Pruning of memories by context-based prediction error

Ghootae Kim[a,b], Jarrod A. Lewis-Peacock[c,d], Kenneth A. Norman[a,b], and Nicholas B. Turk-Browne[a,b,1]

[a]Department of Psychology and [b]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544; and [c]Department of Psychology and [d]Institute for Neuroscience, University of Texas at Austin, Austin, TX 78712

Edited by Daniel L. Schacter, Harvard University, Cambridge, MA, and approved May 8, 2014 (received for review October 16, 2013)

The capacity of long-term memory is thought to be virtually unlimited. However, our memory bank may need to be pruned regularly to ensure that the information most important for behavior can be stored and accessed efficiently. Using functional magnetic resonance imaging of the human brain, we report the discovery of a context-based mechanism for determining which memories to prune. Specifically, when a previously experienced context is reencountered, the brain automatically generates predictions about which items should appear in that context. If an item fails to appear when strongly expected, its representation in memory is weakened, and it is more likely to be forgotten. We find robust support for this mechanism using multivariate pattern classification and pattern similarity analyses. The results are explained by a model in which context-based predictions activate item representations just enough for them to be weakened during a misprediction. These findings reveal an ongoing and adaptive process for pruning unreliable memories.

forgetting | learning | multivariate pattern analysis | perception | temporal context

Our experience of the world is recorded in long-term memory every moment of every day. Such memory formation occurs continuously and incidentally, resulting in a potentially astronomical number of memory traces. This cluttering can be problematic for the efficient functioning of memory systems in the brain. At retrieval, irrelevant memories will compete with the sought-after memory and can prevent it from coming to mind. To avoid these costs, we propose that memory systems regulate themselves by adaptively "pruning" item representations.

How does the brain decide which items to prune from memory? We hypothesize that the brain makes this determination based on how accurately an item is predicted by its context. Specifically, the brain may automatically predict which items are likely to appear in a given context based on prior experience and then compare this prediction against the actual contents of experience. When the prediction is wrong, the representations of the expected items in long-term memory are weakened. This weakening is manifested in a graded reduction of the accessibility of these items during later retrieval. In this way, the brain can use context-based prediction error to determine when an item is not a stable aspect of the world, and this determination may, in turn, mark the item for pruning.

The notion that the brain automatically predicts which items should appear in a context is supported by previous research. Such predictions result from learning relationships between items (1), which in turn allow the appearance of one item to cue the reactivation of other associated items (2–4). Behavioral studies have obtained evidence for automatic prediction by showing that task performance is facilitated for items that are predictable in the current context (5, 6). More recently, neuroimaging studies have obtained evidence for automatic prediction by showing that the medial temporal lobe and sensory cortex reinstate representations of predicted items (7–9).

Although context-based prediction has been found previously, its consequences for item learning are unknown. As noted above, such prediction provides an opportunity to discover whether an item is a stable aspect of the world. When an item fails to appear in a context with which it previously has been associated, the resulting error signal may cause the item to be pruned from memory. To test this pruning hypothesis, we set out to relate prediction strength to subsequent memory. Specifically, our hypothesis posits that in situations where a predicted item does not appear as expected, prediction strength should be negatively related to subsequent memory for the item: Stronger predictions lead to larger prediction errors, which in turn lead to more weakening of the predicted item's memory trace and ultimately to reduced confidence in having seen the item before and a greater likelihood of forgetting the item altogether.

Numerous studies have investigated how prediction error shapes learning in the brain (10–12). Our study differs from these studies in two important ways. First, existing studies have focused primarily on learning to predict future rewards. That is, the specific identities of the predicting stimuli were irrelevant other than in terms of how much reward they predicted (13). In contrast, we examine an unsupervised form of stimulus–stimulus learning in which relationships are formed between, and predictions made about, stimulus identities with no inherent motivational significance. Second, prediction error typically is viewed as a way of updating associative strength between cues and outcomes (14). Here, that process would correspond to learning how strongly an item (the "outcome") should be predicted by its context (the "cue"). Instead, we test the idea that context-based prediction error weakens the long-term memory representation of the predicted item itself.

To relate the context-based prediction of an item to its later accessibility in memory, we developed a trial-by-trial measure of prediction strength. There is no clear behavioral signature of automatic prediction that can be measured for a single trial. Thus, we used the output of a multivariate pattern classifier applied to functional magnetic resonance imaging (fMRI) data

Significance

Forgetting is often considered to be bad, but selective forgetting of unreliable information can have the positive side effect of reducing mental clutter, thereby making it easier to access our most important memories. Prior studies of forgetting have focused on passive mechanisms (decay, interference) or on effortful inhibition by cognitive control. Here we report the discovery of an active mechanism for forgetting that weakens memories selectively and without burdening the conscious mind. Specifically, we show that the brain automatically generates predictions about which items should appear in familiar contexts; if these items fail to appear, their memories are weakened. This process is adaptive, because such memories may have been encoded incorrectly or may represent unstable aspects of the world.

Author contributions: G.K., J.A.L.-P., K.A.N., and N.B.T.-B. designed research; G.K. performed research; G.K. and J.A.L.-P. analyzed data; and G.K., J.A.L.-P., K.A.N., and N.B.T.-B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: ntb@princeton.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1319438111/-/DCSupplemental.

(15). With this technique, we quantified how strongly an item was predicted when its previously associated temporal context was repeated and then related this prediction strength to subsequent recognition memory for that item. We also tested our hypothesis with a second multivariate technique, pattern similarity analysis (16). This approach allowed us to replicate and extend the main findings, confirming that the relationship between classifier output and subsequent memory reflected prediction of the previously associated item per se (*SI Text*).

Twenty-four participants completed an fMRI session in two phases: an incidental encoding phase and a subsequent memory test phase. In the incidental encoding phase, participants were exposed to a continuous stream of photographs of faces and scenes while performing a categorization cover task (discriminating male/female or indoor/outdoor, respectively). Unbeknownst to participants, the stream was generated from triplets (Fig. 1*A*). The first two context items in each triplet were from one category, and the final item was from the other category (e.g., $A_{face}\rightarrow B_{face}\rightarrow C_{scene}$). These two context items were repeated later in the stream but this time were followed by a novel final item from their same category (e.g., $A_{face}\rightarrow B_{face}\rightarrow D_{face}$). The triplets were used to construct the sequence, but the triplet structure was not overtly signaled to participants; items appeared continuously every 4.5 s. Additional single items from each category (i.e., items whose preceding context items never repeated) were inserted into the stream between triplets (e.g., X_{scene}).

We expected that the repeated context items would automatically trigger a prediction that the original final item (C) would appear next — a prediction that then would be violated by the appearance of the novel final item (D). During these repeated triplets, we assessed the strength of the prediction of C by measuring how much information about the C item's category was available in the brain. Note that C always came a category different from the A, B, and D items in the corresponding repeated triplet, making it possible to resolve how much participants were expecting C using a category-based classifier.

In the subsequent memory test phase, participants performed a recognition task for items from the encoding phase (Fig. 1*B*).

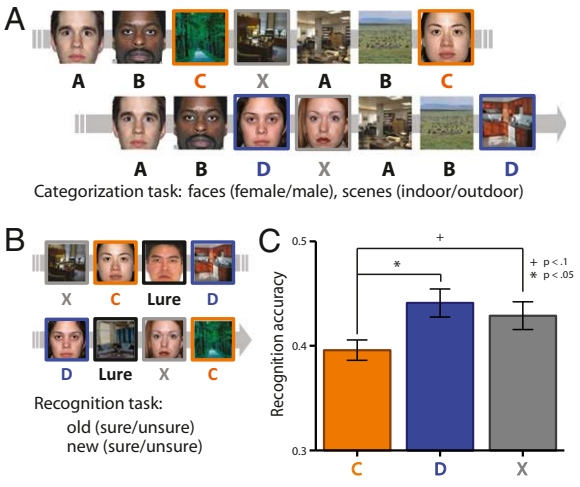

Fig. 1. Experimental design and behavioral results. (*A*) During incidental encoding, the trial sequence was constructed from triplets (A→B→C) that repeated once with a novel final item (A→B→D) and unrepeated single items (X). The categorization task was orthogonal to the triplet structure. (*B*) In the subsequent memory test, old/new judgments were collected for final items from initial (C) and repeated (D) triplets, single control items (X), and novel lure items. All old items appeared once during incidental encoding. (*C*) The high-confidence hit rate for C items was significantly lower than for D items and was marginally lower than for X items. Error bars reflect ± 1 SEM.

In addition to new lures, there were three types of old items: the final items from the initial (C) and repeated triplets (D), and control items that were not part of a triplet (X). Importantly, all old items appeared only once during encoding, and thus differences in memory performance between conditions must result from the repetition of context items and any associated prediction errors.

Results

Subsequent Memory Behavior. Participants displayed reliable memory performance overall, successfully discriminating old items in every condition from lures (all Ps < 0.001) (Fig. S1). Although the essential test of our hypothesis relies on relating memory to prediction strength on an item-by-item basis, we first considered whether there was overall memory suppression for C items relative to both D and X items irrespective of prediction strength (Fig. 1*C*). Indeed, the high-confidence hit rate for C items was significantly lower than for D items ($t_{23} = 2.33$, $P = 0.029$) and was marginally lower than for X items ($t_{23} = 1.73$, $P = 0.096$). These results did not depend on the greater novelty of items preceding C items during encoding nor on the earlier average serial position of C items in the trial sequence (*SI Text*).

For the fMRI analyses below, we treat high-confidence hits as "remembered" and other items as "forgotten." Requiring high confidence for an item to be considered as remembered is consistent with prior studies (e.g., refs. 17 and 18) and is supported by participants exhibiting greater sensitivity for high-confidence responses (mean $A' = 0.81$) than for low-confidence responses (mean $A' = 0.58$; $t_{23} = 6.07$, $P < 0.001$).

Relating Prediction Strength to Subsequent Memory. Our hypothesis does not state that memory will be impaired for all mispredicted items. Rather, context-based prediction error triggers pruning, and thus only C items that were strongly predicted in the repeated triplet should be more likely to be forgotten. That is, we expected that the amount of prediction for a given C item should be negatively related to its likelihood of being remembered: Weak predictions should generate small errors with little impact, whereas strong predictions should generate large errors and lead to weakening of the memory. To test this hypothesis, we measured prediction strength with an fMRI-based pattern classifier.

For classification, we used regularized logistic regression (penalty = 1) to identify face and scene information in brain-activity patterns from anatomical regions of interest (ROIs) in bilateral ventral temporal cortex. A separate classifier was trained for each participant using an independent functional localizer (Fig. S2*A*). The classifier then was applied to continuous brain patterns from the incidental encoding phase. For each C item, we calculated the relative amount of classifier evidence for its category during the repeated triplet. We interpreted information about C's category as evidence of prediction for two reasons: (*i*) The C item (e.g., C_{scene}) was not shown in the repeated triplet, and (*ii*) the items that did appear in the repeated triplet were all from the other category (e.g., $A_{face}\rightarrow B_{face}\rightarrow D_{face}$). For the main analysis, we averaged the relative evidence for C's category over all time points in the repeated triplet. To test the hypothesis that greater misprediction of a C item increases the likelihood of its being forgotten, we related this average category evidence to the memory outcome for the same C item using logistic regression (Fig. 2*A*).

Consistent with our hypothesis, there was a significant negative relationship between prediction strength and memory (Fig. 2*B*): Greater category evidence was linked to an increased likelihood of forgetting ($\beta_1 = -0.85$, $P < 0.001$). To examine how this negative trend developed, we performed this same analysis over different time windows in the repeated triplet (Fig. 2*C*). We binned the three time points around each of the three items in the repeated triplet and around the triplet's onset (as a baseline).

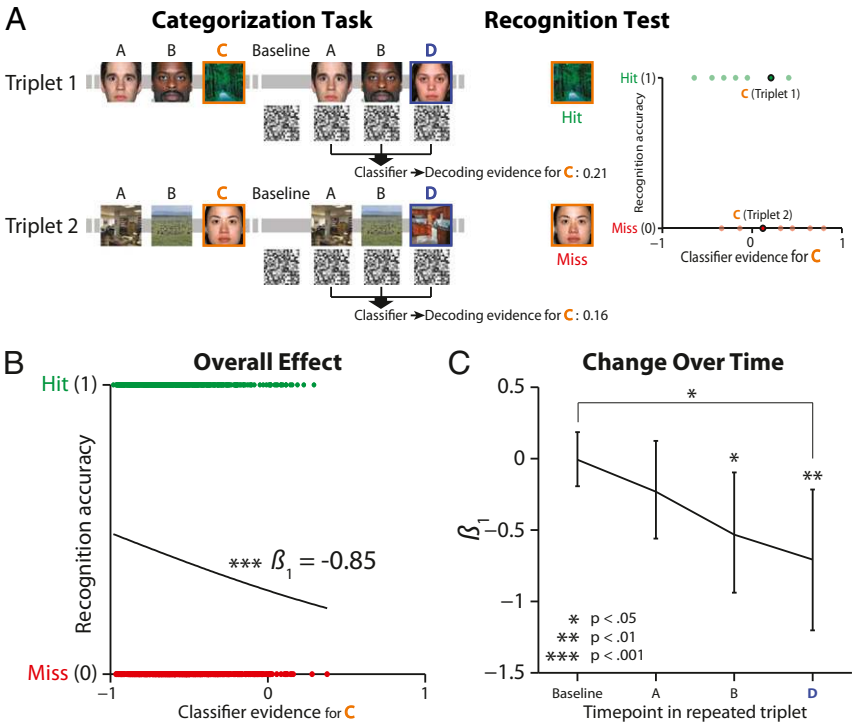

Fig. 2. Results of pattern classification analysis. (*A*) A classifier trained to decode visual categories was applied to brain patterns from the repeated triplets. Classifier evidence for the category of a C item was related to subsequent memory for that item. This category was not presented in the repeated triplet, and thus classifier evidence for it was interpreted as prediction. (*B*) Category evidence for C was first averaged over all time points in the repeated triplet. Dots indicate the distribution of classifier evidence for remembered (green) and forgotten (red) items. Logistic regression analyses revealed a reliable negative relationship (i.e., negative β_1) in which greater classifier evidence was associated with more forgetting. (*C*) The same analysis was performed separately during the baseline, A, B, and D time periods. The negative relationship was maximal during the anticipated time of C but was evident earlier in the repeated triplet (during B). Error bars reflect 95% bootstrap CIs.

If the overall negative trend reflects context-based prediction, the trend should be most negative later in the triplet. Namely, there should be no trend before the context items appear, and the trend should grow to maximum negativity at the time of D, after both predictive context items have appeared and at the anticipated time of C. Indeed, the beta coefficient during the D time period was reliably negative ($P = 0.002$) and was significantly lower than baseline ($P = 0.014$). All these findings were validated with a second multivariate analysis approach based on pattern similarity (Fig. S3), which confirmed that prediction of the C item per se was negatively related to subsequent memory. Note that because only high-confidence hits were treated as remembered, increased forgetting may reflect reduced confidence in recognition (e.g., low-confidence hits that were treated as forgotten). Nonetheless, this interpretation still would be consistent with the hypothesis that strong prediction of the C item weakened its memory.

To test the specificity of these results, we examined whether classifier evidence for categories other than the C category was related to subsequent memory. We first considered whether the observed negative relationship would persist after controlling for evidence of the D category with partial correlation and found that it remained robust ($P = 0.008$). When this analysis was reversed, testing how D category evidence was related to memory for C after controlling for C category evidence, there was no relationship ($P = 0.99$). Moreover, classifier evidence for neither the object category nor "rest" (from the localizer) predicted C memory ($Ps > 0.36$). Thus, worse memory for C items was explained only by evidence for the C category in repeated triplets.

Ruling out an Alternative Interpretation. The observed negative relationship between prediction strength and subsequent memory is consistent with our hypothesis that context-based prediction error leads to pruning. However, a potential alternative explanation is that forgetting occurred because of poor encoding of C in the initial triplet rather than because of a strong prediction of C in the repeated triplet. In this scenario, repeated context items triggered the attempted retrieval of C, which was slower and more difficult for items that were poorly encoded. During this memory search, other items from the same category were examined and rejected, giving rise to greater category information. Under this interpretation, the negative relationship between classifier evidence and memory is spurious: The quality of the encoding of the C item in the initial triplet may determine both the amount of category information in the repeated triplet (poor quality means more search and thus more category information) and the subsequent memory outcome (poor quality means more forgetting), without the former having any direct effect on the latter.

According to this alternative, the observed negative relationship should be eliminated if the quality of initial encoding is controlled. We operationally defined encoding quality as the amount of classifier evidence for C's category when it was perceived in the initial triplet. This definition is based on the finding of enhanced activation during initial encoding for subsequently remembered vs. forgotten items (17, 19). We then repeated all analyses after controlling for the effects of this "perception strength" measure on both prediction strength and subsequent memory (Fig. 3). Contrary to the alternative explanation, the negative relationship between prediction strength and subsequent memory remained robust ($P = 0.004$).

Evidence of Prediction. Having ruled out a contribution of initial encoding to the negative relationship between prediction strength and memory, we are left with the interpretation that prediction of C items during the repeated triplet per se was responsible for their worse memory. To support this interpretation, we performed two additional tests of whether C items were predicted overall.

First, as would be expected, there was more classifier evidence for the C category during A and B in repeated triplets (when C could be predicted) than during A and B in initial triplets ($P = 0.038$). There is a potential confound with this analysis, however: A and B were novel in initial triplets but not in repeated triplets. If repetition suppression reduced the amount of activity for A

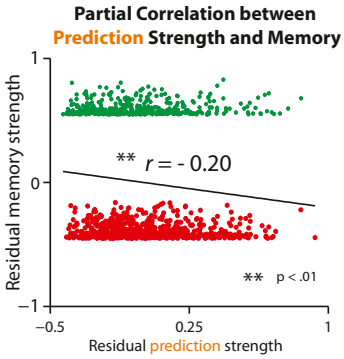

Fig. 3. Partial logistic regression. Scatterplot and correlation of residual prediction strength and residual memory outcome after partialing out item-by-item perception strength. Green and red dots indicate the distributions of prediction strength values for subsequently remembered and forgotten items, respectively.

and B (20), then classifier evidence for their category may decrease, spuriously increasing evidence for other categories, including C's. This confound is improbable because repetition suppression has been linked to increased, not decreased, classifier evidence for repeated items (7). Indeed, more repetition suppression for A and B (initial minus repeated univariate activity; *SI Text*) was associated with greater evidence for their category ($P = 0.040$) and lower evidence for C's category ($P < 0.001$). Thus, repetition suppression worked against the observed increase in C category evidence during context items in repeated vs. initial triplets, so the increase can be interpreted more soundly as prediction of C.

Second, insofar as C was predicted in repeated triplets, then the amount of classifier evidence for C's category in initial and repeated triplets should be correlated across items. That is, idiosyncratic variance in the prototypicality of the faces and scenes should lead to systematic variance across items in the amount of evidence for their category; because we posited that the same C item was processed in matching initial and repeated triplets, prediction strength should be correlated with perception strength. Indeed, this correlation was reliably positive at the time of C in the repeated triplet ($P = 0.022$). There also is a potential confound in this analysis: A and B were present in both initial and repeated triplets, and carryover activity from these items may have driven the correlation. However, the relationship remained marginally significant after controlling for the context items with partial correlation ($P = 0.060$). These results also are consistent with our interpretation that C items were predicted.

Nonmonotonic Plasticity Mechanism. Our discovery of a negative relationship between memory activation and subsequent memory is striking, given that positive relationships have been reported routinely (17, 19). How does context-based prediction lead to memory weakening? We interpret these findings in terms of the nonmonotonic plasticity hypothesis (21, 22), which posits a U-shaped relationship between memory activation and learning: Low activation does not lead to learning, moderate activation leads to memory weakening, and high activation leads to strengthening. This hypothesis is based on neurophysiological findings (23, 24): Moderate postsynaptic depolarization causes long-term depression (i.e., synaptic weakening), whereas greater postsynaptic depolarization causes long-term potentiation (i.e., synaptic strengthening).

Neural network modeling suggests that this learning principle should scale up from synapses to neural ensembles (25, 26). In these models, memories are composed of distributed populations of neurons and the strength of the memory is proportional to the degree of interconnectivity between neurons in the ensemble. After incorporating nonmonotonic plasticity, moderate activation of these distributed ensembles weakens synapses within the ensemble (thereby reducing behavioral expression of the memory), and stronger activation strengthens synapses within the ensemble (thereby increasing behavioral expression of the memory). Importantly, these models predict that nonmonotonic plasticity should apply to the initial presentation of an item as well as to subsequent presentations: When an item is first presented, its neural representation will have some (nonzero) level of interconnectivity arising from prior experience with related stimuli; this initial level of interconnectivity can be reduced (in the case of moderate activation) or increased (in the case of strong activation). In keeping with the predictions of these models, prior studies have found a U-shaped relationship between activation (in EEG or fMRI) and subsequent accessibility (21, 22). For example, moderate levels of activation elicited by the first presentation of a stimulus lead to negative priming (i.e., slower subsequent responding), whereas higher levels of activation lead to positive priming (i.e., faster subsequent responding) (21).

This model can explain the observed relationship between prediction and memory if strong but unconfirmed predictions trigger moderate item activation and lesser predictions trigger weaker activation. Indeed, even the strongest predictions in the current task may result in only moderate activation: There is only a single opportunity to associate C items with their context, predictions were incidental with respect to the categorization task during encoding, and prediction is internally generated activation — such activation is weaker than perceptual activation in general (27). When activation levels fall in this low-to-moderate range, the nonmonotonic plasticity hypothesis posits that the relationship of item activation to memory will be negative: Weak predictions will be neutral, whereas stronger predictions (leading to moderate activation) will induce forgetting. By this account, learning does not depend on the explicit computation of an error signal. Rather, prediction error is realized implicitly in the brain as moderate activation of the unconfirmed prediction.

It was hard to know a priori that context-based predictions would elicit low-to-moderate activation values. Nevertheless, based on this claim, we can evaluate the nonmonotonic plasticity hypothesis by exploring the consequences of higher levels of activation: According to nonmonotonic plasticity, the relationship of item activation to memory should be positive when activation values range from moderate (which causes weakening) to high (which causes strengthening). To test this idea, we related perception strength—i.e., classifier evidence for the C category in initial triplets—to subsequent memory. We expected that the perception of an item would elicit substantially higher activation levels than its later context-based prediction (27), leading to a positive relationship with subsequent memory.

As expected, perception strength was robustly greater on average than prediction strength ($t_{23} = 38.57$, $P < 0.001$) (Fig. 4*B*, *Lower*). Furthermore, the relationship between perception strength and subsequent memory was reliably positive ($\beta_1 = 0.81$, $P = 0.002$) (Fig. 4*A*). As shown in Fig. 4*B*, moderate activation (derived from strong prediction or from weak perception) led to worse memory relative to both low activation (weak prediction) and high activation (strong perception).

To model the combined influence of perception and prediction, we used the probabilistic curve induction and testing (P-CIT) Bayesian curve-fitting algorithm (22). This algorithm estimates the posterior distribution over "plasticity curves" relating activation (classifier evidence during incidental encoding) to learning (subsequent memory behavior). For each C item, perception in the initial triplet and prediction in the repeated triplet were treated as separate learning events whose effects were summed to model recognition. The curves recovered by P-CIT

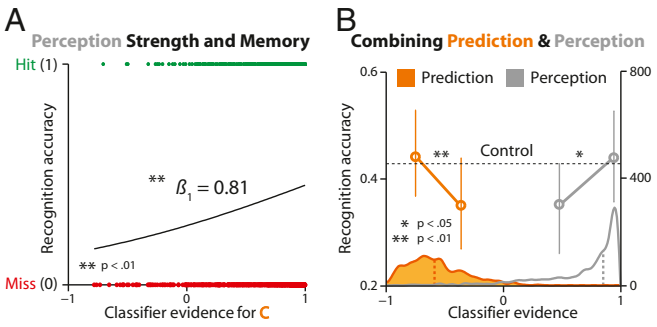

Fig. 4. Nonmonotonic relationship. (*A*) Logistic regression fit of perception strength to subsequent memory. (*B*) Relationship between classifier evidence and subsequent memory. (*Lower*) Histograms of classifier evidence for prediction (orange) and perception (gray) over C items. Vertical dotted lines indicate the median. (*Upper*) Memory for C in each half of median splits. Error bars reflect 95% bootstrap CIs.

were reliably U-shaped, consistent with our interpretation (Fig. S4).

Discussion

We obtained evidence that context-based prediction error can lead to forgetting using multivariate pattern analyses of fMRI data. This finding supports our pruning hypothesis that item memories are weakened when they are mispredicted by their context. The nonmonotonic plasticity hypothesis provides a neurobiologically plausible mechanism for these findings: A large prediction error is associated with moderate activation of the mispredicted item; this moderate activation, in turn, triggers weakening of the synapses that support the item's representation in memory.

There are some boundary conditions for this account. As we showed, weak predictions induce low activation and leave memory intact. At the same time, very strong predictions (e.g., after extensive experience) may induce activation that is high enough to shield against pruning or even enhance memory (28). Moreover, if a prediction is correct, the sum of bottom-up perception and top-down prediction likely will yield activation that is high enough to shield the memory from pruning.

Note that our use of the term "pruning" is not meant to imply that traces are being deleted completely from memory. Rather, the pruning we describe refers to a mnemonic regulation process in which the accessibility of less-reliable memories is reduced in a graded manner. This graded decrease in accessibility also might show up as a decrease in recognition confidence. Indeed, we obtained the same negative relationship between prediction strength and memory when we used a continuous measure of memory defined linearly over the range of high-confidence new, low-confidence new, low-confidence old, and high-confidence old ($P = 0.004$). We interpret reduced confidence for mispredicted items as reflecting weakening of their representation rather than interference during retrieval from a new memory trace created during the repeated triplet (*SI Text*).

Forgetting seems disadvantageous but plays an essential role in maintaining the efficiency of memory operations (29). Our study sheds light on the adaptive role of forgetting. Previous studies examined the impact of controlled retrieval on forgetting, whereby executive control processes inhibit or suppress undesirable memories competing for retrieval (30). Here we demonstrate, for the first time to our knowledge, that such forgetting can occur without deploying control processes, simply as a result of automatic retrieval during context-based prediction. Participants reported being unaware that contexts were repeating, suggesting that automatic retrieval occurs constantly in the background, pruning invalid memories without burdening our conscious mind.

Materials and Methods

Participants. Twenty-four adults (14 women; 19 right-handed; mean age, 22.8 y) participated for monetary compensation. All participants had normal or corrected-to-normal vision and provided informed consent. The study protocol was approved by the Princeton University Institutional Review Board.

Stimuli. Participants were shown color photographs of male and female faces (including from www.macbrain.org/resources.htm), indoor and outdoor scenes (including from http://cvcl.mit.edu/MM/sceneCategories.html), and natural and manmade objects. Stimuli were displayed on a projection screen behind the scanner, viewed with a mirror on the head coil (subtending 8.8 × 8.8°). Participants fixated a central dot that remained onscreen.

Procedure. Participants completed one scanning session, including the incidental encoding phase, the subsequent memory test, and a functional localizer. During incidental encoding, participants viewed a sequence of faces and scenes and made male/female and indoor/outdoor judgments. Unbeknownst to them, this sequence was generated from triplets. The first two "context" items in a triplet were either both faces or both scenes, and the final item was from the other category. Each triplet used novel exemplars. The context items were shown again after other intervening items (average lag = 12.5 items), with the final item in the triplet replaced by a new item from the context category. Other single items, whose preceding context items never repeated, were inserted between triplets. Three runs of incidental encoding were collected, each lasting 513 s and containing 16 initial triplets (eight $A_{face}\rightarrow B_{face}\rightarrow C_{scene}$ and eight $A_{scene}\rightarrow B_{scene}\rightarrow C_{face}$), their repetitions (eight $A_{face}\rightarrow B_{face}\rightarrow D_{face}$ and eight $A_{scene}\rightarrow B_{scene}\rightarrow D_{scene}$, respectively), and 16 control items (eight X_{face} and eight X_{scene}). The resulting 112 trials started with a blink of fixation to signal an upcoming trial, followed by the stimulus for 1 s and a blank interval of 3.5 s.

The subsequent memory test phase began ~10 min after encoding. The memory test was a surprise to participants. It consisted of a recognition task for 144 old items and 48 novel lure items. There were three types of old items: 48 final items each from initial (C) and repeated triplets (D), and 48 control items (X). Old items were interleaved with lures randomly. Participants judged familiarity on a 4-point scale: 1 = sure old, 2 = unsure old, 3 = unsure new, and 4 = sure new. As in previous studies (e.g., 17, 18), items receiving high-confidence "sure old" responses were treated as remembered, and items receiving other responses were treated as forgotten.

After the test, participants completed two runs of a functional localizer. Each run contained 15 blocks, with five blocks from each of three categories: faces, scenes, and objects. Participants judged faces as male or female, scenes as indoor or outdoor, and objects as manmade or natural. Each stimulus was presented for 500 ms, followed by a blank interval of 1,000 ms. There were 10 trials per block. Each 15-s block was followed by 10.5 s of fixation, which was treated as a rest category. Total run duration was 400.5 s.

Data Acquisition. Experiments were run with the Psychophysics Toolbox (http://psychtoolbox.org). Neuroimaging data were acquired using a 3-T MRI scanner (Siemens Skyra) with a 16-channel head coil. We collected a scout anatomical to align axial functional slices. Whole-brain functional images for the encoding phase and functional localizer were acquired with a gradient-echo echo planar imaging sequence (TR = 1.5 s; TE = 28 ms; flip = 64°; integrated parallel acquisition technique = 2; matrix = 64 × 64; slices = 27; thickness = 4 mm, resolution = 3 × 3 mm). High-resolution (magnetization-prepared rapid-acquisition gradient echo) and coplanar (fast low-angle shot) T1 anatomical scans were acquired for registration, along with field maps to correct B0 inhomogeneity.

Preprocessing. The fMRI data were preprocessed with FSL (http://fsl.fmrib.ox.ac.uk). Functional scans were corrected for slice-acquisition time and head motion, high-pass filtered (128-s period cutoff), and aligned to the first volume. All multivariate pattern analyses were conducted on voxels within anatomically delineated ventral temporal cortex (31). We generated these ROIs in standard space by summing left and right masks of temporal fusiform cortex and parahippocampal gyrus from the Harvard–Oxford cortical atlas in FSL. We converted the ROIs to subject space by inverting the transformations obtained from registering functional scans to standard space.

Classification Analyses. Classification was conducted with the Princeton Multi-Voxel Pattern Analysis Toolbox (www.pni.princeton.edu/mvpa), using

penalized logistic regression with L2-norm regularization (penalty = 1). To validate our classifier, we first performed a cross-validation analysis within the localizer data. We trained a separate model for each of four categories—face, scene, object, and rest—using one of the localizer runs and tested it on the other run (and then swapped training and test runs). All regressors were shifted forward in time by 4.5 s to adjust for the hemodynamic lag. For each fMRI volume in the test set, the classifier estimated the extent to which the activity pattern matched the activity patterns for the four categories on which it was trained (from 0 to 1). We refer to these category-level pattern match values as "classifier evidence."

To classify the incidental encoding data, we trained a model on both localizer runs. We operationalized prediction strength as the activation of a C item's category during the repeated triplet. Note that all items in the repeated triplet were from the other category; for example, if C was a scene, all the items in the repeated triplet were faces. We averaged classifier evidence over the repeated triplet (3–16.5 s after trial onset, adjusted for hemodynamic lag) and calculated the difference between the evidence for C's category and A/B/D's category (e.g., scene minus face evidence for A_{face}→B_{face}→D_{face}).

We used this relative measure of category information because it has proven more sensitive in prior studies (22). The pattern of results was identical when we used evidence for C's category alone (without subtracting evidence for A/B/D's category). To quantify how category information changed over time, we locked trial regressors for the incidental encoding phase to trial onset time (Fig. S2*B*). We then binned three volumes around the onset of the repeated triplet (baseline) and around each item in the repeated triplet shifted forward by 3 s to account for the hemodynamic lag: baseline = −1.5–1.5 s; A = 3–6 s; B = 7.5–10.5 s; and D = 12–15 s.

Logistic Regression Analyses. The main goal of our study was to examine the relationship between classifier evidence for C in the repeated triplet and subsequent memory for that item. Characterizing this relationship in a quantitatively precise manner required a substantial amount of data; therefore we pooled trials across participants before performing the logistic regression analyses.

There are two potential concerns with this kind of "super-subject" analysis. First, effects could be driven by a subset of participants and not generalize to the population. Therefore we assessed the reliability of our results across participants (random-effects) using a bootstrap test in which we resampled entire participants with replacement and performed the same logistic regression analyses on the resampled data (32). This resampling provided a population-level confidence interval (CI) for each effect in terms of the proportion of bootstrap samples (out of 1,000) in which the effect was present.

The second potential concern is that effects may not be reliable within individual participants. For example, the supersubject relationship between prediction strength and subsequent memory might reflect variance across rather than within participants. To address this concern, we standardized prediction strength values within each participant (by z-scoring classifier evidence across C items) and reran all supersubject analyses. Because this step eliminates differences in mean prediction strength across participants, any remaining effects are attributable to within-participant variance in prediction strength. The pattern of results was the same as what we obtained when not standardizing: There was a negative relationship between average classifier evidence and subsequent memory ($\beta_1 = -0.26$, $P = 0.006$), and the relationship in the D time period was reliably negative ($P = 0.004$) and significantly more negative than the relationship in the baseline time period ($P = 0.026$).

We also examined the relationship between the perception of C and subsequent memory for that item. Similar to prediction strength, perception strength was defined as the relative classifier evidence for C's category. The only difference is that we measured this evidence during C's presentation in the initial triplet, 4.5 s after its onset. We then performed the same logistic regression analyses to relate perception to memory.

ACKNOWLEDGMENTS. We thank Lila Davachi and Per Sederberg for helpful conversations. This work was supported by National Institutes of Health Grants R01 EY021755 (to N.B.T-B.) and R01 MH069456 (to K.A.N.).

1. Cohen NJ, Eichenbaum H, eds (1993) *Memory, Amnesia, and the Hippocampal System* (MIT Press, Cambridge, MA).
2. Hirsh R (1974) The hippocampus and contextual retrieval of information from memory: A theory. *Behav Biol* 12(4):421–444.
3. Miyashita Y (1993) Inferior temporal cortex: Where visual perception meets memory. *Annu Rev Neurosci* 16:245–263.
4. Howard MW, Kahana MJ (2002) A distributed representation of temporal context. *J Math Psychol* 46(3):269–299.
5. Nissen MJ, Bullemer P (1987) Attentional requirements of learning: Evidence from performance measures. *Cognit Psychol* 19(1):1–32.
6. Olson IR, Chun MM (2001) Temporal contextual cuing of visual attention. *J Exp Psychol Learn Mem Cogn* 27(5):1299–1313.
7. Kok P, Jehee JFM, de Lange FP (2012) Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron* 75(2):265–270.
8. Schapiro AC, Kustner LV, Turk-Browne NB (2012) Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Curr Biol* 22(17): 1622–1627.
9. Turk-Browne NB, Simon MG, Sederberg PB (2012) Scene representations in parahippocampal cortex depend on temporal context. *J Neurosci* 32(21):7202–7207.
10. O'Doherty J, et al. (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304(5669):452–454.
11. Pagnoni G, Zink CF, Montague PR, Berns GS (2002) Activity in human ventral striatum locked to errors of reward prediction. *Nat Neurosci* 5(2):97–98.
12. Schultz W, Dickinson A (2000) Neuronal coding of prediction errors. *Annu Rev Neurosci* 23:473–500.
13. Niv Y, Schoenbaum G (2008) Dialogues on prediction errors. *Trends Cogn Sci* 12(7): 265–272.
14. Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory*, eds Black AH, Prokasy WF (Appleton Century Crofts, New York), pp 64–99.
15. Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci* 10(9):424–430.
16. Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis - connecting the branches of systems neuroscience. *Front Syst Neurosci* 2:4.
17. Wagner AD, et al. (1998) Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science* 281(5380):1188–1191.
18. Turk-Browne NB, Yi D-J, Chun MM (2006) Linking implicit and explicit memory: Common encoding factors and shared representations. *Neuron* 49(6):917–927.
19. Brewer JB, Zhao Z, Desmond JE, Glover GH, Gabrieli JDE (1998) Making memories: Brain activity that predicts how well visual experience will be remembered. *Science* 281(5380):1185–1187.
20. Grill-Spector K, Henson R, Martin A (2006) Repetition and the brain: Neural models of stimulus-specific effects. *Trends Cogn Sci* 10(1):14–23.
21. Newman EL, Norman KA (2010) Moderate excitation leads to weakening of perceptual representations. *Cereb Cortex* 20(11):2760–2770.
22. Detre GJ, Natarajan A, Gershman SJ, Norman KA (2013) Moderate levels of activation lead to forgetting in the think/no-think paradigm. *Neuropsychologia* 51(12): 2371–2388.
23. 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(6288):69–72.
24. Hansel C, Artola A, Singer W (1996) Different threshold levels of postsynaptic [Ca2+]i have to be reached to induce LTP and LTD in neocortical pyramidal cells. *J Physiol Paris* 90(5-6):317–319.
25. Norman KA, Newman E, Detre G, Polyn S (2006) How inhibitory oscillations can train neural networks and punish competitors. *Neural Comput* 18(7):1577–1610.
26. Norman KA, Newman EL, Detre G (2007) A neural network model of retrieval-induced forgetting. *Psychol Rev* 114(4):887–953.
27. Johnson MR, Mitchell KJ, Raye CL, D'Esposito M, Johnson MK (2007) A brief thought can modulate activity in extrastriate visual areas: Top-down effects of refreshing just-seen visual stimuli. *Neuroimage* 37(1):290–299.
28. Smith TA, Hasinski AE, Sederberg PB (2013) The context repetition effect: Predicted events are remembered better, even when they don't happen. *J Exp Psychol Gen* 142(4):1298–1308.
29. Anderson MC (2003) Rethinking interference theory: Executive control and the mechanisms of forgetting. *J Mem Lang* 49:415–445.
30. Kuhl BA, Dudukovic NM, Kahn I, Wagner AD (2007) Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nat Neurosci* 10(7): 908–914.
31. Kuhl BA, Rissman J, Wagner AD (2012) Multi-voxel patterns of visual category representation during episodic encoding are predictive of subsequent memory. *Neuropsychologia* 50(4):458–469.
32. Efron B (1979) Bootstrap methods: Another look at the jackknife. *Ann Stat* 7(1):1–26.

Supporting Information

Kim et al. 10.1073/pnas.1319438111

SI Text

Overview of Pattern Similarity Analyses. In the pattern classification analyses reported in the main text, we quantified the prediction of C by measuring the amount of classifier evidence for the category of the C item. To provide converging evidence for our claim that context-based prediction error induced forgetting, we repeated these analyses using pattern similarity as a complementary measure of item prediction. Here, we quantified the prediction of C by taking the pattern of activity elicited by the perception of C in the initial triplet and measuring how strongly this pattern was reinstated in the repeated triplet. Pattern similarity was quantified as the spatial correlation in activity over voxels in category-selective ventral temporal cortex during the perception of C and during the repeated triplet (when participants might be predicting C). Using logistic regression, we then related the correlation for a given C item to whether it was remembered or forgotten in the subsequent memory test.

One potential advantage of this approach is that our measure of prediction of C was based on the pattern match to the perception of the same item. This analysis stands in contrast to the main pattern classification analyses, in which the classifier was trained on different items from the same category presented in a separate localizer run. Because the pattern similarity approach computes the match to patterns for specific C items, it can, in principle, be used to assess how prediction of the C item itself affects subsequent memory. In practice, accomplishing this goal is complicated by the fact that pattern similarity reflects both item-specific information (1, 2) and also category information. For example, if a pattern is defined over both face- and scene-selective visual cortex, then different face exemplars will have higher pattern similarity with each other than with any scene.

To verify that the relationship between pattern similarity and subsequent memory reflected item-specific information, we performed a within-category permutation test in which we repeatedly shuffled the pairing of initial and repeated triplets for all C items from the same category. This approach allowed us to test whether the relationship between pattern similarity and subsequent memory is stronger when the item matches than when only the category matches.

Relating Pattern Similarity to Subsequent Memory. To perform the logistic regression for pattern similarity, we extracted the patterns of activity evoked by the perception of C in the initial triplet (4.5 s after its onset) and by the prediction of C in the repeated triplet (3–16.5 s after triplet onset). These patterns of activity for C items were obtained from within the corresponding category-sensitive region of interest (ROI). That is, we extracted the patterns from the temporal fusiform cortex when C was a face and from the parahippocampal gyrus when C was a scene. This selection of the category-selective ROIs reduced the amount of variance over voxels in the pattern attributable to category information (thereby isolating item-specific information). These anatomical regions have been used previously as face- and scene-selective ROIs, respectively (3); they include the peaks of the fusiform face area and parahippocampal place area, respectively, as well as surrounding category-selective voxels that might contribute to representing individual exemplars (4–6).

We then correlated the patterns from perception and prediction for a given C item to obtain a measure of pattern similarity. We interpreted greater correlation as more prediction of C, given that C was never presented in the repeated triplet and that patterns were obtained from the ROI selective for the category of C (note that A, B, and D were from the other category). As before, we used logistic regression to relate this new measure of prediction strength to subsequent memory (Fig. S3*A*). As hypothesized, greater pattern similarity for C was associated with more forgetting ($P = 0.034$) (Fig. S3*B*). Separately correlating the initial C pattern with the same baseline, A, B, and D time windows used for the classification analyses revealed that this trend grew more negative over time in the repeated triplet (Fig. S3*C*). These pattern similarity results provide a clear replication of the pattern classification results.

Ruling out Carryover Effects. One potential concern with the pattern similarity analysis is that the same A and B context items were shown in both the initial and repeated triplets. If the pattern obtained during the perception of C in the initial triplet was contaminated by lingering traces of A and B, then the observed pattern similarity in the repeated triplet could reflect the repeated perception of A and B rather than the prediction of C. The use of an ROI selective for C but not A and B mitigates this concern. To rule out this possibility further, we performed additional control analyses that seeded the pattern similarity analysis with patterns obtained from the time of A and B in the initial triplet. If the repeated perception of these items drove the negative relationship, then A and B seeds should yield the same results. However, no reliable negative relationship with memory for C was observed when pattern similarity was based on A or B seed patterns (all Ps > 0.27).

Item-Specific Permutation Analysis. In addition to replicating the main results, the pattern similarity analysis provides additional explanatory leverage. By definition, the pattern classification analysis described in the main text identified information only about the category of C items. In contrast, the pattern similarity analysis also might be sensitive to information about specific exemplars within each category. To examine this possibility, we conducted a permutation analysis by shuffling the pairing of initial and repeated triplets so that the category of C was preserved. A negative relationship between pattern similarity and subsequent memory that was stronger when the exemplar matched than when only the category matched would decisively support our interpretation that forgetting reflects misprediction of the C item itself.

To perform the permutation analysis, we first identified pairs of initial and repeated triplets that shared the same context items and separated them based on whether the C item was a face or a scene. We then shuffled these pairings 1,000 times within each category, each time calculating pattern similarity across the scrambled pairs and relating it to subsequent memory for the C item in the initial triplet with logistic regression. Based on the resulting null distribution of 1,000 beta coefficients, a z-score for the true beta coefficient (when the context items were aligned) was calculated. To assess the random-effects reliability of this z-score across participants, we used the same kind of bootstrap test that was performed for the basic logistic regression analysis of classifier evidence and pattern similarity. Specifically, we resampled entire participants with replacement and performed the same permutation test on the resampled data. The distribution of z-scores across bootstrap samples provides a population-level confidence interval (CI) on these z-scores.

We found a significantly stronger negative trend at the time of D for the true pairing of initial and repeated triplets, relative to the null distribution acquired from shuffled data (bootstrap

$P = 0.036$). Because triplet pairings were shuffled within category, this result provides evidence that activation of the specific C exemplar (above and beyond activation of C's category) contributed to forgetting.

Curve-Fitting Analysis. We used the probabilistic curve induction and testing (P-CIT) Bayesian curve-fitting algorithm (7) to estimate the shape of the "plasticity curve" relating item activation during the incidental encoding phase (indexed by category classifier evidence) to recognition during the subsequent memory test. The P-CIT algorithm approximates the posterior distribution over plasticity curves (i.e., which curves are most probable, given the neural and behavioral data). P-CIT generates this approximation via the following three steps: First, the algorithm defines a parameterized family of curves (piecewise-linear curves with three segments) and randomly samples 100,000 curves from this parameterized space. Importantly, this family of curves includes some curves that fit with the nonmonotonic plasticity hypothesis and other curves that do not fit. Second, for each randomly generated curve, the algorithm assigns an importance weight to the curve that explains how well the curve explains the observed relationship between neural and behavioral data. Finally, these importance weights are used to compute the probability of each curve, given the neural and behavioral data.

Perception and prediction were treated as distinct learning events, both of which could affect subsequent memory. For each one of the randomly sampled curves, we used that curve, coupled with perception- and prediction-strength values (measured using the classifier), to generate predictions about which C items would be remembered or forgotten. Specifically, for each item, we separately computed the expected effect of perception (by taking the measured perception strength and evaluating the sampled plasticity curve at that value) and the expected effect of prediction (by taking the measured prediction strength and evaluating the sampled plasticity curve at that value). To estimate the probability that the item would be remembered or forgotten, we summed the expected effects of perception and prediction and fed this sum into a logistic function (the parameters of which were estimated by the model), giving us an estimated probability of successful recognition for that item. For each sampled curve, we compared these estimated probabilities of successful recognition (for each item) with the actual recognition outcomes and assigned an importance weight to the curve reflecting how well the estimated recognition outcomes fit with the actual outcomes. This importance-weight value summarizes how well that particular curve explains the observed relationship between neural data (i.e., classifier measurements of perception and prediction) and behavioral data.

After assigning importance weights to each of the 100,000 sampled curves, we generated a new set of samples by taking the best curves from the previous generation (i.e., the curves with the highest importance weights) and distorting them slightly. From this point forward, we alternated between assigning importance weights to sampled curves and generating new sampled curves based on these importance weights. We repeated this process for 100,000 iterations (7).

The collection of weighted curves generated by this process can be interpreted as an approximate posterior probability distribution over curves; the weight of a curve is thus proportional to its probability. To generate a mean predicted curve, we averaged together the sampled curves in the final generation of samples, weighted by their importance values (Fig. S4). We also computed credible intervals to indicate the spread of the posterior probability distribution around the mean curve. We did so by evaluating the final set of sampled curves at regular intervals along the x axis (i.e., item activation). For each x coordinate, we computed the 90% credible interval by finding the range of y values that contained the middle 90% of the curve probability mass.

We also computed $P(\textit{theory consistent})$, the overall posterior probability that the true plasticity curve fits with our theory (i.e., that it is U-shaped). To compute this probability, we first labeled each sampled curve as theory-consistent or -inconsistent. Curves were labeled as theory-consistent if they showed a "dip," i.e., the curve dropped below its starting point and then rose above that starting point, moving from left to right. We then calculated the proportion of posterior probability mass taken up by theory-consistent samples. To compute this value, we summed together the importance weights associated with theory-consistent samples. This number provides an efficient summary of how well the data support the nonmonotonic plasticity hypothesis. The Matlab code used to perform the analyses can be downloaded from http://code.google.com/p/p-cit-toolbox.

Relating Repetition Suppression to Classifier Evidence. We quantified repetition suppression by first extracting the activity evoked by the presentation of A and B in the initial and repeated triplets (4.5–9 s after triplet onset) for each voxel in category-selective anatomical ROIs (i.e., in temporal fusiform cortex when A and B were faces and in parahippocampal gyrus when they were scenes), and then we averaged over voxels within each ROI and performed the subtraction of initial minus repeated triplets. Classifier evidence for C and D was obtained in the same manner as in the main analysis but only during the time window when A and B were processed. As before, we used logistic regression to relate the amount of repetition suppression for A and B to the classifier evidence for each of the C and D categories.

Ruling out Effects of Novelty. The behavioral data nicely fit our memory-pruning hypothesis. However, one concern is that the pattern of results (lower memory for C items than for D items) might be confounded by the novelty of the preceding items in the trial sequence (i.e., A and B were novel before C and were repeated before D). For example, the difference between C and D may reflect enhanced encoding of D because it stood out as novel against a context of old items and/or reduced encoding of C because the preceding new items captured attention. There are theoretical and empirical reasons to think that contextual novelty cannot explain our results.

First, it was shown recently (8) that novel items facilitate the formation of new memory representations for a subsequent item (pattern separation), whereas preceding familiar items engage retrieval of existing memory representations and thus reduce encoding (pattern completion). Because C was preceded by novel items and D by familiar items, this study would predict better memory for C than D, whereas we observed the exact opposite.

Second, a behavioral pilot study we ran controlled for novelty but observed the same forgetting effect for mispredicted items. The design was quite similar to the reported functional magnetic resonance imaging study, but we used pairs of scene images (e.g., A→B and C→D) instead of triplets, and these pairs repeated several times. The prediction of the second item (B) based on the first item (A) was violated by swapping the first and the second items across pairs on the fourth repetition (A→B, A→B, A→B, A→D and C→D, C→D, C→D, C→B). Other pairs (e.g., E→F) were repeated four times intact as a control condition (E→F, E→F, E→F, E→F). We measured subsequent memory for the second items in the violation condition (B and D) and the control condition (F). Critically, context items in both conditions (A, C, and E) had equal frequency on the fourth repetition when predictions could be violated. Nevertheless, memory in the violation condition was significantly lower than in the control condition ($P = 0.017$).

Third, although C and D were, by definition, preceded by novel and familiar items, the X items had variable contexts. We

therefore examined memory for X items as a function of the number of preceding novel items: (*i*) preceded by a repeated ABD triplet (one novel, two repeated); (*ii*) preceded by a repeated ABD triplet and another X item (two novel, two repeated); (*iii*) preceded by an initial ABC triplet (three novel); and (*iv*) preceded by an initial ABC triplet and another X item (four novel). If contextual novelty impaired encoding (e.g., for C vs. D), then memory for X items should decrease as a function of the number of preceding novel items. However, there was no effect of these conditions (Ps > 0.17). Indeed, X memory was numerically highest in the fourth condition, which had the most preceding novelty.

Finally, to test further whether the familiarity of items preceding D boosted its encoding, we measured repetition priming for the A and B items in the initial minus repeated triplets and then related these priming scores to D memory using logistic regression. Not surprisingly, we obtained overall repetition priming in response times (Ps < 0.001): Participants judged the subcategory of A and B items faster when they were repeated than when they were novel (45.46 and 47.13 ms faster, respectively). The familiarity account predicts a positive relationship between repetition priming and D memory, which was not obtained; also, there was no relationship with C memory (Ps > 0.16).

All these findings suggest that our behavioral data result from a detrimental effect of prediction violation rather than an effect of the novelty/familiarity of preceding context items.

Ruling out Effects of Serial Position. During incidental exposure, C items appeared earlier in the trial sequence than D items (by definition) and X items (which were distributed uniformly), resulting in a systematic serial position difference across conditions (Ps < 0.001). This potential confound may explain why memory was worse for C items than for D and X items.

We were sensitive to this issue when designing the experiment and attempted to minimize it in two ways. First, the incidental encoding phase was divided into three runs (each lasting around 9 min), and triplets repeated within run such that the conditions were spread across all runs. Second, there was a 10-min rest period between the encoding and test phases to attenuate recency effects.

Nevertheless, we conducted control analyses to rule out a contribution of serial position to our results empirically. One analysis examined (across trials) whether memory for C or D items could be predicted from their serial position. The logic was that if serial position was solely responsible for the observed behavioral differences, then serial position should be related to subsequent memory within these conditions. However, there was no relationship for either C or D items (Ps > 0.15).

We also conducted a subsampling analysis in which we reversed the serial position bias by selecting C and D items so that the average serial position of D items was earlier than C items. Specifically, we deleted pairs of the earliest remaining C item and the last remaining D item within subject until the reversal occurred. As a manipulation check, this procedure did result in an earlier serial position for D than C (P < 0.001). Nevertheless, the behavioral results from the remaining trials were identical to the original pattern, with worse memory for C than for D (P = 0.040).

These findings rule out serial position as an explanation of the behavioral data and remain consistent with our pruning hypothesis.

Arbitrating Between Memory Weakening and Interference During Retrieval. Our preferred explanation for the decreased recognition of C items relative to D and X items is that memory for C items was weakened. However, other potential explanations exist. For example, it is possible that when C was predicted (but did not appear) during the repeated triplet, participants encoded that "C is absent"; later, during the recognition test, this "C is absent" memory trace might have been activated, competing with the original C memory and reducing recognition confidence.

When we consider the largely implicit nature of our study, however, we think that this alternative "memory for absence" account cannot fully explain our results. Stimuli in the incidental encoding phase were presented in a continuous stream, and thus it was impossible to know a priori whether an item was the A, B, or C item in a triplet. Meanwhile, participants performed an orthogonal categorization task on these stimuli, and they thought that measuring performance on this task was the purpose of the study. Additionally, the context items (A and B) were repeated only once. Thus, it was extremely hard for a participant to detect any structure in the stimulus sequence. After completing the study, we anecdotally asked the participants whether they noticed any regularity during the study phase. Although their answers were not recorded systematically, no participant reported explicit awareness of the triplet structure. In other words, participants were likely not aware of the absence of C, and, consequently, it is unlikely that they formed a declarative trace of the thought "C is absent."

Furthermore, even if participants did form such a trace, knowing that C was predicted-but-absent necessarily implies that C was presented earlier in the experiment. This knowledge should increase rather than decrease confidence that C was studied. For example, imagine that you met a person at a bus stop yesterday, and you notice that the person is not at the bus stop today. Later, you run into the person again unexpectedly. Intuitively, the additional declarative trace of the absence would help you remember that you have met that person before, rather than impairing the memory.

Another, related possibility is that when C was predicted during the repeated triplet, the C representation was bound to the D representation. Later, when C was presented at test, it activated the D representation, thereby causing interference and reducing recognition confidence. This account seems unlikely for two reasons. First, recognition memory tests are thought to provide direct access to stored memory traces and thus to be relatively impervious to these kinds of retrieval-interference effects (9), compared with tests of cued recall. Second, this account also predicts that C should interfere with D (i.e., when D was presented at test, the C representation should have come to mind, causing interference and reducing recognition confidence for D), but this was not the case: Memory for D items did not differ statistically from memory for control X items (P = 0.63) and was, in fact, numerically higher.

1. Xue G, et al. (2010) Greater neural pattern similarity across repetitions is associated with better memory. *Science* 330(6000):97–101.
2. Schapiro AC, Kustner LV, Turk-Browne NB (2012) Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Curr Biol* 22(17): 1622–1627.
3. Kuhl BA, Rissman J, Wagner AD (2012) Multi-voxel patterns of visual category representation during episodic encoding are predictive of subsequent memory. *Neuropsychologia* 50(4):458–469.
4. Nestor A, Plaut DC, Behrmann M (2011) Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proc Natl Acad Sci USA* 108(24): 9998–10003.
5. Bonnici HM, et al. (2012) Decoding representations of scenes in the medial temporal lobes. *Hippocampus* 22(5):1143–1153.
6. Verosky SC, Todorov A, Turk-Browne NB (2013) Representations of individuals in ventral temporal cortex defined by faces and biographies. *Neuropsychologia* 51(11):2100–2108.
7. Detre GJ, Natarajan A, Gershman SJ, Norman KA (2013) Moderate levels of activation lead to forgetting in the think/no-think paradigm. *Neuropsychologia* 51(12):2371–2388.
8. Duncan K, Sadanand A, Davachi L (2012) Memory's penumbra: Episodic memory decisions induce lingering mnemonic biases. *Science* 337(6093):485–487.
9. Tomlinson TD, Huber DE, Rieth CA, Davelaar EJ (2009) An interference account of cue-independent forgetting in the no-think paradigm. *Proc Natl Acad Sci USA* 106(37): 15588–15593.

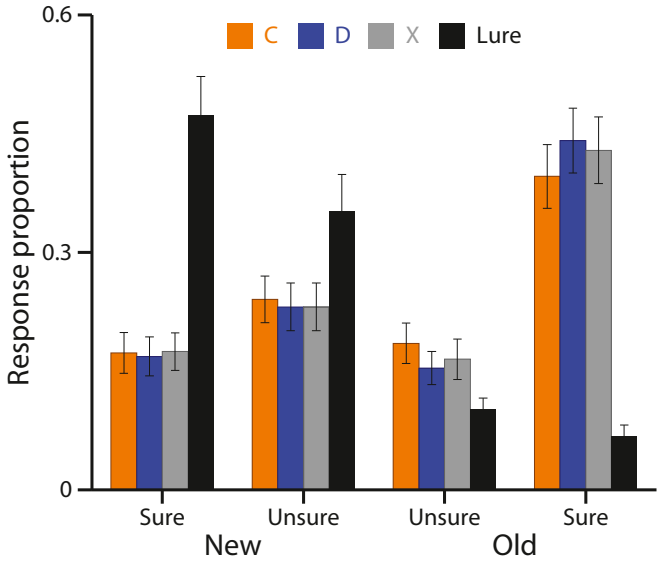

Fig. S1. Behavioral recognition memory. Response proportions for old items from incidental encoding (C, D, X) and new items (Lure). The four response options are shown on the *x* axis. Error bars reflect ±1 SEM.

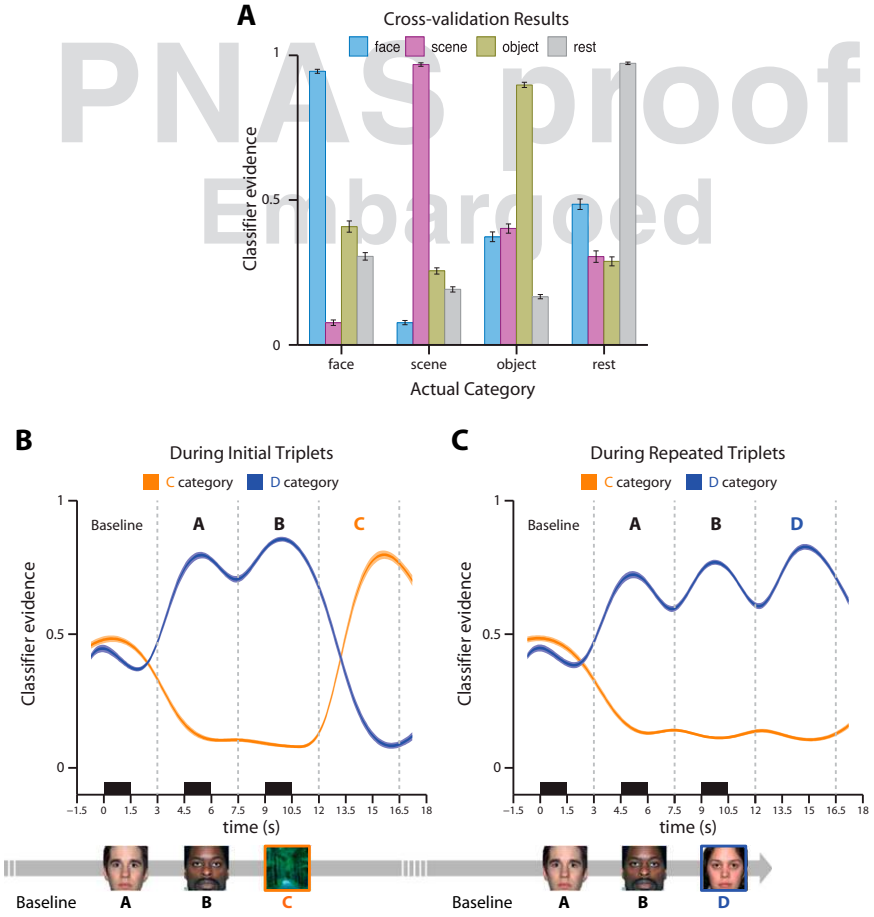

Fig. S2. Details of pattern classification analyses. (*A*) Cross-validation results by category from the localizer. Error bars reflect ± 1 SEM. (*B*) Trajectories over time of classifier evidence for stimulus categories in initial triplets. On the *x* axis, time = 0 indicates the actual time of stimulus onset (i.e., not shifted to account for hemodynamic lag). Classifier evidence peaked around 4.5 s after stimulus onset. Discrete data points were interpolated for visualization. Ribbons reflect ±1 SEM. (*C*) Same trajectories in the repeated triplets.

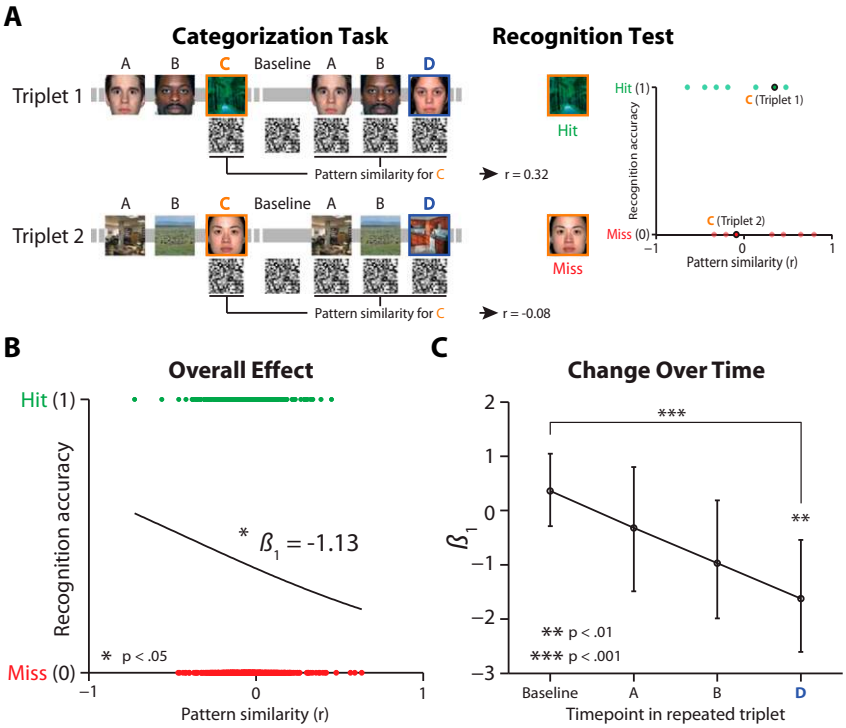

Fig. S3. Pattern similarity analyses. (*A*) Pattern similarity was computed as the Pearson correlation between the patterns of voxel activity from the initial triplet when C was perceived and from the repeated triplet when C could have been predicted. The resulting coefficient for each triplet then was related to subsequent memory for C. (*B*) The pattern similarity for C was first averaged over all time points in the repeated triplet. Dots indicate the distribution of similarity for remembered (green) and forgotten (red) items. There was a reliably negative logistic trend, with greater pattern similarity associated with more forgetting. (*C*) The same analysis was performed separately during the baseline, A, B, and D time periods. The negative relationship was maximal during the anticipated time of C. Error bars reflect 95% bootstrap CIs.

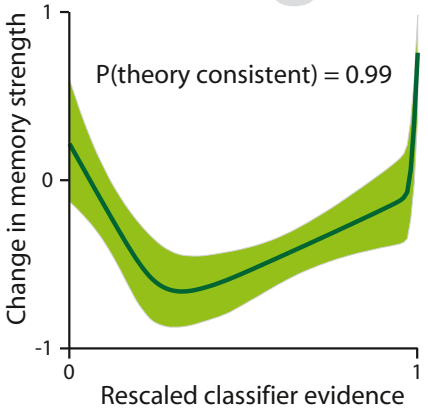

Fig. S4. Curve-fitting analysis. Empirically derived estimate of the plasticity curve relating classifier evidence to subsequent memory performance obtained using the P-CIT curve-fitting algorithm (7). Behavioral outcomes on the recognition memory test were modeled as the summed effects of perception strength (during the initial triplet) and prediction strength (during the repeated triplet). The *x* axis shows rescaled classifier evidence (0 = minimum observed classifier evidence; 1 = maximum observed classifier evidence), and the *y* axis represents the change in subsequent memory strength. The solid green line depicts the mean of the posterior distribution over curves, and the ribbon shows the 90% credible interval (so that 90% of the curve probability mass lies within the ribbon). P-CIT also returns the overall posterior probability that the curve has a U-shape (as predicted by the nonmonotonic plasticity hypothesis); in this case *P*(*theory consistent*) = 0.99.