject contributed only task accuracy information. Critically, these three subjects completed the task, contributed recognition memory task data, and were included in the main analyses.

Stimuli

Experimental materials comprised color pictures of scenes, faces, and objects. A large collection of face stimuli was drawn from a previously published experiment (Lewis-Peacock and Norman, 2014) and their sources (including the NimStim face stimulus database, Tottenham et al., 2009). Faces had neutral expressions, were cropped from the neck down, and shown over a white background. A subset of these faces was chosen based on moderate memorability (2.33–4.10; mean: 3.17; Lewis-Peacock and Norman, 2014). A subset of scenes from the Fine-Grained Image Memorability Dataset was used in the present experiment (Bylinskii et al., 2015). Scenes were chosen by taking images comprising moderate memorability ratings (2.28–4.38; mean: 3.278; scaled from 1 to 5) for the task. Objects were drawn from various on-line sources, including Google Images, cropped to exclude any original background, and displayed over a white background. All items were sized to 300 × 300 pixels and presented using Psychophysics Toolbox Version 3 in Matlab 2014a on an Apple MacBook Air computer running OS X 10.5.

Experimental design and statistical analysis

Procedures

Each subject completed three phases in the experiment: localizer, encoding, and recognition, in that order (Fig. 1*a*,*c*,*d*). The first two phases were administered in the MRI scanner, while the recognition phase was administered outside the scanner ~10–15 min after the encoding phase was completed. In the localizer phase, subjects performed a perceptual localizer task to train fMRI pattern classifiers on categories of scenes, faces, objects, and rest. Subjects performed a one-back task with 18 s miniblocks of items from the three stimulus categories. They also observed 18 s miniblocks of a blank screen with a fixation cross, which served as the baseline "rest" condition. A miniblock consisted of nine items from the same category shown in succession with 8 s in between miniblocks. For each miniblock, one or two items repeated, thus there were 7–8 unique items per miniblock. For each item, subjects were required to respond "not a repeat" with their right index finger button or "repeat" with their right middle finger button. Within the miniblocks, each trial began with the presentation of a single item for 1.5 s, followed by three horizontally aligned fixation crosses for 50 ms. The localizer phase consisted of three localizer runs. Each run included four blocks, and each block included one miniblock of each category type. Across all three runs, the localizer included 90 faces, 90 scenes, 90 objects, and 12 miniblocks of rest. For each participant, stimuli were randomly selected for the localizer and then not presented again. The entire localizer phase lasted ~15 min.

The second phase, the encoding phase, comprised an item-method directed-forgetting task with a random selection of new face and scene images. In this task, subjects were shown either a face or a scene for 3 s. During the presentation of each item, subjects were instructed to give a subcategory judgment. If a face was presented, subjects were to indicate whether the face was that of a male or female. If a scene was presented, subjects were to indicate whether the scene was indoors or outdoors. Following the presentation of the item, an instruction cue was given for 6 s in the form of a yellow fixation cross ("forget," TBF) or black fixation cross ("remember," TBR) presented in the center of the screen. Unlike the task diagram in Figure 1*c*, the actual size of this fixation cross in the experiment was quite small—occupying an area of only 24 × 24 pts, or 0.07% of the total screen size from a projector configured at a resolution of 1024 × 768 pts. Subjects were instructed to apply the instruction represented by the cue to the preceding item. Subjects were not encouraged to use any particular strategy. Rather, they were instructed to simply forget or remember the previously presented item. Importantly, critical TBF trials were always preceded by a TBR trial of the opposing category (e.g., if a face was presented on a TBF trial, a scene preceded it on a TBR trial) so that our category-specific pattern classification analyses could distinguish trial-specific memory processing (see Multivariate). To discourage subjects from anticipating the forget instruction, we included 60 additional TBR trials distributed across the experiment to precede other TBR trials. Therefore, there were instances in which a TBR trial was followed by another TBR trial, but a forget trial was never followed by another forget trial. None of the participants reported any explicit awareness of any predictable sequence of stimulus category or memory instruction. Note that for our analyses comparing the trial conditions, we excluded data from these additional TBR trials. This allowed equal sampling of trials from both TBR and TBF conditions (96 trials each). Supporting analysis found no differences in subsequent memory ($t_{(19)} = 1.01$, $p = 0.325$) or classification of the neural data for the 60 TBR trials versus the 96 TBR trials used in the main analysis (data not shown). Each of six runs of the directed-forgetting phase included 42 trials (16 TBF, 26 TBR; 21 faces, 21 scenes) and lasted ~6 min. Across all runs, there were 252 total trials (96 TBF, 156 TBR; 126 faces and 126 scenes), which lasted ~38 min.

The third phase of the experiment was a self-paced recognition test conducted outside the scanner. Subjects performed a recognition memory task on a large set of 504 items, which included 252 items from the study task (half faces, half scenes) and 252 new items. Subjects were asked to give confidence judgments ("definitely old," "probably old," "probably new," or "definitely new") to each item presented at test. To encourage recognition responses that reflect actual memory of the items, and to discourage responses that reflected the instruction cue given at study (e.g., to discourage a "definitely old" response being given to an item for which the subject remembers being told to forget), confidence responses were assigned points. Subjects were informed of the point system, instructed to maximize their points, and the total point sum was reported to the subject at the end of the test. The point system was as follows: for each old item, a new response ("probably new" or "definitely new") was penalized with −1 point while an old response ("probably old" or "definitely old") was awarded +1 point. For each new item, an old response ("probably old" or "definitely old") was penalized −1 point while a new response ("probably new" or "definitely new") was awarded +1 point. Practice items for both localizer and study items were administered before the scan session. Test items were not practiced.

Subsequent memory analysis

We calculated subsequent memory sensitivity using d′, treating only "high-confidence old" responses as hits. This metric accounts for both hit rates and false-alarm rates.

fMRI data acquisition

Functional and anatomical MRI data were acquired on a 3 T MRI scanner (Magnetom Skyra, Siemens AG) equipped with a 32-channel parallel imaging head coil. Functional scans were acquired with a T2*-weighted echo-planar image (EPI) sequence with the following parameters: TR = 1 s; TE = 30 ms; flip angle, 63°; 2.4 mm slices; no gap; matrix size, 110 × 110; field of view (FOV), 230 mm; 56 oblique axial slices; multiband acceleration factor, 4. Slices were acquired in interleaved order. Automatic high-order shim was used to orient acquisition parallel to the anterior commissure–posterior commissure line for full coverage of the brain with limited coverage of the cerebellum. Data were acquired for both localizer and study phases, while the test phase was acquired outside the scanner. High-resolution T1-weighted anatomical images were acquired for all subjects using a three-dimensional magnetization-prepared rapid acquisition gradient echo (MPRAGE) pulse sequence (TR = 1.9 s; TE = 2.43 ms; flip angle, 9°; FOV, 256 mm; matrix size, 256 × 256; voxel size, 1 mm^3; 192 slices; sagittal acquisition).

fMRI data analysis

Univariate. Functional EPI images were preprocessed and analyzed using SPM12 (http://www.fil.ion.ucl.ac.uk/spm/) implemented under Matlab R2014a. EPI images were spatially aligned to the mean volume and reoriented parallel to the anterior-to-posterior commissure plane before normalization. All volumes were normalized to the Montreal Neurological Institute (MNI) template EPI* brain and further smoothed 6 mm full-width at half-maximal. We implemented a mass univariate, general linear model (GLM) analysis primarily to confirm the presence of directed-forgetting effects found in previous experiments that used the item-method directed-forgetting paradigm (Wylie et al., 2008; Rizio and Dennis, 2013). We implemented a two-stage mixed-effects model by first convolving the onset of each TBR and TBF instruction with a canonical hemodynamic response function with its temporal and dispersion derivatives. Stimulus onsets were not modeled due to their consistent temporal proximity with the instruction. An alternative model including these regressors was created, and the results were not qualitatively different from those reported here. In the first stage, we used the subsequent memory procedure to sort trials from study into items subsequently forgotten or subsequently remembered. Further, we segregated these items into those previously presented with a TBF or a TBR instruction. In the second stage, we carried these effects of interest forward into a random-effects analysis. We were interested in two primary comparisons: (1) successful forgetting effects: regions demonstrating greater activity for subsequently forgotten TBF items than for subsequently remembered TBR items; (2) successful remembering effects: regions demonstrating greater activity for subsequently remembered TBR items than for subsequently forgotten TBF items. All effects are reported with an uncorrected threshold of $p < 0.001$ (one-tailed) with a cluster extent threshold of 237 voxels determined by Monte-Carlo simulations to control for type-1 errors using 3dClustSim (Cox, 1996) with the mixed-model autocorrelation function option (Cox et al., 2017) to account for the noise smoothness structure.

Multivariate. For multivoxel pattern analyses (MVPA; Lewis-Peacock et al., 2014), functional EPI images were preprocessed and analyzed using FSL [FMRIB (fMRI of the Brain) Software Library] 5.0 (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/) subroutines implemented under Matlab R2014a. Functional images were realigned to the middle volume of the middle (fifth overall) run to correct for motion, and high-pass filtered (128 s) to eliminate slow drift. All MVPA analyses were done in native space for each participant [using the Princeton MVPA toolbox (https://github.com/PrincetonUniversity/princeton-mvpa-toolbox) and subsequent analysis in custom code in Matlab R2014a].

All MVPA analyses were conducted within an anatomical ventral temporal mask for each participant. We focused our classifiers on activity in the VTC, a region that serves as input to convergence zones (for example, in the medial temporal lobes) responsible for storing long-term memories (Lavenex and Amaral, 2000); this enabled us to treat our scene and face classifier evidence scores as reflecting the strength of the excitatory inputs into memory regions. The ventral temporal mask (in MNI space) was defined using boundaries delineated by Grill-Spector and Weiner (2014) and created by merging the temporal fusiform cortex, parahippocampal gyrus, occipital fusiform gyrus, and temporal occipital fusiform cortex regions from the Harvard–Oxford atlas (Frazier et al., 2005; Desikan et al., 2006; Makris et al., 2006) found in FSL 5.0. To create subject-specific masks, we coregistered EPI volumes to their own MPRAGE structural volume using FSL FMRIB's Linear Image Registration Tool (Jenkinson and Smith, 2001; Jenkinson et al., 2002). We then used FSL FMRIB's Nonlinear Image Registration Tool to register structural volumes to MNI space. Individual, native-space ventral temporal masks were created by combining (with the registration parameters for the MPRAGE) and applying a reversed transformation matrix from EPI to MNI stereotaxic space on the ventral temporal mask described above.

We used MVPA to quantify the degree of face and scene category-specific neural activity associated with items on TBF and TBR trials in the VTC (as already defined). To ensure accurate decoding of face and scene categories, we trained four binary L2-penalized logistic regression classifiers (with a penalty of 50) on faces, scenes, objects, and rest-related activity from the category localizer task. Each of these one-versus-other classifiers produced a class evidence score for the class on which it was trained. Therefore, the four evidence values did not need to sum to one. For each miniblock, we trained and tested the classifier on the preprocessed BOLD data from the 18 TRs after the onset of the first item. We shifted regressors by 5 s to account for hemodynamic delay. Classifier training consisted of using the leave-one-run-out method on the three localizer runs, in which the classifier trains on one run, and tests on the two others, rotating through until all runs were tested. Figure 1*b*, which shows the mean classifier evidence for each category, demonstrates that the classifiers have sufficient sensitivity to discriminate each category of interest.

To decode the directed-forgetting task for each participant, we trained classifiers on all localizer data (separately for each participant) and applied them to each TR of the TBR and TBF trials. Here, we produced classification evidence output scores for each 1 s TR after the trial onset (uncorrected for hemodynamic delay). From the decoded classifier evidence at each time point in each trial, we calculated "target" and "nontarget" evidence by appropriately relabeling the data (e.g., "face" evidence became target evidence, and "scene" evidence became nontarget evidence on a face trial). Finally, we calculated the differences between target and nontarget evidence to reflect the relative balance of trial-relevant and trial-irrelevant processing at every time point.

Critically, each TBF item was followed by a TBR item of the opposite category (e.g., a TBF face followed by a TBR scene). Meanwhile, TBR items could be followed by either type of item, including a TBR item from the same category. These trials (38.4% of all TBR trials) were not included in any of the analyses to ensure that all items were preceded by an item that was given an opposite instruction (TBF or TBR) and was drawn from the opposite category (scene or face). This procedure enabled us to cleanly differentiate face and scene categories in relation to their memory instruction, thus serving as a proxy for neural activity associated with each TBR and TBF item.

Representational similarity analysis. We performed representational similarity analysis (RSA; Kriegeskorte et al., 2008) that compared each item's activation pattern before and after the memory instruction in the encoding phase to a "category template pattern" for faces and scenes from the localizer phase. For faces, a face-sensitive region (mean, 347 voxels; SEM, 77.6 voxels) was identified with the GLM contrast (face > all other categories) in the localizer data, masked by the anatomically defined VTC region. A scene-sensitive region (mean, 210 voxels; SEM, 47.0 voxels) was identified by a similar GLM contrast (scenes > all other categories) masked by the same VTC region. Contrasts were thresholded at $p < 0.005$, corrected for familywise error rate, unless voxel counts were <100 for an individual. For these subjects ($N = 3$ each for face and scene), more liberal thresholding ($p < 0.05$ to $p < 0.001$ uncorrected) was used to identify ≥ 100 voxels for the contrast.

From these two category-selective regions of interest (ROIs), we computed a category template pattern for faces and for scenes by averaging, for each voxel, the data from the category-specific trials in the localizer data (e.g., face trials for the face template). Then, using these same ROIs, we extracted item-specific activity patterns for each trial in the encoding phase. For each trial, we computed a "baseline" pattern from the first 3 TRs (3 s) of the trial, and an "encoding" pattern from 3 TRs after the memory instruction (6–8 s after trial onset). Both of these trial-unique patterns were compared with the appropriate category template using Pearson correlation, and all correlations were Fisher's *z*-transformed before statistical analysis.

Relating classifier evidence to subsequent memory performance

We used the Probabilistic Curve Induction and Testing Toolbox (P-CIT; Detre et al., 2013; Lewis-Peacock and Norman, 2014; https://code.google.com/p/p-cit-toolbox) developed in Matlab which uses a Bayesian curve-fitting algorithm to estimate the shape of a "plasticity curve" relating neural data (category-specific classifier evidence) and behavioral data (recognition memory confidence scores). The P-CIT algorithm approximates the posterior distribution over plasticity curves (that is, which curves are most probable, given the neural and behavioral data). P-CIT generates this approximation by randomly sampling curves (piecewise-linear curves with three segments) and then assigning each curve an importance weight that quantifies how well the curve explains the observed relationship between the neural and behavioral data. Finally, these importance weights are used to compute the probability of each curve. To assess evidence for the NMPH, P-CIT labels each sampled curve as theory consistent (if it shows a U shape, dropping below its starting point and then rising above its minimum value) or theory inconsistent, and

Table 1. Subsequent memory for TBR and TBF items by stimuli category

	Hits		Misses		False Alarms		Correct Rejections	
Confidence: proportion (SEM)	High	Low	High	Low	High	Low	High	Low
TBR faces	0.26 (0.03)	0.35 (0.03)	0.08 (0.02)	0.31 (0.03)				
TBF faces	0.21 (0.02)	0.40 (0.03)	0.07 (0.02)	0.32 (0.04)	0.07 (0.01)	0.30 (0.03)	0.16 (0.03)	0.47 (0.04)
TBR scenes	0.34 (0.05)	0.24 (0.02)	0.13 (0.02)	0.29 (0.04)				
TBF scenes	0.27 (0.04)	0.26 (0.02)	0.15 (0.02)	0.32 (0.04)	0.07 (0.01)	0.19 (0.03)	0.29 (0.04)	0.46 (0.03)

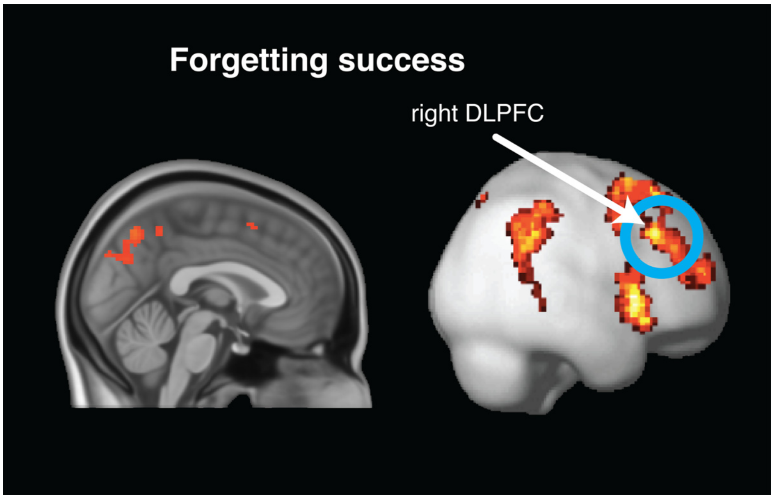

Figure 2. Univariate results. GLM results for forgetting success (greater activity for successful intentional forgetting relative to successful intentional remembering, $p < 0.001$, $k = 237$). See Table 2 for complete univariate results. DLPFC, Dorsolateral prefrontal cortex.

then computes a log Bayes factor score that represents the log ratio of evidence for versus against the NMPH; positive values of this score indicate a balance of evidence in support of nonmonotonic plasticity. P-CIT also computes a χ^2 test that assesses how well the curve explains the data overall, regardless of its shape; the P value for this χ^2 test indicates the probability of obtaining the observed level of predictive accuracy, under a null model where classifier evidence is unrelated to memory behavior.

In this analysis, we used a "super-subject" procedure in which each participant contributed 96 trials for a total of 1824 trials for each TBF and TBR instruction condition. We used this fixed-effects analysis because data from each individual subject were insufficient for random-effects analysis (Lewis-Peacock and Norman, 2014) across subjects. Additionally, for this application of P-CIT, we treated preinstruction (and postitem onset, 1–3 s) and postinstruction (4–9 s) time intervals as separate events with distinctive processes (perceptual encoding vs mnemonic processing). This approach to modeling each trial with two neural data points uses the "net effects" procedure (Lewis-Peacock and Norman, 2014) to sum their individual contributions to the single behavioral outcome of remembered or forgotten (see the P-CIT manual for further details). To evaluate the reliability of these results, we also implemented a bootstrap resampling procedure (Efron, 1979) with 1000 iterations.

Visualization of results

GLM surface results are visualized by the SPM12 canonical render. All GLM subcortical findings are visualized over the FSL MNI T1 − 1 mm anatomical standard.

Data and code availability

The data from this study will be made available upon reasonable request. All code and stimuli used in this study can be found in a public GitHub repository (LewisPeacockLab/imdif).

Results

Behavioral results

The perceptual localizer consisted of a one-back task on mini-blocks of same-category items (i.e., face, scene, object, and rest) for category-level decoding. We obtained accuracy (successfully identifying a repeated image) and response latency behavioral performance measures. Outlier trials for which response latencies were >3 SDs from each subject's mean were removed from the analysis (2.2% of all trials). Accuracy on the one-back task was at ceiling for faces (97.7%; SEM, 0.5%), scenes (98.2%; SEM, 0.4%), and objects (98.5%; SEM, 0.4%). Rest trials did not require a response. One-way ANOVA revealed no accuracy differences (p = 0.530) between these three categories. Further, a one-way ANOVA test of reaction times across faces (593 ms; SEM, 25 ms), scenes (614 ms; SEM, 33 ms), and objects (600 ms; SEM, 28 ms) revealed no differences between these categories (p = 0.871).

Subcategory identification accuracy in the directed-forgetting task was high for both faces (97.7%, SEM 0.4%) and scenes (98.0%, SEM 0.4%), with no significant differences between them (p = 0.570, two-tailed paired t test). Participants responded well within the 3 s response deadline, and were faster to identify faces (1061 ms; SEM 42 ms) than scenes (1236 ms; SEM 45 ms; $p < 0.001$, two-tailed paired t test).

Performance on the subsequent memory test of these items is shown in Figure 1*d* and Table 1. For all subsequent memory analyses of old items described below, we treated recognition responses as a graded measure of memory strength (sure old, 1; probably old, 0.667; probably new, 0.333; sure new, 0), in which the "old" responses corresponded to remembered items and "new" responses corresponded to forgotten items. The proportion of old responses for TBR trials was 0.594 (SEM, 0.158; for high-confidence only: 0.297; SEM, 0.163), and for TBF trials was 0.434 (SEM, 0.162; for high-confidence only: 0.110; SEM, 0.060). The forgetting effect (based on d′ scores) depicted in Figure 1*d* was significant for both scenes (p = 0.001) and faces (p = 0.028) with no difference between categories (p = 0.281).

Neural measures of directed forgetting

We conducted univariate fMRI analyses based on the GLM to contrast brain regions engaged for TBF items that were forgotten and TBR items that were remembered (see Materials and Methods). Consistent with prior work (Wylie et al., 2008; Rizio and Dennis, 2013), we found increased activity for successful forgetting in the dorsolateral prefrontal cortex, posterior cingulate, and precuneus (Fig. 2). For a detailed report of the univariate results, please refer to Table 2. Together with the behavioral results reported above, these data confirm that our experiment is producing directed-forgetting univariate results consistent

Table 2. GLM regions identified during item-method intentional forgetting by study memory instruction and subsequent memory outcome[a]

Successful forgetting					
MNI coordinates			Cluster size	*z*-value	Brodmann's area number
49	−38	50	567	5.14	39/40
25	8	55	1324	5.02	6/9
42	13	−3	554	4.59	13/44
−4	−66	50	877	4.63	7

[a]Regions with greater activity for subsequently forgotten items after forget-cue than for subsequently remembered items after remember-cue with a threshold of $p < 0.001$ and cluster size of 237.

with prior findings, allowing for interpretation of multivariate analysis outcomes.

Measuring memory processing with multivariate fMRI analyses

To assess the degree of memory processing on each trial, we applied pattern classifiers to the fMRI data (Rissman and Wagner, 2012; Lewis-Peacock et al., 2014; D'Esposito and Postle, 2015) in the VTC. Group-averaged results for the classifiers, trained separately for each participant's localizer data, are shown in Figure 1*b*. The classifier confusion matrix shows the mean classifier evidence for all categories (columns) on localizer blocks featuring stimuli from one target category (rows). The cross-validation procedure used to evaluate classifier performance entailed training a classifier on two runs of data and then applying that classifier to independent data from the held-out third run; the runs were then rotated and this procedure was repeated until all three runs had been tested. Decoding accuracy for all categories was well above chance (25%). Face evidence was reliably higher than scene evidence for face blocks, and vice versa (both P's < 0.001), but face and scene scores were not dissociable during rest periods ($p = 0.650$). These analyses were repeated in the bilateral hippocampus and cuneus (two regions found in the univariate analyses to be differentially active for TBF and TBR conditions), but the classifiers were not sufficiently sensitive to category differences in these regions. To analyze data from the memory task, we applied classifiers retrained on all localizer data, again separately for each participant. For each trial, we computed a "target–nontarget" classifier evidence score, which reflects the relative balance between trial-relevant processing and trial-irrelevant processing (e.g., face evidence minus scene evidence on a face trial). These neural measures served as a proxy for processing of category information for the observed item.

Averaged across trials, the classification results show that evidence for the memory item was higher after a TBF instruction compared with after a TBR instruction (between 6 and 8 s after stimulus onset, two-tailed t test, $p = 6.789\text{e-}8$; Fig. 3*a*). This result, found for both face trials ($p = 2.293\text{e-}8$) and scene trials ($p = 2.680\text{e-}5$), suggests that there was stronger processing of TBF items. To evaluate representational changes as a function of memory instruction (Fawcett et al., 2016), we conducted an RSA (see Materials and Methods), in which we computed the pattern similarity between the multivoxel representation of the memory item on each trial to the average category-specific pattern of voxel activity (the category template) from the localizer data. Before the memory instruction, there was no difference in pattern similarity for TBF items versus TBR items (paired t test, $p = 0.365$). However, after the memory instruction, the TBF items were less similar to (i.e., more distinct from) the category template than were TBR items (paired t test, $p = 2.96 \times 10^{-11}$; Fig. 3*b*). Together, the across-category pattern classification analysis and the within-category pattern similarity analysis provide complementary insights into the neural consequences of item-method directed forgetting: the intention to forget leads to increased processing of the item (characterized by higher classifier evidence for the item's category), which in turn is associated with a more distinct representation of the item (characterized by lower representational similarity to the item's category template).

This result is inconsistent with a prominent view that item-method directed forgetting of TBF items results from stronger encoding (e.g., selective rehearsal) of TBR items (Gelfand and Bjork, 1985; Basden et al., 1993; Basden and Basden, 1998). Instead, it agrees with more recent reports that deliberate forgetting is associated with effortful processing following an instruction to forget (Pastötter and Bäuml, 2007; Fawcett and Taylor, 2008; Fawcett et al., 2013). Forgetting effects have been linked to a decrease in memory sensitivity for TBF items, as we found here, rather than outright forgetting of those items. For example, Zwissler and colleagues (2015) found that forget instructions result in active processing that reduces the false-alarm rate but does not impair memory beyond an uncued baseline where only incidental encoding occurs. Here, our central hypothesis is that the degree of memory processing after a forget instruction will predict the degree of forgetting success for that item (Norman et al., 2006; Newman and Norman, 2010; Detre et al., 2013; Lewis-Peacock and Norman, 2014). We now address this hypothesis by linking the neural measures of memory processing during directed forgetting with the behavior measures of memory sensitivity from the recognition test at the end of the experiment.

Relating classifier evidence to subsequent memory

We hypothesized that across items, there would be a nonmonotonic (U-shaped) relationship (Newman and Norman, 2010; Detre et al., 2013) between target–nontarget classifier evidence for TBF items and subsequent recognition memory for those items at the end of the experiment. To formally test for the nonmonotonic pattern in these data, we used the P-CIT (Detre et al., 2013; Lewis-Peacock and Norman, 2014) Bayesian curve-fitting algorithm to estimate the shape of the plasticity curve relating post-instruction memory processing in the directed-forgetting task (indexed by classifier evidence) and subsequent memory performance (indexed by recognition confidence).

For our P-CIT analyses, the pre-instruction interval and the post-instruction interval (for each item) were treated as separate learning events whose effects were summed to model recognition of that item. The fitted curves shown in Figure 3*c* explained a significant amount of variance in subsequent recognition outcomes on both TBF trials ($\chi^2 = 21.34$, $p = 3.840\text{e-}6$) and TBR trials ($\chi^2 = 56.6$, $p = 5.274\text{e-}14$). We found a U-shaped mapping between classifier evidence scores and subsequent memory outcomes such that moderate levels of target–nontarget evidence were associated with worse subsequent memory than higher and lower levels of target–nontarget evidence in TBR trials. This outcome is consistent with the findings of Lewis-Peacock and Norman (2014), who showed incidental forgetting of moderately active items in working memory following encoding. The curves recovered by P-CIT on TBF trials also revealed a U-shaped mapping between classifier evidence scores and subsequent memory outcomes. That is, deliberate forgetting was most successful when the TBF item's memory activation was sufficiently enhanced (but not too high) so as to produce moderate levels of activity during the forgetting attempt. This result also held when using raw classifier evidence scores for the target category were used to quantify memory processing on each trial (e.g., "face" evidence instead of

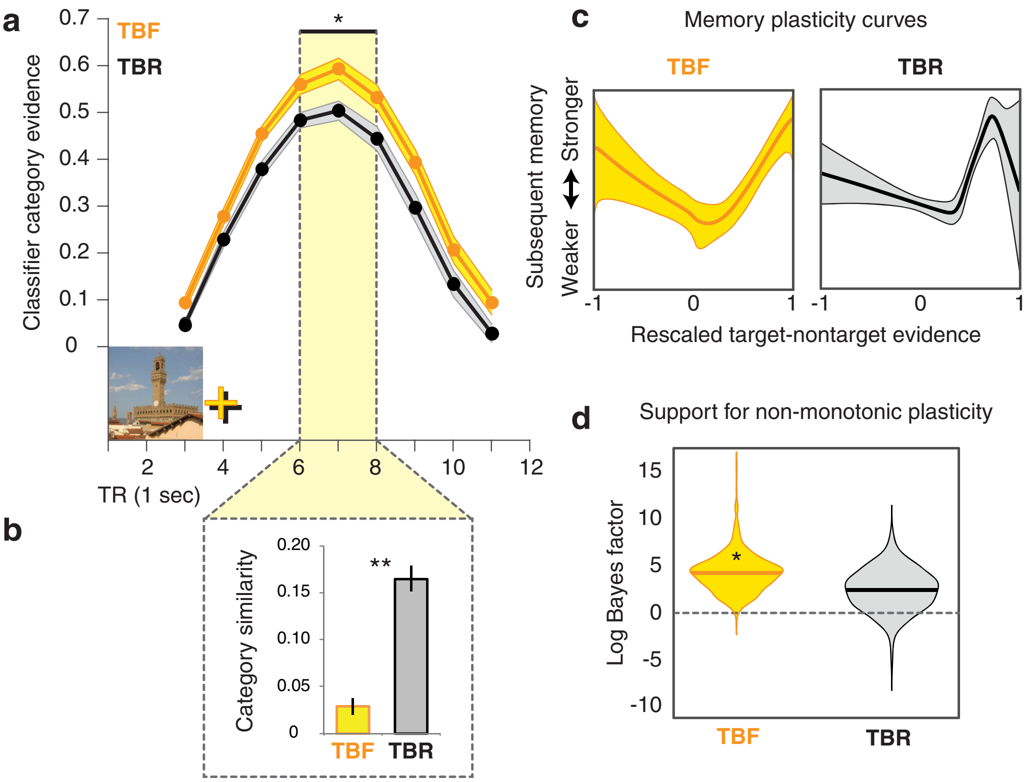

Figure 3. Pattern classification of fMRI data from directed-forgetting task. ***a***, Target–nontarget category classifier evidence for TBF (yellow) and TBR (black) trials. Classifier evidence scores were not shifted to account for hemodynamic lag (ribbon thickness indicates SEM across participants, $n = 20$; $*p = 6.789\text{e-}8$). ***b***, Category similarity analysis showing RSA correlations (Fisher's z-transformed) which compare trial-specific activity patterns to average category-specific activity patterns from the localizer in category-selective voxels in the VTC. $**P = 2.96 \times 10^{-11}$. ***c***, Empirically derived estimates (generated using the Bayesian P-CIT algorithm) of the plasticity curve characterizing the relationship between target–nontarget classifier evidence and subsequent memory performance (recognition confidence). Behavioral outcomes are modeled as depending on the summed effects of pre-instruction (1–3 s) and post-instruction (4–9 s) classifier evidence. Within each box, the line shows the mean of the posterior distribution over curves, and the ribbon shows the 90% credible interval (such that 90% of the curve probability mass lies within the ribbon). The horizontal axis shows target–nontarget classifier evidence scores rescaled so that the minimum classifier evidence value equals −1 and the maximum classifier evidence value equals 1; the vertical axis represents the subsequent memory strength. ***d***, Violin plots describing the balance of evidence (operationalized in terms of log Bayes factor) in favor of the NMPH, shown separately for the two conditions. These plots show the probability density (using kernel density estimation) of the log Bayes factor derived from 1000 bootstrap iterations. The thick marker inside each plot indicates the mean. Positive values of the log Bayes factor correspond to evidence in favor of the NMPH and negative values correspond to evidence against the hypothesis. ($*p = 0.019$, nonparametric bootstrap analysis, 1000 iterations).

"face–scene" evidence on a face trial). This suggests that deliberate forgetting does not require competition per se between two or more items in memory, but rather depends on moderate activity of the targeted memory alone. This result is predicted by the NMPH, which links moderate activation with memory weakening (Newman and Norman, 2010). Note that competition between memories is one way to achieve moderate activation, but it is not required (Newman and Norman, 2010; Detre et al., 2013; Poppenk and Norman, 2014).

To assess the population-level reliability of the U-shaped curve (that is, to determine whether the results were driven by a small subset of participants), we also ran a bootstrap resampling test in which we resampled data from participants with replacement and recomputed the log Bayes factor for the resampled data (Efron, 1979). For TBF trials, 98% of these bootstrap samples (out of 1000 total) showed evidence in support of the NMPH (that is, a positive log Bayes factor), thereby indicating a high degree of population-level reliability in the shape of the curve (Fig. 3*d*). There was less population-level reliability in the shape of the curve for TBR trials (only 81% of 1000 bootstrap samples showed evidence in support of a U-shaped curve). A final P-CIT analysis that combined all trials, regardless of memory instruction, found a reliable U-shaped memory plasticity curve ($\chi^2 = 49.6$, $p = 1.875\text{e-}12$) that held across 94% of bootstrap samples.

In summary, this analysis linking neural data with behavior found a U-shaped relationship between memory activation strength (as measured by fMRI pattern classifiers) and subsequent recognition memory for images of faces and scenes in an item-method directed-forgetting paradigm, with greater processing and more forgetting associated with motivated forgetting trials.

Discussion

Here, we applied machine-learning methods to human fMRI data to reveal a novel mechanism involved in intentional forgetting: the weakening of moderately active representations of TBF items in the VTC. Compared with the intention to remember, intention to forget is associated with higher MVPA classifier evidence and worse subsequent memory. While an instruction to remember an item may not demand much additional processing after a sufficiently long encoding period (3 s), the instruction to forget an item does. This boost in activation can render the memory vulnerable to disruption and more susceptible to subsequent forgetting.

Trial-by-trial analysis revealed a U-shaped relationship between the neural activation of a memory item and its subsequent memory strength (Fig. 3*c*) such that moderately activated items (vs weakly or strongly activated items) were most likely to be intentionally forgotten. The present result for intentionally forgotten items is also true of incidentally forgotten TBR items and is consistent with our prior work (Lewis-Peacock and Norman, 2014). While prior research has linked forgetting to suppression of sensory representations during the repeated retrievals (Wimber et al., 2015; Hulbert et al., 2016) or simulations (Benoit et al., 2016) of episodic events, the present study demonstrates that activation-dependent forgetting effects can be seen during encoding after a single exposure of an item. We believe these results provide a first step to understanding how memory representations in the sensory cortex are modified to facilitate their deliberate forgetting.

These findings are predicted by the NMPH (Newman and Norman, 2010; Detre et al., 2013), and they converge with work that describes intentional forgetting as an active and effortful cognitive process (Zacks et al., 1996; Pastötter and Bäuml, 2007; Fawcett et al., 2013). Many learning theories describe a strictly linear and positive relationship between memory activation and learning, but this hypothesis posits a U-shaped relationship such that moderate levels of memory activation lead to weakening of the memory, whereas higher levels of activation lead to strengthening. The NMPH receives support from neurophysiological data showing that moderate postsynaptic depolarization leads to long-term depression (that is, synaptic weakening) and stronger depolarization leads to long-term potentiation (Artola et al., 1990; Hansel et al., 1996; Bear, 2003; that is, synaptic strengthening). It also has received support from human neuroimaging studies showing a U-shaped relationship between how strongly a representation comes to mind and the subsequent accessibility of that representation (Newman and Norman, 2010; Detre et al., 2013; Lewis-Peacock and Norman, 2014; Poppenk and Norman, 2014; Kim et al., 2017).

Our findings are compatible with and extend existing explanations for intentional forgetting. In one prominent view on the neural mechanics that support intentional forgetting, Anderson and colleagues (for review, see Anderson and Hanslmayr, 2014) have described distinct neural mechanisms associated with two common behavioral strategies described as “direct suppression” and “thought substitution.” Behaviorally, direct suppression is thought to be a termination of the rehearsal of items, such as that elicited by a forget cue (Basden et al., 1993; Basden and Basden, 1998). Direct suppression is thought to occur when inhibitory signals from the dorsolateral prefrontal cortex downregulate hippocampal engagement related to memory encoding. Thought substitution, on the other hand, occurs when subjects replace a TBF item with some alternative item, such as an item that was studied previously or any other random thought. During thought substitution, the left ventral prefrontal cortex engages in cognitive control processes that result in demonstrative increases in hippocampal engagement (Benoit and Anderson, 2012). In the current experiment, we did not constrain behavioral strategy to avoid any mitigation of directed-forgetting effects due to self-evaluation of instructed strategies (Sahakyan et al., 2004). Therefore, participants may have attempted either or both strategies, and perhaps other idiosyncratic strategies, too. Despite potentially varied strategy choices, our results show a consistent increase in memory processing following a forget instruction relative to a remember instruction. The degree of this boost in processing, specifically when it resulted in moderate activation of the item, was predictive of successful forgetting. This suggests a new route to successful forgetting: to forget a memory, its mental representation should be enhanced to trigger memory weakening (Newman and Norman, 2010; Detre et al., 2013; described by nonmonotonic plasticity) via local inhibitory processes governing homeostatic regulation of neural activity.

A possible limitation to this interpretation of our data comes from the categorical nature of fMRI pattern classifiers used to measure memory processing (Zeithamova et al., 2017). An increase in category-specific memory processing observed for TBF items could result, not from an increase in processing of the TBF item, but from the selective retrieval and rehearsal of another item from the same category (e.g., rehearsing a previously studied TBR face when instructed to forget a different face on the current trial). This would be an example of a thought-substitution strategy (described above). We argue, however, that if people were to engage in selective rehearsal of previous items, it is unlikely that they would be able to limit this process to only same-category items rather than rehearsing a mixture of previous TBR items from both categories. There are relatively few recent same-category items to choose from: the proportion of same-category items in the span of preceding TBR trials was 0, 11.2, and 25.5% across the most recent one, three, and five TBR trials before a given TBF trial. Therefore, we believe the task procedures made it unlikely that the increased target processing on forget trials was due to thought substitution of another same-category TBR item. Empirically, if the increase in target-related processing of TBF items reflected rehearsal of same-category alternatives, this would predict a linear and negative relationship with subsequent memory such that higher processing (indicating more same-category substitution) would lead to more forgetting. Our analyses, however, revealed a U-shaped relationship with memory processing and forgetting success such that a moderate level of processing (vs low or high levels) was associated with more forgetting.

Another possible explanation for weaker neural measures of target-specific processing on TBR trials (vs TBF trials) is that participants retrieved and rehearsed other recent TBR items after a memory instruction. Turning to our data, we found that TBF items were associated with higher classifier evidence for the target category (e.g., “face” on a face trial), and also lower classifier evidence for the nontarget category (“scene”) relative to TBR items. Following an instruction to forget, activation of the TBF item was selectively enhanced. On the other hand, there was greater evidence for task-irrelevant processing (associated with the nontarget category) after a TBR instruction. This could be consistent with the idea that, following an instruction to remember, a “covert rehearsal” strategy was used in which a mixture of previously studied TBR items (some faces, some scenes) were rehearsed. However, this strategy should be expected to also increase the amount of processing for the target category. In fact, for a given TBR trial, most of the recent prior TBR items belonged to the same category as the current item (83% of the most recent item, and 69% of the three most recent items). Therefore, a recency-weighted rehearsal of prior TBR items should on balance show more classifier evidence for the target category (relative to TBF trials), but the data do not support this. In relevant work, Fawcett and Taylor (2008) explicitly instructed participants to rehearse only the current item, and not to engage in any cumulative rehearsal of prior TBR items. Memory outcomes were unaffected, indicating that multiple-item rehearsal is not essential for task performance on TBR trials. Therefore, we believe it is unlikely that a covert-rehearsal strategy can account for the lower measures of target-category processing on TBR trials.

To address more directly the potential ambiguity of the category-level classification results, we performed an item-specific RSA in which we compared the activity patterns associated with processing each TBF and TBR item in the encoding phase to the average category-specific activity pattern from the localizer phase. Figure 3*b* shows that, despite higher category classifier evidence for TBF items, the activity patterns on these trials were less typical of the category compared with the activity patterns for TBR items, suggesting that TBF items elicited greater processing of item-specific details. This result disambiguates two possible interpretations: if the increase in category classifier information for TBF items reflected a neural process that weakened the idiosyncratic, item-specific features of the item, we should expect higher category typicality for these items (due to the erosion of their item-specific features). If, on the other hand, the increase in category information on TBF trials was a result of increased attentional focus on the specific features of the item to modulate its degree of neural activation, we should expect lower category typicality for these items. These RSA results are more consistent with the latter. Altogether, our results support the idea that an increase in attentional focus on TBF items during a deliberate forgetting attempt increases their memory activation, which in turn, facilitates their forgetting.

The strength of the current study is the identification of a relationship between lingering activation of TBF memories in the VTC and their subsequent memory strength. We found a perhaps counterintuitive result that the intention to forget a memory is associated with increased activation of that memory compared with the intention to remember a memory. However, in accordance with the NMPH (Norman et al., 2006; Newman and Norman, 2010; Detre et al., 2013), we found that forgetting occurs more often when a memory has a moderate degree of activation (vs too high or too low) following the instruction to forget. This highlights the contribution of an automatic memory-weakening mechanism to deliberate forgetting, and it suggests an alternative strategy for successful forgetting: to weaken an unwanted memory, raise (rather than suppress) its level of activation.

References

Anderson MC, Green C (2001) Suppressing unwanted memories by executive control. Nature 410:366–369.

Anderson MC, Hanslmayr S (2014) Neural mechanisms of motivated forgetting. Trends Cogn Sci 18:279–292.

Artola A, Bröcher S, Singer W (1990) Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex. Nature 347:69–72.

Basden BH, Basden DR, Gargano GJ (1993) Directed forgetting in implicit and explicit memory tests: a comparison of methods. J Exp Psychol Learn Mem Cogn 19:603–616.

Basden BH, Basden DR (1998) Directed forgetting: A contrast of methods and interpretations. In: Intentional forgetting: Interdisciplinary approaches (Golding JM, MacLeod CM, eds), pp 139–172. Mahwah, NJ: Erlbaum.

Bear MF (2003) Bidirectional synaptic plasticity: from theory to reality. Philos Trans R Soc Lond B Biol Sci 358:649–655.

Benoit RG, Anderson MC (2012) Opposing mechanisms support the voluntary forgetting of unwanted memories. Neuron 76:450–460.

Benoit RG, Davies DJ, Anderson MC (2016) Reducing future fears by suppressing the brain mechanisms underlying episodic simulation. Proc Natl Acad Sci U S A 113:E8492–E8501.

Bjork EL, Bjork RA, Anderson MC (1998) Varieties of goal-directed forgetting. In: Intentional forgetting: Interdisciplinary approaches (Golding JM, MacLeod CM, eds), pp 103–137. Mahwah, NJ: Erlbaum.

Bylinskii Z, Isola P, Bainbridge C, Torralba A, Oliva A (2015) Intrinsic and extrinsic effects on image memorability. Vision Res. 116:165–78.

Christophel TB, Klink PC, Spitzer B, Roelfsema PR, Haynes JD 2017 The distributed nature of working memory. Trends Cogn Sci 21:111–124.

Cox RW (1996) AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162–173.

Cox RW, Chen G, Glen DR, Reynolds RC, Taylor PA (2017) fMRI Clustering in AFNI: False-positive rates redux. Brain Connect 7:152–171.

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31:968–980.

D'Esposito M, Postle BR (2015) The cognitive neuroscience of working memory. Annu Rev Psychol 66:115–142.

Detre GJ, Natarajan A, Gershman SJ, Norman KA (2013) Moderate levels of activation lead to forgetting in the think/no-think paradigm. Neuropsychologia 51:2371–2388.

Efron B (1979) Bootstrap methods: Another look at the jackknife. Ann Stat 7:1–26.

Fawcett JM, Taylor TL (2008) Forgetting is effortful: Evidence from reaction time probes in an item-method directed forgetting task. Mem Cognit 36:1168–1181.

Fawcett JM, Taylor TL, Nadel L (2013) Intentional forgetting diminishes memory for continuous events. Memory 21:675–694.

Fawcett JM, Lawrence MA, Taylor TL (2016) The representational consequences of intentional forgetting: Impairments to both the probability and fidelity of long-term memory. J Exp Psychol Gen 145:56–81.

Frazier JA, Chiu S, Breeze JL, Makris N, Lange N, Kennedy DN, Herbert MR, Bent EK, Koneru VK, Dieterich ME, Hodge SM, Rauch SL, Grant PE, Cohen BM, Seidman LJ, Caviness VS, Biederman J (2005) Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. Am J Psychiatry 162:1256–1265.

Gelfand H, Bjork RA (1985) On the locus of retrieval inhibition in directed forgetting. Paper presented at the 26th annual meeting of the Psychonomic Society, Boston, MA, November.

Grill-Spector K, Weiner KS (2014) The functional architecture of the ventral temporal cortex and its role in categorization. Nat Rev Neurosci 15: 536–548.

Hansel C, Artola A, Singer W (1996) Different threshold levels of postsynaptic $[Ca^{2+}]_i$ have to be reached to induce LTP and LTD in neocortical pyramidal cells. J Physiol (Paris) 90:317–319.

Hulbert JC, Henson RN, Anderson MC (2016) Inducing amnesia through systemic suppression. Nat Commun 7:11003.

Jenkinson M, Smith S (2001) A global optimisation method for robust affine registration of brain images. Med Image Anal 5:143–156.

Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved optimisation for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825–841.

Johnson HM (1994) Processes of successful intentional forgetting. Psychol Bull 116:274.

Kim G, Norman KA, Turk-Browne NB (2017) Neural differentiation of incorrectly predicted memories. J Neurosci 37:2022–2031.

Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis—connecting the branches of systems neuroscience. Front Syst Neurosci 2:4.

LaRocque JJ, Lewis-Peacock JA, Postle BR (2014) Multiple neural states of representation in short-term memory? It's a matter of attention. Front Hum Neurosci 8:5.

Lavenex P, Amaral DG (2000) Hippocampal-neocortical interaction: a hierarchy of associativity. Hippocampus. 10:420–430.

Lewis-Peacock JA, Norman KA (2014) Competition between items in working memory leads to forgetting. Nat Commun 5:5768.

Lewis-Peacock JA, Postle BR (2012) Decoding the internal focus of attention. Neuropsychologia 50:470–478.

Lewis-Peacock JA, Drysdale AT, Oberauer K, Postle BR (2012) Neural evidence for a distinction between short-term memory and the focus of attention. J Cogn Neurosci 24:61–79.

Lewis-Peacock, Jarrod A, Norman KA (2014) Multivoxel pattern analysis of functional MRI data. In: The cognitive neurosciences, Ed 5 (Gazzaniga MS George R. Mangun GR, eds), pp 911–920. Cambridge, MA: MIT.

MacLeod CM (1999) The item and list methods of directed forgetting: Test differences and the role of demand characteristics. Psychon Bull Rev 6:123–129.

Makris N, Goldstein JM, Kennedy D, Hodge SM, Caviness VS, Faraone SV, Tsuang MT, Seidman LJ (2006) Decreased volume of left and total anterior insular lobule in schizophrenia. Schizophr Res 83:155–171.

Newman EL, Norman KA (2010) Moderate excitation leads to weakening of perceptual representations. Cereb Cortex 20:2760–2770.

Norman KA, Newman E, Detre G, Polyn S (2006) How inhibitory oscillations can train neural networks and punish competitors. Neural Comput 18:1577–1610.

Norman KA, Newman EL, Detre G (2007) A neural network model of retrieval-induced forgetting. Psychol Rev 114:887–953.

Pastötter B, Bäuml KH (2007) The crucial role of postcue encoding in directed forgetting and context-dependent forgetting. J Exp Psychol Learn Mem Cogn 33:977–982.

Poppenk J, Norman KA (2014) Briefly cuing memories leads to suppression of their neural representations. J Neurosci 34:8010–8020.

Rissman J, Wagner AD (2012) Distributed representations in memory: Insights from functional brain imaging. Annu Rev Psychol 63:101–128.

Rizio AA, Dennis NA (2013) The neural correlates of cognitive control: Successful remembering and intentional forgetting. J Cogn Neurosci 25:297–312.

Sahakyan L, Delaney PF, Kelley CM (2004) Self-evaluation as a moderating factor of strategy change in directed forgetting benefits. Psychon Bull Rev 11:131–136.

Tottenham N, Tanaka JW, Leon AC, McCarry T, Nurse M, Hare TA, Marcus DJ, Westerlund A, Casey BJ, Nelson C (2009) The NimStim set of facial expressions: judgments from untrained research participants. Psychiat Res 168:242–249.

Wimber M, Alink A, Charest I, Kriegeskorte N, Anderson MC (2015) Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. Nat Neurosci 18:582–589.

Wylie GR, Foxe JJ, Taylor TL (2008) Forgetting as an active process: an FMRI investigation of item-method-directed forgetting. Cereb Cortex 18:670–682.

Zacks RT, Radvansky G, Hasher L (1996) Studies of directed forgetting in older adults. J Exp Psychol Learn Mem Cogn 22:143–156.

Zeithamova D, de Araujo Sanchez MA, Adke A (2017) Trial timing and pattern-information analyses of fMRI data. Neuroimage 153:221–231.

Zwissler B, Schindler S, Fischer H, Plewnia C, Kissler JM (2015) 'Forget me (not)?'—remembering forget-items versus un-cued items in directed forgetting. Front Psychol 6:1741.