Suppressing unwanted memories reduces their unconscious influence via targeted cortical inhibition

Pierre Gagnepain[a,b,c,d], Richard N. Henson[e], and Michael C. Anderson[e,f,1]

[a]Institut National de la Santé et de la Recherche Médicale and [d]Centre Hospitalier Universitaire, Unité 1077, 14033 Caen, France; [b]Université de Caen Basse–Normandie and [c]Ecole Pratique des Hautes Etudes, Unité Mixte de Recherche S1077, 14033 Caen, France; [e]Medical Research Council Cognition and Brain Sciences Unit, Cambridge CB2 7EF, England; and [f]Behavioural and Clinical Neuroscience Institute, University of Cambridge, Cambridge CB2 3EB, England

Edited by Jonathan Schooler, University of California, Santa Barbara, CA, and accepted by the Editorial Board February 10, 2014 (received for review June 20, 2013)

Suppressing retrieval of unwanted memories reduces their later conscious recall. It is widely believed, however, that suppressed memories can continue to exert strong unconscious effects that may compromise mental health. Here we show that excluding memories from awareness not only modulates medial temporal lobe regions involved in explicit retention, but also neocortical areas underlying unconscious expressions of memory. Using repetition priming in visual perception as a model task, we found that excluding memories of visual objects from consciousness reduced their later indirect influence on perception, literally making the content of suppressed memories harder for participants to see. Critically, effective connectivity and pattern similarity analysis revealed that suppression mechanisms mediated by the right middle frontal gyrus reduced activity in neocortical areas involved in perceiving objects and targeted the neural populations most activated by reminders. The degree of inhibitory modulation of the visual cortex while people were suppressing visual memories predicted, in a later perception test, the disruption in the neural markers of sensory memory. These findings suggest a neurobiological model of how motivated forgetting affects the unconscious expression of memory that may be generalized to other types of memory content. More generally, they suggest that the century-old assumption that suppression leaves unconscious memories intact should be reconsidered.

inhibitory control | repetition suppression | think/no-think | dynamic causal modeling | representational dissimilarity analysis

Remembering the past is not always a welcome experience. The years of our lives bring unpleasant and even traumatic events that most people would prefer to forget. When reminded of such an event, people often intentionally limit awareness of the unwelcome memory. Over the last decade, evidence has grown that people's efforts to suppress unwelcome memories can impair conscious recall of the suppressed event (1, 2). Suppression engages control mechanisms that reduce momentary awareness of a memory and impair its later conscious recall, a process supported by the right middle frontal gyrus (MFG) (3–6). Suppressing retrieval in this manner reduces hippocampal activity (3–6), especially when unwanted memories intrude into awareness and need to be purged (6). Thus, suppression impairs conscious retention by modulating brain activity in structures known to be involved in recollection. The capacity to control retrieval in this manner may be essential to adapting memory in the aftermath of traumatic life experience.

Although people have some control over whether memories are consciously remembered, suppression's effects on unconscious expressions of memory remain largely unknown. Determining how suppression affects unconscious memory is important to understand its impact on mental health. On the one hand, disrupting conscious access to an experience may leave unconscious memory intact. Research on organic amnesia indicates that even when conscious memory is lacking, an experience can influence behavior through learning supported by brain systems outside the medial temporal lobes (7–9). The learning underlying affective conditioning and repetition priming, for example, can occur without conscious memory (8, 9). Thus, in healthy individuals, modulating hippocampal activity during suppression might disrupt conscious memory, leaving perceptual, affective, and even conceptual elements of an experience intact. Importantly, the distressing intrusions that accompany posttraumatic stress disorder have, in some theoretical accounts, been attributed to the failure of encoding to integrate sensory neocortical traces into a declarative memory that is subject to conscious control (10). If so, disrupting episodic memory may leave persisting neocortical and subcortical traces that trigger intrusive imagery, thoughts, and emotional responses. A similar concern arises in classical psychoanalytic theory, according to which excluding memories from awareness left them fully intact in the unconscious, where they perniciously influenced behavior and thought (11). Thus, unconscious remnants of a suppressed memory may persist and harm mental health.

On the other hand, it may be premature to presume that unconscious forms of memory would survive efforts to suppress conscious memories. Dissociations between explicit and implicit retention arising in neuropsychological patients may not be good precedents for predicting the effects of motivated forgetting in healthy individuals. In healthy individuals, for example, both hippocampal and neocortical systems are intact and are likely to interact during retrieval, influencing how suppression is accomplished. Suppressing an unwanted memory rich in sensory detail may, for example, involve inhibitory control targeted at both hippocampal and neocortical traces. Targeted neocortical inhibition may arise because in healthy individuals retrieval involves hippocampally driven reinstatement of cortical and subcortical

Significance

After a trauma, people often suppress intrusive visual memories. We used functional MRI to understand how healthy participants suppress the visual content of memories to overcome intrusions, and whether suppressed content continues to exert unconscious influences. Effective connectivity, representational similarity, and computational analyses revealed a frontally mediated mechanism that suppresses intrusive visual memories by reducing activity in the visual cortex. This reduction disrupted neural and behavioral expressions of implicit memory during a later perception test. Thus, our findings indicate that motivated forgetting mechanisms, known to disrupt conscious retention, also reduce unconscious expressions of memory, pointing to a neurobiological model of this process.

Author contributions: P.G., R.N.H., and M.C.A. designed research; P.G. performed research; P.G. performed modeling; R.N.H. contributed to the modeling; P.G. and M.C.A. analyzed data; and P.G., R.N.H., and M.C.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. J.S. is a guest editor invited by the Editorial Board.

[1]To whom correspondence should be addressed. E-mail: michael.anderson@mrc-cbu.cam.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1311468111/-/DCSupplemental.

traces that represent the content of an experience (12–14). This reinstatement of neocortical traces via the hippocampus may arise rapidly and involuntarily, as suggested by recent models of retrieval (15). Indeed, intrusive memories are widely known to evoke unwanted visual, auditory, and even tactile memories of the event (10, 16, 17), probably by reactivating traces in the sensory neocortex (12–14). Theoretical accounts of memory control view inhibition as an activation-dependent mechanism that suppresses intrusive traces (6, 18, 19). Thus, cue-driven reinstatement of sensory features may trigger inhibitory control mechanisms that directly target neocortical traces instead of, or in addition to, hippocampal modulation. Critically, if cortical traces reactivated during intrusions also underlie indirect expressions of memory, suppressing those traces should disrupt implicit memory as well. Supporting this possibility, retrieval suppression recently was found to impair repetition priming for visual objects (20). This finding suggests that inhibitory control is targeted at suppressing intrusive neocortical representations, although the neural basis was not examined.

These considerations led us to hypothesize the existence of an inhibitory control process that directly targets neocortical traces reactivated by cues and that may undermine unconscious expressions of memory. To test this hypothesis, we investigated how suppression might hinder later performance on a task that made no reference to memory, but that could benefit indirectly from neocortical traces (21–24). Following on recent work, we examined whether the content of suppressed memories would become less visible to observers on a later visual perception test (20). To detect difficulties in perception, we adapted the "think/no-think" procedure developed to study how people suppress unwanted memories (1–6, 20) (Fig. 1*A*). The procedure had three steps (*Materials and Methods*): the study phase, think/no-think phase, and perceptual identification phase. During the study phase, participants studied pairs made of a word cue and a photo of an object, and then were trained until they could correctly select the object that went with each cue. Next, they performed the think/no-think task while being scanned with functional MRI (fMRI). On each trial, participants received the cue from one of the pairs (e.g., "duty"), and were asked either to recall its paired object (e.g., "binoculars;" think trials), or instead to prevent the object from entering conscious awareness (no-think trials). For no-think trials, we asked participants not to generate distracting thoughts, but to focus on the reminder, and to suppress the object from awareness if it intruded (5).

After the think/no-think phase, we tested how easily participants could identify the objects amid visual noise. The aim of this perceptual identification phase was to see whether suppressing awareness of the objects had made them harder to see, and whether neural markers of those visual memories would be reduced. We scanned participants with fMRI as they observed changing displays that presented either studied or new objects. Each display first appeared with visual noise obscuring the object, but the object grew visible gradually as we reduced the noise. While observing these displays, participants pressed a button when they could first see and identify the object, and we recorded the time it took them to do so. In general, viewing a stimulus makes it easier to identify the same stimulus later on, a form of repetition priming (21–24). Although this speeded perception may be followed by conscious memory for the repeated stimulus, the repetition priming benefit does not depend on such recognition, occurring undiminished in patients with organic amnesia (8) and in neurologically normal subjects with no conscious memory of the repetition (25, 26). Thus, we expected participants to identify the studied objects more quickly than the new (unprimed) objects, and we interpret this repetition benefit to reflect the unintended influence of memory on perception. We measured repetition priming for objects in the think and no-think conditions, but also for baseline objects that had been studied, but not cued during the think/no-think phase. The latter objects provided a baseline measure of repetition priming

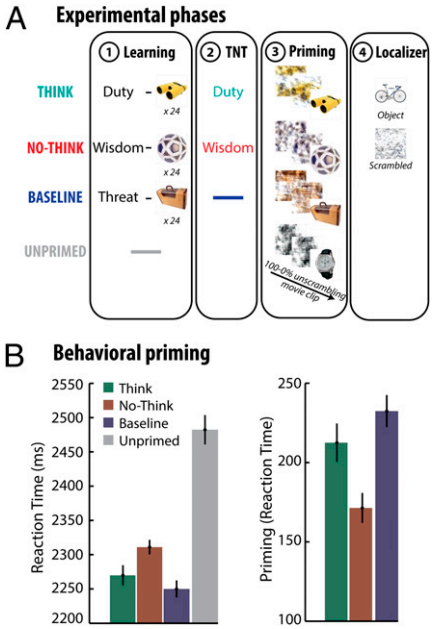

Fig. 1. Behavioral methods and results. (*A*) The procedure. After learning pairs consisting of a word and object picture, participants were scanned during the think/no-think (TNT) task. For think items (in green), participants recalled the associated picture. For no-think items (in red), they tried to prevent the picture from entering awareness. Next, participants were scanned while think, no-think, and baseline (old) objects plus new unprimed objects appeared in a perceptual identification task for degraded objects. In a localizer session, fMRI data from a comparison of new intact and scrambled objects were used to isolate object perception regions. (*B*) Reaction times to identify the scrambled object during the perceptual task (*Left*) and priming effects for studied objects (unprimed − old reaction time; *Right*). Error bars represent within-participant SEs. No-think objects exhibited less repetition priming than did think or baseline objects indicating that suppression disrupted perceptual memory.

in the absence of suppression or retrieval. Consistent with recent work (20), we expected to find that participants were slower to identify no-think than baseline or think objects, reflecting the disruptive effects of retrieval suppression on perceptual identification.

To gauge the existence of cortical inhibition, we first examined whether, during suppression, neocortical regions involved in object processing showed reduced activation for no-think relative to think items. Then we determined whether, in the later perceptual identification test, those same visual regions exhibited aftereffects of suppression on neural priming, an index of perceptual memory (23, 24). If inhibitory control truly disrupts sensory traces in the visual cortex, this disruption should emerge during the perceptual identification test in the form of reduced neural priming for no-think objects, compared with that observed for baseline or think objects. Importantly, we used effective connectivity analyses to evaluate the role of top-down inhibitory modulation of the visual cortex by right MFG during retrieval suppression, and to examine whether individual differences in inhibitory modulation were linked to the predicted disruptions in neural priming during the later perception task. If disrupted neural priming during perceptual identification is an aftereffect of inhibitory control, inhibitory modulation during retrieval suppression should predict this effect. Finally, we tested whether the hypothesized inhibition mechanism was targeted toward neural populations initially most activated by reminders, through a computational model of memory inhibition that we tested with pattern-information analyses (27).

Results

Suppression of Memory Impairs Later Perception. Participants took less time to identify studied objects [Mean (M) = 2,276 ms] than they did unprimed objects (M = 2,482 ms) [$t(23) = -7.2$, $P < 0.001$]. This repetition priming effect indicates that participants could identify studied objects more readily amid distortion, confirming the indirect influence of memory on perception. A one-way ANOVA showed a main effect [$F(1.83,42.03) = 5.744$, $P < 0.01$, Greenhouse–Geiser correction] of the retrieval condition for primed items (think versus no-think versus baseline). The amount of priming was reduced for no-think objects (M = 2,310 ms), which participants identified more slowly than objects from the baseline (M = 2,249 ms) or think (M = 2,269 ms) conditions (Fig. 1*B*). These findings parallel a recent report that suppression impaired the identification of line drawings (20). Thus, as previously shown, when objects reappeared in participants' visual worlds, they found the objects that they had suppressed to be harder to perceive than other recently encountered objects, consistent with the hypothesized disruption of the unconscious influence of visual memories on perception.

Controlled Modulation of the Visual Cortex. Next we investigated whether control processes interacted with the visual cortex to disrupt later memories of the suppressed objects. To do this, we related activation observed when participants suppressed retrieval to the neural signatures of memory during our later perception test. Previous work has found that suppressing retrieval engages a right lateralized frontoparietal network and has highlighted the role of the right MFG in reducing hippocampal activation (3–6). Consistent with this, when we contrasted no-think and think trials at the whole-brain level [*P* family-wise error (P_{FWE}) < 0.05], we observed more activation in a large right-lateralized network (Fig. 2*A* and Table S1), including the MFG, inferior frontal gyrus (IFG), superior frontal gyrus, and inferior parietal cortex. Although we did not observe less hippocampal activation during no-think than think trials in the whole-brain analysis, we did find reductions in activity ($P_{FWE} < 0.05$) when we restricted the search volume to anatomically defined regions of interest (ROIs), i.e., the left and right hippocampus as defined by the Automated Anatomical Labeling (AAL) atlas (28). When we averaged activation across all voxels within those ROIs, this effect was marginal in the left hippocampus [$t(23) = 1.41$, $P = 0.086$] and absent in the right [$t(23) = 0.61$, $P = 0.27$] (however, see *SI Data* regarding an outlier which compromised the significance of this effect). Thus, suppression robustly engaged the brain regions associated with memory control, and this was accompanied by reduced activation in the hippocampus.

Importantly, however, retrieval suppression also reduced activation in the left fusiform gyrus, relative to retrieval ($P_{FWE} < 0.05$ whole brain, Fig. 2*B*). Fusiform gyrus activation has been associated with the perception of visual objects (29, 30), and so reduced activation during no-think trials suggests that suppressing retrieval modulated neocortical regions involved in object perception. To verify that the region in which activation was reduced by suppression was the same as that associated with visual perception, we identified areas associated with object perception in an independent localizer task contrasting activation for intact versus scrambled images of objects (*Materials and Methods*; Fig. 2*B*). Using ROIs defined with this task [fusiform gyrus and lateral occipital complex (LOC); *Materials and Methods*], we found that activity was indeed greater during think than during no-think trials in the left fusiform [$t(23) = 3.18$, $P < 0.01$] but not in the right fusiform [$t(23) = 0.92$, $P = 0.18$] (Fig. 2*C*). We found the opposite pattern of more activation for no-think than think trials in both the left [$t(23) = -2.98$, $P < 0.01$, two-tailed] and right [$t(23) = -2.63$, $P < 0.05$, two-tailed] LOC. Although unexpected, we speculate that this increased LOC activity during no-think trials may have arisen from our instructions to attend to the word cue while suppressing retrieval, and may therefore reflect sustained neural activity in populations coding for word form (31), which may overlap with those coding for objects.

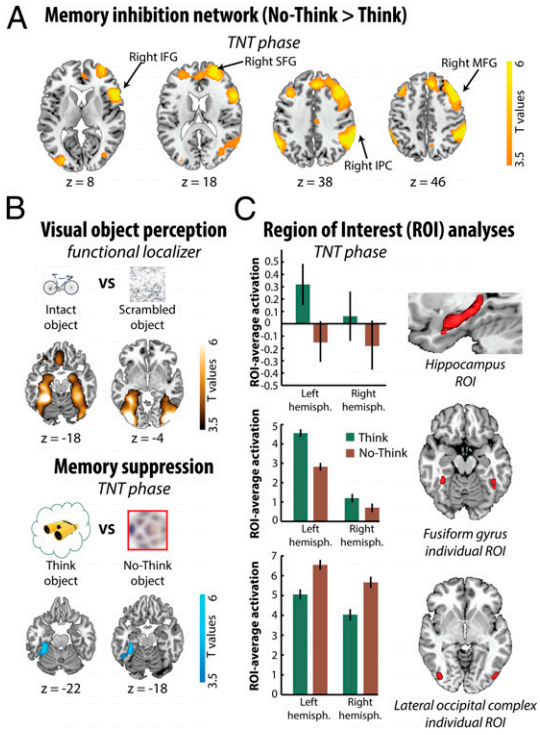

Fig. 2. Brain activity as participants controlled unwanted visual memories. (*A*) Brain areas more engaged by retrieval suppression (no-think > think, thresholded $P < 0.001$ uncorrected, for visualization). (*B*) Suppressing visual object memories modulated neocortical object perception areas. When viewing meaningful objects, people showed more LOC and fusiform cortex activity (*Upper*) compared with viewing visual noise. In overlapping fusiform regions, we observed reduced activity when people suppressed object memories compared with when they retrieved them (*Lower*). (*C*) Suppressing object memories reduced activity in the left hippocampus (*Top*; anatomical ROI) as well as in the fusiform cortex (*Middle*; ROI defined by object perception localizer). LOC (*Bottom*; ROI defined by object perception localizer) showed increased activity during object suppression. (*Right*) Fusiform and LOC ROIs identified for one participant using independent functional localizer. Error bars represent within-participant SEs.

Inhibitory Basis of Memory Control. The amount of fusiform gyrus activation is linked to the degree of conscious awareness that people experience for visual objects during perception (29, 30), and also to whether people remember visual details about consciously remembered objects (13, 14, 32–34). Therefore, reduced activation in this area during no-think trials suggests that participants successfully limited the sensory reinstatement of object memories. One possibility is that the fusiform cortex was simply not engaged during no-think trials. Alternatively, inhibitory control mechanisms mediated by the MFG may have actively suppressed activity in the fusiform cortex during no-think trials, and thereby disrupted perceptual traces. The robust involvement of MFG during retrieval suppression is consistent with the latter possibility. If inhibitory control is involved, it might be possible to measure its aftereffects on the neural signatures of perceptual memory in the fusiform gyrus during our perceptual identification task. Effective connectivity analyses should also reveal that MFG modulates the fusiform gyrus during no-think trials, and that this negative modulation is related to later inhibitory aftereffects observed in the fusiform gyrus.

To determine how retrieval suppression impaired visual perception, we first identified those regions involved in our object perception task that were affected by memory. One robust

marker of perceptual memory is neural priming, or reduced brain activity in areas that process a stimulus, when the stimulus is repeated (23–25). Neural priming has been observed with fMRI in humans (23–25), and with single unit recording in nonhuman primates (35). Strikingly, neural priming occurs even when people do not report conscious memory for previous presentations of a stimulus (25). In our perception task, we observed robust neural priming for studied items (that is, activation in response to think, no-think, or baseline items was less than for new items) in ventral visual stream areas involved in object perception ($P_{FWE} < 0.05$; Fig. 3*A* and Table S2). Confirming this, we also observed neural priming in ROIs identified in our object perception localizer task, including left the LOC [$t(23) = -5.3$, $P < 0.001$], right LOC [$t(23) = -3.97$, $P < 0.001$], left fusiform gyrus [$t(23) = -4.97$, $P < 0.001$], and right fusiform gyrus [$t(23) = -4.94$, $P < 0.001$].

We then examined how retrieval suppression had affected this signature of memory. Importantly, perceptual identification was associated with greater activation for no-think objects than for think and baseline objects when we restricted the search volume to perceptual memory sites in the ventral stream ($P_{FWE} < 0.05$; Fig. 3*B* and Table S2). An additional analysis in which mean identification time differences across conditions were covaried out yielded the same findings (Fig. S1). In ROI analyses, we found greater activation for no-think objects than for baseline objects in the left [$t(23) = 4.13$, $P < 0.001$] and right LOC [$t(23) = 6.6$, $P < 0.001$] and in the left [$t(23) = 3.23$, $P < 0.01$] and right fusiform gyrus [$t(23) = 3.27$, $P < 0.01$] (Fig. 3*C*). Identifying no-think objects also yielded greater activation than did identifying think items in the left [$t(23) = 1.91$, $P < 0.05$] and right [$t(23) = 1.91$, $P < 0.05$] LOC, as well as in the right fusiform gyrus [$t(23) = 2.01$, $P < 0.05$], although not in the left fusiform gyrus [$t(23) = 1.28$, $P = 0.11$]. Activity did not differ reliably between think and baseline objects in the fusiform gyrus ($t < 1$) or the left LOC ($t < 1.2$) (although there was a marginal increase for think relative to baseline objects in the right LOC; $t(23) = -1.8$, $P = 0.08$, two-tailed). It might have been expected that think objects would show greater neural priming than baseline objects, owing to their repeated retrieval from memory during the think/no-think phase; the absence of this effect might be because representations retrieved from memory are not as effective as visual presentations in driving neural priming, or because of saturation of neural priming effects from the repeated exposures of all items during training (36). In summary, neural priming was reduced selectively for no-think items in both the fusiform gyri and the left LOC.

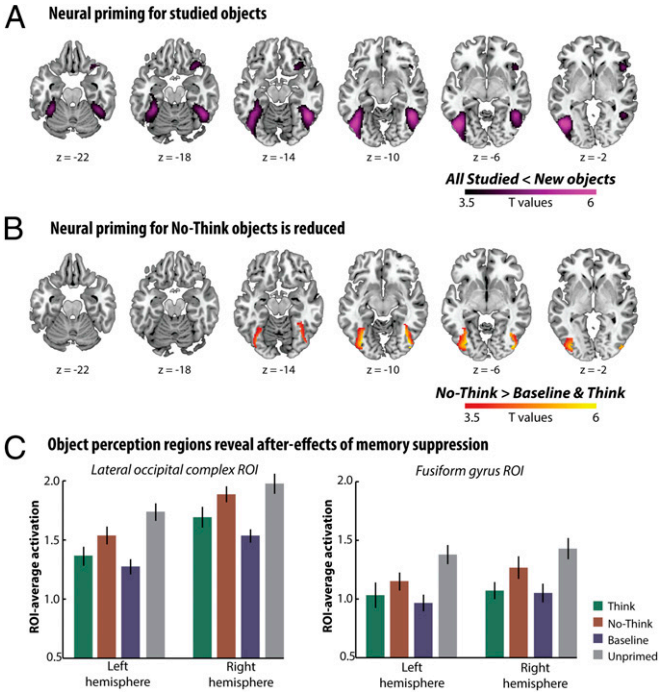

Fig. 3. Brain activity observed during the indirect influence of visual memories on perception. (*A*) The fusiform cortex and the LOC showed less activation during perceptual identification of studied, compared with new objects (neural priming), illustrating the benefits of object memory on neural processes contributing to perception (SPM, thresholded at $P < 0.001$ uncorrected, for visualization). (*B*) Both the LOC and the fusiform cortex show greater activation during the perception of suppressed (i.e., no-think) objects, compared with other studied objects, reflecting the degraded benefit of memory on perception. The SPM showing no-think activation greater than baseline and think activity is masked using the main effect of priming. (*C*) In independently localized ROIs for object-responsive regions of the visual cortex (LOC and fusiform cortex), neural priming for studied objects compared with new objects is partially reversed for no-think objects. Thus, neural activation markers of perceptual memory in the visual cortex were disrupted by memory suppression. Error bars represent within-participant SEs.

Reduced neural priming in the visual cortex suggests that retrieval suppression disrupted the neocortical memory traces for no-think objects, altering their effect on perception. If so, we should find that activity in the fusiform gyrus was reduced by some control process during no-think trials in the think/no-think phase. To test this hypothesis, we used dynamic causal modeling (DCM) (37) and Bayesian model selection (BMS) (38) to test whether the right MFG region, which was previously implicated in memory inhibition, down-regulated activity in the hippocampus and neocortex during no-think trials (*Materials and Methods*). DCM evaluates the effective connectivity between brain areas through a network composed of a small number of key brain regions. A model space is defined by combining (*i*) intrinsic connections between regions in the network, (*ii*) modulatory influences on connections by experimental manipulations, and (*iii*) input sources that drive network activity. These models are mapped onto the fMRI time series using a hemodynamic model of the blood oxygenation level-dependent (BOLD) response, and each of the connectivity parameters estimated. We focused on the potential top-down modulatory influence of the right MFG on the left LOC, fusiform, and hippocampus. The right MFG may modulate the LOC and fusiform gyrus through the inferior frontooccipital fasciculus (39) and modulate the hippocampus via limbic fibers (i.e., cingulum and fornix fibers) (40). In addition, all models were composed of bidirectional connections between the LOC and fusiform, and between fusiform and hippocampus, to respect the hierarchical processing stages of the visual ventral stream.

To focus on top-down modulation, we compared models including an additional top-down modulation during no-think trials to ones in which activation differences across conditions in posterior regions could be solely explained in terms of intrinsic coupling without further modulation (i.e., null models). To generalize the validity of this comparison across distinct patterns of connection or driving inputs, we defined a large model space of 84 networks of differing connections between the four nodes (right MFG, left hippocampus, left fusiform, and left LOC). We then partitioned this model space into four families defined by key model dimensions: (*i*) the pattern of intrinsic connections, which could either be unidirectional or bidirectional between MFG and targeted regions; (*ii*) the entry point of driving input, which could be either the LOC, the MFG, or both: (Fig. S2); (*iii*) the modulatory influence of MFG on targeted regions during no-think trials, which could either be present or absent; and (*iv*) the configuration of regions targeted by MFG, which could include any of the individual regions (LOC, fusiform, or hippocampus), any combination of two regions (LOC + fusiform, LOC + hippocampus, or fusiform + hippocampus), or all three. The differing entry points of driving inputs are meant to represent both the potential influence of visually presented cues in the visual

system and the influence of memory control task instructions in the right MFG. Having defined this model space, we then used family-based inference (38) to restrict the space to the most plausible models (Fig. S2). This method excluded models with unidirectional connections between the MFG and other regions, and also all models that had driving input only to the LOC or only to the MFG. Fourteen models remained (Fig. 4*A*). These models had bidirectional MFG intrinsic connections to either the LOC, fusiform, hippocampus, or some combination of these regions, and input that entered both the LOC and the MFG.

To test the hypothesis that the MFG caused the reduced activation during no-think trials, we compared the remaining models that modulated the connection between MFG and posterior regions during no-think trials (top row in Fig. 4*A*) to those that did not (bottom row in Fig. 4*A*) (i.e., the third family distinction explained above). This analysis overwhelmingly favored models with modulation over models without modulation (exceedance probability = 0.95, expected posterior probability = 0.73). Exceedance probability refers to the extent to which a model is more likely in relation to other models considered, whereas expected posterior probability is the probability of a model generating the observed data. When we did this same comparison separately for each network configuration, five of the modulatory models won decisively against their respective network without modulation, including models in which MFG targeted any one of the individual regions (LOC, fusiform, or hippocampus), one of the two-region models (fusiform and hippocampus), and the model including all three sites (LOC, fusiform, and hippocampus), with an exeedance probabilities of 0.92, 0.85, 0.97, 0.90, and 0.78 and expected posterior probabilities of 0.70, 0.62, 0.76, 0.66, and 0.60, respectively (Fig. 4*A*). We then compared these five remaining modulatory models to assess if there was a preferential pathway by which memory inhibition was achieved, but we found no clear winner (exeedance probabilities of 0.025, 0.05, 0.38, 0.17, and 0.375 for the LOC, fusiform, hippocampus, fusiform + hippocampus, and LOC + fusiform + hippocampus models, respectively). Thus, although our data do not resolve a preferred target of modulation, they do provide strong evidence that retrieval suppression during the think/no-think phase is associated with modulatory signals from the right MFG to posterior brain regions.

These DCM results are consistent with our hypothesis that an inhibitory influence of MFG on the visual cortex disrupts visual object memories, in turn reducing neural priming for those objects in our later perception test. To further evaluate this possibility, we tested whether the degree of negative coupling between MFG and the visual cortex predicted reductions in neural priming. We used Bayesian model averaging (BMA) to

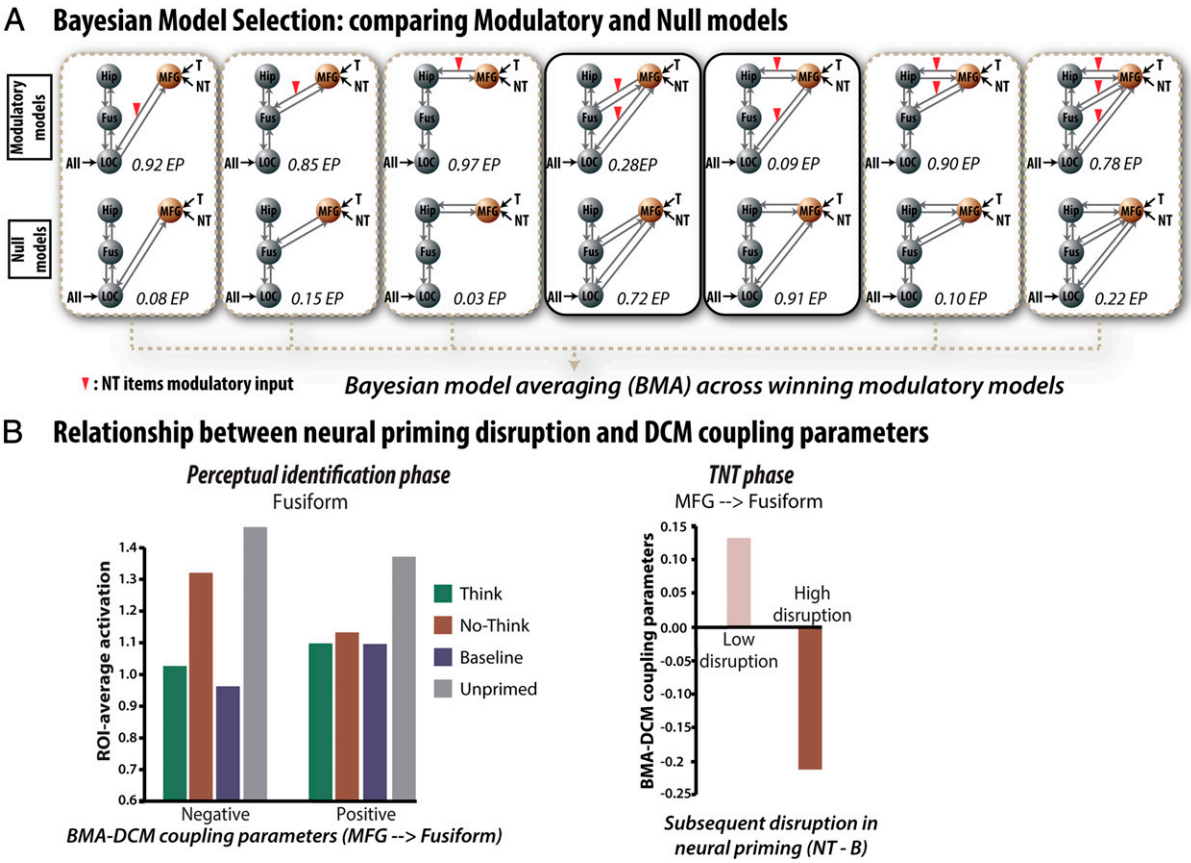

Fig. 4. Effective connectivity underlying the suppression of visual memories, and its impact. (*A*) Potential connectivity from MFG during suppression, and its modulation. Fourteen DCMs remained after model family selection (Fig. S2). DCMs included (*Upper*) or did not include (*Lower*) top-down modulation of activity during no-think trials (red triangles) originating from the MFG, affecting either the left hippocampus, fusiform gyrus, or the LOC or a combination of these regions. Model exceedance probabilities (EPs) comparing the modulatory and nonmodulatory families separately for each modulated site are displayed below each model. BMA was applied to extract and weight coupling parameters for each connection among the successful modulatory models. (*B*) Relationship between modulatory parameters for the MFG→fusiform model and disrupted neural priming for no-think items (i.e., no-think − baseline). (*Left*) Fusiform activation during perceptual identification for participants with low and high negative modulation during the TNT task (n = 12 in each group). (*Right*) The modulatory parameters for participants with either a low or high reduction of neural priming (n = 12 in each group). These findings indicate that the MFG is negatively coupled with the visual cortex during memory suppression, and, importantly, that the degree of negative coupling predicts disruptions in neural priming for suppressed objects during the later perceptual identification task (see also Fig. S3).

extract the DCM coupling parameters for each target region across modulatory models that won against their respective null model (38). BMA weights the parameter estimates within a family (here successful modulatory models) by the posterior probability of the model, and thus provides a single estimate of the coupling between two regions across different network architectures that include those regions. We calculated, for each participant, how much neural priming was reduced for no-think items compared with baseline items (i.e., no-think − baseline averaged ROI activation). As predicted by the inhibition hypothesis, the more negative the coupling between MFG and fusiform gyrus, the more neural priming for no-think items was later reduced compared with baseline priming; robust Spearman $r = -0.56$, $P < 0.05$, 95% CI after bootstrapping: [−0.83 to 0.17]—a method (41) that identifies bivariate outliers and removes them, while accounting for them in calculating CIs. Fig. S3 illustrates this relationship, whereas Fig. 4*B* shows the same relationship when simply performing median splits. Thus, the degree of inhibitory modulation between the MFG and the fusiform gyrus during no-think trials predicted reduced neural priming for those objects on our perceptual test.

Targeted Nature of the Inhibition. If inhibitory control is triggered when reminders elicit unwanted sensory memories, inhibition may be targeted at reducing this unwanted activation, as we have hypothesized (6, 18, 19). Although the evidence described above indicates that inhibition is taking place, it does not imply that this inhibition is activity dependent. For example, instead of inhibiting the most activated voxels in a targeted manner to terminate reinstatement (targeted modulation), the MFG may simply inhibit specific regions of the fusiform gyrus to which the inferior frontooccipital fasciculus projects, irrespective of their level of activation (fixed modulation). Moreover, the capacity to terminate retrieval could reflect an inhibitory mechanism implemented as a down-scaling effect on affected voxels (i.e., inhibition); alternatively, memory reinstatement could initially begin as it does for think items but be stopped instead of directly inhibited, resulting in fewer voxels activated (i.e., truncated reactivation).

To assess the role of these mechanisms, we went beyond overall activation differences between the think and no-think conditions to analyze patterns of activity during retrieval suppression. We used simulation modeling to examine how well the foregoing mechanisms fit the pattern of activity observed in the fusiform cortex. To achieve this, we defined a virtual fusiform cortex and simulated object memories by assigning them different random sets of activation (*SI Simulation Methods*). Each activation pattern is meant to reflect the emerging activity for an object memory, given the initial appearance of its reminder. To simulate the full retrieval of an object, we increased the activation of selected voxels during think trials (i.e., retrieval enhancement). Of interest then is whether the modulatory mechanism that is recruited during no-think trials is targeted at the most active voxels, to terminate this emerging activity, whether it is inhibitory in nature, and if so, how it affects distributed object representations in the fusiform cortex.

We therefore examined nine models that arise from crossing two model characteristics (Fig. 5*A*): (*i*) the voxel selection criterion (either activation based, fixed, or random) and (*ii*) the type of modulation applied to no-think item voxels [inhibition (i.e., down-scaling of activity), truncated reactivation (i.e.,

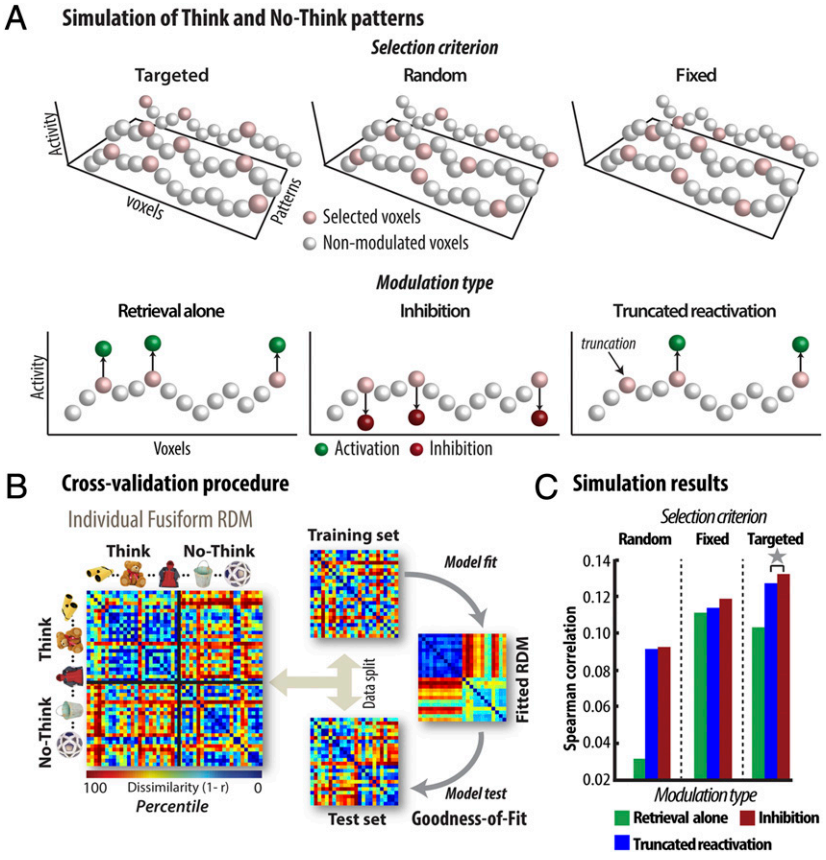

Fig. 5. Simulation of the neocortical activity pattern during the TNT phase. (*A*) Simulation modeling. Initial activation patterns reflect a memory's emerging activity, given its cue (white and pink spheres). White spheres are voxels not modulated by retrieval or suppression. Pink spheres are voxels selected and modulated by retrieval or suppression. The green and the red spheres correspond to the activity level of voxels modulated for think and no-think patterns, respectively, according to modulation type (retrieval alone, inhibition, and truncated reactivation). A targeted model modulates the initially most active voxels in response to reminders; random and fixed models modulate a randomly selected set of voxels, with the former modulating a different set for each pattern, and latter, a consistent set across patterns. Memory control was either absent (retrieval alone mechanism in which no-think items were not modulated) or was implemented by inhibition (down-scaling of activity), or truncated reactivation (activation for some voxels is stopped but not inhibited). Applying combinations of these voxel selection rules and modulation mechanisms to the simulated fusiform cortex allowed us to generate predicted activation data that could be tested against real fusiform data, via a cross-validation procedure (*B*). (*B*) To evaluate each model, we computed each participant's RDM for their left fusiform cortex (e.g., *Left*, displayed here as the rank-ordered RDM of a single participant). In this RDM, each small square represents the dissimilarity (1 − correlation) between the activity patterns across all of the fusiform voxels for two objects that were either recalled (think) or suppressed (no-think). We then randomly split this data into a training set and a test set. We used the training set to estimate the parameters of a given model, applying these parameters to generate a predicted RDM used to test the model against the RDM for the test set (see *Targeted Nature of the Inhibition*, *SI Simulation Methods*, and Fig. S4). (*C*) Simulation outcomes. Each model's GOF averaged across participants (see *Targeted Nature of the Inhibition*, *SI Simulation Methods*, and Fig. S4). These findings show that the pattern of representational similarity in the fusiform gyrus during the think/no-think task is best predicted by a model that posits inhibition, and, in particular, targeted inhibition of a pattern's most active voxels. Gray star illustrates significant difference at 95% confidence interval level tested by bootstrapping.

interruption of retrieval enhancement), or absence of suppression (i.e., retrieval alone)]. To address our activity-dependent hypothesis, we contrasted models in which we modulated those voxels most activated by reminders (targeted, activation-based models), with models in which we modulated randomly chosen voxels (random voxel models), or a fixed set of voxels across items (fixed voxel models), irrespective of their activity. To address the inhibition hypothesis, we contrasted models in which modulation was inhibitory (i.e., down-scaling) with models in which we either stopped reinstatement of no-think activity without further down-regulation (truncated reactivation), or did not modulate no-think activity at all (retrieval alone). We evaluated these models by assessing how each mechanism affected the activity pattern across voxels in its simulated fusiform cortex, and how well this matched what was observed in the data at the individual subject level. Because models varied in how activity was distributed over the voxels assumed to represent each memory, we hypothesized that they would differ in how modulatory mechanisms would alter the similarity relationships between patterns.

To test this, we applied representational similarity analysis (RSA) (42) to the simulated fusiform cortices of each model, and to the data from the left fusiform gyrus of each participant. We computed representational dissimilarity matrices (RDMs), which plotted the degree of representational dissimilarity (1 minus the correlation) between the activation patterns for each of our 48 objects (rows) to the patterns for every other object (columns). To test which of our models provided a better fit to the RDM observed in the left fusiform gyrus, we applied a cross-validation approach, dividing the observed fusiform RDM into a training set and a test set (by splitting half of the items in each condition; Fig. 5*B*). We then fitted each model's parameters to the training set RDM. To measure how well the models fit the data, we computed a theoretically predicted RDM for each model and compared this RDM to the test set RDM (see *SI Simulation Methods* and Fig. S4, Fig. S5, and Fig. S6 for details on the fitting procedure). Repeating this fitting and testing procedure 100 times with a different random split of the data each time allowed us to compute an average goodness of fit (GOF) for each participant and each model that generalized from one half of the data (training set) to the other half (test set). Pairs or families of models were then compared by testing the mean difference in their GOFs across participants (i.e., treating participants as a random effect). The significance of the mean difference was tested by bootstrapping, giving the corresponding 95% confidence intervals (CIs).

To determine the modulation mechanism that best fit the data, we compared inhibition models, as a family, to the truncated reactivation and retrieval alone families (Fig. 5*A*). Both the inhibition model (CI: [0.013, 0.055]) and truncated reactivation (CI: [0.014, 0.048]) families performed reliably better than did retrieval alone model family. Thus, retrieval enhancement of think items on its own is not enough to explain representational dissimilarity in the fusiform cortex and modulation of no-think patterns is necessary. Among models that included a mechanism that modulated the retrieval of no-think items (i.e., inhibition or truncated reactivation), the targeted voxel selection family outperformed both the fixed (CI: [0.0013, 0.027]) and random (CI: [0.016, 0.052]) selection families. This suggests that voxels are selected for modulation based on their level of activity. Critically, the inhibition model also performed reliably better than the truncated reactivation model within the targeted selection family (CI: [0.0009, 0.01]). Taken together, these findings indicate that an inhibition mechanism provides a reliably better account of our data, and, importantly, that this mechanism targets the most active voxels, which are critical to the reinstatement of a memory. The outcome of our DCM analysis indicates that retrieval suppression is best explained by inhibition, and the pattern similarity analyses further specify that this mechanism is activation dependent, consistent with the hypothesis that targeted inhibition is involved in overcoming the influence of intruding sensory experiences.

Discussion

Our findings indicate that when reminders trigger unwanted visual memories, inhibitory control modulates the visual cortex in a targeted way to reduce sensory reactivation. This mechanism limits awareness of the visual memory in the present moment, but also reduces its influence on later indirect tests of memory. The role of inhibitory control is especially clear because we tracked activity in brain regions representing the content of visual memories, both at the moment when awareness was being suppressed, and in a later test in which the neural aftereffects of the inhibition process could be measured. The data from both phases implicate an inhibition mechanism. Suppressing visual memories reduced neural activity in the fusiform cortex, which is linked to awareness of visual objects in perception (29, 30). This reduction arose from negative modulation of this area by the right MFG, and the MFG has been implicated in overriding prepotent responses in general, and in suppressing retrieval in particular (3–6). Reduced neural activity in the fusiform cortex during the suppression of mnemonic awareness was accompanied, in our later perceptual test, by selective reductions in neural priming for the visual objects that had been excluded from awareness. The two phenomena were related: the strength of inhibitory modulation between the MFG and the fusiform cortex during no-think trials in the think/no-think phase predicted the degree of disruption to neural priming in the later perceptual identification phase. Given that neural priming has been associated with unconscious influences of memory (25, 26), these findings are consistent with the possibility that suppressing awareness inhibits sensory memory traces in the visual cortex, thereby reducing their unconscious influence on later perception. On average, inhibition only partially reduced these unconscious influences, however, suggesting an imperfect process that may vary across individuals. This possibility is consistent with the strong individual differences in frontal cortical coupling observed here (Fig. 4*B*), and the relationship of that coupling to the neural aftereffects of suppression.

One unexpected finding was the increased activation in LOC during no-think trials, compared with think trials during the think/no-think task. This finding is unexpected because we had predicted that suppressing retrieval of visual objects would reduce activation in visual object perception regions in general, including LOC. We had not considered, however, that instructions to attend to the visually presented word cue while suppressing retrieval might also influence LOC activation. Focusing intently on the word may have induced attentional gain in populations coding for word form (31). By this hypothesis, reduced neural priming in LOC in the later priming test would reflect the aftereffects of suppression processes that were obscured during the earlier no-think trials by attention to visual word form.

The present evidence for a targeted inhibition mechanism modulating the visual cortex suggests a framework for understanding how memory suppression may influence indirect expressions of memory more generally. Although unwanted memories that intrude into awareness are often visual, the process identified here may extend beyond visual content. A fundamental dynamic of motivated forgetting involves the intrusion of unwelcome content, coupled with the goal of excluding that content from consciousness (1, 3–6). Consistent with this intrusion dynamic, pattern-information analyses favored a targeted, activation-dependent suppression model in which inhibition affected the elements of visual cortical traces reactivated by reminders (43). Personal experiences, however, typically include many different sensory, conceptual, and emotional features, aside from visual attributes. Reminders also reinstate nonvisual features by reactivating the neocortical or subcortical areas representing that content (13, 14, 44, 45). This cue-driven reinstatement of cortical activity would likely arise from reentrant activation driven by the hippocampus.

Activity-dependent inhibitory control processes supported by the MFG (or another control region) may also reduce activity in nonvisual areas. The traces suppressed by this process may underlie other forms of implicit memory, such as nonvisual perceptual priming, conceptual priming, or even affective conditioning. Thus, the mechanisms specified here could undermine indirect expressions of memory more broadly. Alternatively, some types of content may not be suppressible if the pathways linking prefrontal control regions to a representation site do not support modulation. In either case, disruptions of implicit memory should be reactivation dependent; neocortical traces not reactivated by reminders should not be suppressed and may continue to influence behavior.

Research on memory systems in the brain has emphasized dissociations between explicit and implicit memory, with the former supported by the medial temporal lobe, and the latter by distinct cortical and subcortical systems (21, 22, 46, 47). This emphasis suggests that these types of memory are independent. Supporting this possibility, amnesic patients can show a striking lack of conscious memory for an experience, yet reveal its unconscious influences although intact emotional conditioning (9), repetition priming (7, 8, 21, 22), and other forms of implicit memory (46, 47). Despite these strong dissociations, recent evidence indicates that in healthy brains, the hippocampus interacts with neocortical areas not only to support intentional retrieval, but also various forms of implicit memory (48–50), perhaps via a rapid, involuntary reactivation process (15, 51). The present findings similarly imply an interdependency between conscious and unconscious retrieval. Conscious recollection often depends on hippocampal mechanisms that can reactivate diverse cortical and subcortical traces formed during the original experience (13, 47). Indeed, this reactivation is viewed as a central function of the hippocampus during retrieval (47). Intriguingly, we found that the visual cortex representations that were suppressed during attempts to stop the conscious retrieval process are either the same as or interdependent with traces that support the unconscious influences of memory on perception. This dual role of visual cortical representations in hippocampally driven reinstatement and priming suggests that common representations and processes can contribute to explicit and implicit memory (15, 52, 53).

Whether the mechanisms identified here can reduce the unconscious influence of threatening, personally relevant memories remains unknown. Nevertheless, the current findings suggest that the mental operations believed in classical psychoanalysis to banish unwanted memories into the unconscious (11)—where they are immutable, and free to influence behavior—may achieve something quite different. Our data suggest that these operations may instead often reduce a memory's unconscious influence. We found that inhibitory control degraded sensory traces, making the contents of suppressed memories less visible when they reappeared in people's visual worlds. The neural mechanisms underlying these effects may similarly reduce the unconscious influence of other intrusive mental content. If so, it is necessary to reexamine the century-old assumption that suppressing memories necessarily leaves persisting unconscious influences that undermine mental health. There may be a range of conditions under which suppression is an adaptive response to unwanted memories. As a catalyst to understanding these conditions, the present work provides a neurobiological model of how suppressing unwanted memories affects their unconscious influence on behavior, a model that grounds these dynamics in interactions between the memory systems of the human brain.

Materials and Methods

Participants. Twenty-four right-handed native English speakers (13 males) aged between 20–32 y (M = 22.3, SD = 3.9) were paid to participate. They had no reported history of neurological, medical, visual, or memory disorders. The project was approved by the Cambridge Psychology Research Ethics Committee, and all participants gave written informed consent. Participants were asked not to consume psychostimulants, drugs, or alcohol before the experimental period.

Materials. The stimuli were 104 arbitrary word–object pairs composed of abstract English words and artifact objects selected from the http://cvcl.mit.edu/MM/objectCategories.html database (54). Four lists of 24 pairs (assigned to the four conditions: think, no-think, baseline, and unprimed) were created, plus 8 fillers used for practice. The 4 lists were matched on average covert naming latency (derived from a pilot study), and each appeared equally often in all conditions, across participants.

Procedures. Participants learned 80 word–object pairs through a test–feedback cycle procedure with the learning criterion set to 90%. Eight of these were filler pairs reserved for practice on the think/no-think and perceptual identification tasks. The remaining 72 pairs were divided into 3 lists of 24, assigned to think, no-think, and baseline conditions. An additional 24 objects were assigned to the unprimed condition, which appeared during the perceptual identification task. After studying the first 40 pairs for 5 s each, participants were given trials presenting the cue for 3 s, and asked whether they could recall and fully visualize the associated object. If so, 3 objects then appeared (1 correct, and 2 foils taken from other pairs), and they received 4 s to select which picture went with the cue. After selecting an object, the correct answer appeared and they were asked to use this feedback to increase their knowledge of the pair. After testing all pairs in this manner, further test–feedback cycles through the list continued until they reached the criterion of 90% correct. The remaining 40 pairs were then learned in similar fashion. Once participants had reached the learning criterion for both sets of 40 pairs, their memory was assessed a last time using a final criterion test on all of the pairs. The same procedure was used, without feedback. Average performance for this final test was 91%. Only items correctly recalled on this final test were included in later analyses, except for the RSAs, for which it was more convenient to have symmetrical representational dissimilarity matrices across participants. Following learning, participants entered the MRI scanner. A final reminder of all of the pairs appeared during which participants were asked to refresh their memory. This refresher was performed while T1 structural image was collected (see *Imaging Acquisition Parameters*).

Participants then performed the think/no-think task, which was divided into 4 sessions of about 11 min each. Each session presented 24 think and 24 no-think items, twice. Items appeared for 3 s in either green or red (see below), centered on a gray background. Trials were presented in a stochastic fashion with a 2-s average interstimulus interval (ISI) with 30% additional null events and were separated by a fixation cross. The Genetic Algorithm Toolbox (55) was used to optimize both the efficiency of the think versus no-think contrast as well as the estimation of individual conditions against rest. Think cues appeared in green, and participants were told to generate as detailed and complete an image of the associated picture as possible. No-think items appeared in red and participants were told it was imperative to prevent the picture from coming to mind at all, and that they should fixate and concentrate on the cue word without looking away (they knew their eyes were filmed). They were asked to block thoughts of the picture by blanking their mind and not by replacing the picture with any other thoughts or images. If the object image came to mind anyway, they were asked to push it out of mind.

The perceptual identification task followed the think/no-think phase, and tested whether previous attempts at suppression affected repetition priming. It comprised a single session of about 11 min. Each of the think, no-think, baseline, and unprimed items was presented on one trial in a 300 × 300-pixel frame centered on a gray background, and trials were separated by a fixation cross. On each trial, a single item was gradually presented using a phase-unscrambling procedure that lasted for 3.15 s. Participants were instructed to watch carefully as the object was progressively unscrambled, and to press the button as fast as possible the moment they were able to see and identify the name of the object in the picture (1.1% of the trials did not receive any button press). Unscrambling continued until a complete image appeared, irrespective of when and whether participants pressed a button. The scrambling was achieved by decomposing the picture into phase and amplitude spectra using a Fourier transform. Random noise was added to the phase spectrum starting from 100% and was decreased by 5% steps until 0% (i.e., intact picture) was reached. The picture was presented at each level of noise for 150 ms, yielding a total stimulus duration of 3.15 s. Between trials, there was a 2.4-s average ISI, and there were also 20% additional null events added.

Finally, a functional localizer was performed to isolate, for each participant, brain areas involved in perceiving intact objects. Trials during this phase

presented either intact or scrambled objects, and participants simply judged whether the image they were presently viewing matched the one on the last trial (1-back task). The object pictures were not presented in earlier phases and thus were new to the participant. Each image was presented in a 300 × 300-pixel frame on a gray background together. Object and phase-scrambled objects were presented for 1 s (0.5-s ISI) in a blocked fashion (15 s per block). Stimuli were presented using the Psychophysics Toolbox implemented in MATLAB (MathWorks).

Imaging Acquisition Parameters. Scanning was performed on a 3-T Siemens Tim Trio MRI system using a 32-channel whole-head coil. High-resolution (1 × 1 × 1 mm) T1-weighted (magnetization-prepared rapid acquisition with gradient echo) images were collected for anatomical visualization and normalization. Functional data were acquired using a gradient-echo, echo-planar pulse sequence (repetition time = 2,000 ms, echo time = 30 ms, 32 horizontal slices, descending slice acquisition, 3 × 3 × 3 mm voxel size, 0.75-mm interslice gap). The first eight volumes of each session were discarded to allow for magnetic field stabilization.

Preprocessing. Data were analyzed using Statistical Parametric Mapping software (SPM8) (Wellcome Department of Imaging Neuroscience, London). During preprocessing, images were first spatially realigned to correct for motion, before being corrected for slice acquisition temporal delay. Images were then normalized using the parameters derived from the nonlinear normalization of individual gray-matter T1 images to the T1 template of the Montreal Neurological Institute (MNI, Montreal), and spatially smoothed using a 10-mm FWHM Gaussian kernel for univariate analyses. Note, however, that unsmoothed images were used for RSA. The use of unsmoothed images is important for RSA as it preserves the fine-grained spatial pattern that characterizes the representational structure of a region.

ROI Selection. The preprocessed time series in each voxel from the functional localizer were high-pass filtered using a cutoff frequency set at 1/128 Hz. Regressors within a general linear model (GLM) for each voxel were created by convolving a 15-s epoch (boxcar function) for each block with a canonical hemodynamic response function (HRF). Further regressors of no interest were the six realignment parameters to account for linear residual motion artifacts. For each participant, individual peak maxima ($P < 0.05$ uncorrected) were consistently strongest in the bilateral LOC and posterior fusiform gyrus as usually found with the object > scrambled contrast (56). From each peak, an in-house program was then used to select the 100 most significant contiguous voxels separately for each participant (an example of these ROIs can be found in Fig. 2*C*). The average MNI coordinates for the individual peak maxima were as follows: $x = -41$, $y = -80$, and $z = -8$ for the left LOC; $x = 41$, $y = -81$, and $z = -6$ for the right LOC; $x = -36$, $y = -46$, and $z = -18$ for the left fusiform; and $x = 35$, $y = -45$, and $z = -19$ for the right fusiform. These four ROIs were used in all subsequent analyses. In addition, the left and right hippocampi were anatomically defined using the AAL atlas (28).

Think/No-Think and Perceptual Identification Univariate Analyses. The preprocessed time series in each voxel from the main think/no-think and perceptual identification phases were high-pass filtered using a cutoff frequency set at 1/128 Hz. Regressors within a GLM for each voxel were created by convolving a delta function (modeled as an event for the think/no-think task, and as 3.15 s short-epoch for perceptual identification) at stimulus onset for each condition of interest with a canonical HRF. Only items correctly recalled and recognized during the final criterion test preceding the think/no-think task were included in the analyses of the think/no-think and perceptual identification tasks. Further regressors of no interest were the six realignment parameters to account for linear residual motion artifacts, as well as an additional regressor for items not recalled or recognized during the final criterion test, or with no button press in the priming task. Individual parameter estimates were then extracted and averaged in each ROI, and entered into paired *t* tests. A one-tailed *t* statistic was used to test planned comparisons unless otherwise stated. Additional voxel-based analyses were also performed by entering first-level activation maps for each condition of interest into flexible ANOVAs implemented in SPM, which used pooled error and correction for nonsphericity to create *t* statistics. The statistical parametric maps (SPMs) were thresholded for voxels whose statistic exceeded a peak threshold corresponding to a $P_{FWE} < 0.05$ correction across the whole brain or within the appropriate search volumes of interest using random field theory. In Figs. 2 and 3, SPMs were rendered onto a standard brain in MNI space and thresholded at $P < 0.001$ for visualization purposes only.

Think/No-Think DCM Analyses. DCM (37) explains changes in regional activity in terms of experimentally defined modulations (modulatory input) of the connectivity between regions. Here, we used DCM and BMS (38) to assess (*i*) whether the right MFG modulates brain areas involved in recollection during memory suppression and (*ii*) whether the LOC, fusiform, or hippocampus are preferred targets of control. DCM entails defining a network of a few ROIs and the forward and backward connections between them. The neuronal dynamics within this network are based on a set of simple differential equations (the bilinear state equation was used here) relating the activity in each region to (*i*) the activity of other regions via intrinsic connections in the absence of any experimental manipulation, (*ii*) experimentally-defined extrinsic input (or the driving input) to one or more of the regions, and, most importantly, (*iii*) experimentally-defined modulations (or the modulatory input) of the connectivity between regions. Changes in the network dynamics are caused by these driving (entering regions) or modulatory (between regions) inputs. These neural dynamics are then mapped to the fMRI time series using a biophysical model of the BOLD response. The neural (and hemodynamic) parameters of this DCM are estimated using approximate variational Bayesian techniques to maximize the free energy-bound on the Bayesian model evidence. Here, BMS was used to select the preferred model at the group level treating the optimal model across participants as a random effect.

As think versus no-think differences were generally stronger in the left hemisphere, we restricted our DCM to the left LOC and left fusiform (both defined by our independent functional localizer), and the left hippocampus (anatomically defined). Memory inhibition was assumed to originate from the right MFG (see Introduction). The four think/no-think sessions were concatenated, stimulus onsets defined using a delta function modeled as 3-s short-epoch for each condition of interest, and the first eigenvariate extracted in each of the ROIs (i.e., LOC, fusiform, hippocampus, and MFG) and adjusted for effects of no interest (which included the six realignment parameters, sines, and cosines of up to three cycles per run to capture low-frequency drifts, and constant terms to remove the mean of each run). The right MFG was defined in each individual as a sphere of 6 mm, centered at the individual maxima (given by the no-think > think contrast) located within 2.5 times the FWHM of the smoothing kernel of the group maxima (and within the same anatomical structure). The main goal of this analysis was thus to assess whether or not memory inhibition originating from the right MFG was transmitted to posterior regions (i.e., LOC, fusiform, and hippocampus) and under which pathway. The first eigenvariate in those regions was extracted using all voxels composing these ROIs (i.e., no functional thresholding) to ensure that our inferences across univariate, DCM, and representational similarity (see *Think/No-Think RSAs*) analyses were based on the same data (i.e., ROIs including all voxels).

Eighty-four DCM models were created (for an illustration of the model space, see Fig. S2). All models had bilateral connections between the hippocampus and the fusiform cortex, and between the fusiform cortex and the LOC. These 84 models could be divided into 4 model families. The first family of models (the direction family) was divided into those that could have either a unilateral or a bilateral intrinsic connectivity from the MFG to one of the other regions. The second model family (the input family) divided the model space into three subgroups according to the source of the driving input. In the first subgroup of this family, think and no-think stimulation entered the system separately in the right MFG. In the second subgroup, all stimulus items (irrespective of their think/no-think status) entered the system in the LOC. The third subgroup had a combination of both, with think and no-think stimulation entering the right MFG, and all items entering the LOC. The third model family (the modulation family, as shown in Fig. 4*A*) divided the model space into two subgroups that differed according to whether the intrinsic connection from the right MFG was additionally modulated or not by no-think items (modeled here as 3-s short epochs). In other words, this third family included models with modulation of the connection to the hippocampus, the fusiform cortex, or the LOC by retrieval suppression. Finally, the fourth model family (the intrinsic family) divided the model space into 7 groups according to the pattern of intrinsic connections between the MFG and others regions (Fig. 4*A*) which could target (LOC) versus (fusiform) versus (hippocampus) versus (LOC + fusiform) versus (LOC + hippocampus) versus (fusiform + hippocampus) versus (LOC + fusiform + hippocampus).

After estimating all 84 models for each participant, we performed the group BMS as implemented in SPM8 (DCM 10 version). This produces the exceedance probability (i.e., the extent to which each model is more likely than any other model) and expected posterior probability (i.e., the probability of a model generating the observed data). Model selection can critically depend on the space of models used and higher evidence for a given model may be the result of other implausible models. We therefore used the family

inference method (38) to identify the preferred subgroup. Models from the most likely subgroup were then entered into a subsequent BMS, and so on, restricting the model space more and more to plausible models (those with the highest exceedance probability).

Think/No-Think RSAs. Normalized but unsmoothed time series in each voxel were used for this analysis. These time series were concatenated across sessions to improve first-level *t* statistics, which are used to compute the brain RDM (42). Regressors within a GLM for each voxel were created by convolving a delta function (modeled as an event) at stimulus onset of each item separately (with 8 stimulus onsets for a given item), with a canonical HRF. In addition to the 48 regressors corresponding to each individual item, further regressors of no interest were the 6 realignment parameters, sines, and cosines of up to 3 cycles per run to capture low-frequency drifts, and constant terms to remove the mean of each run. Individual *t* maps were then computed by contrasting each item against the rest, and used to compute RDMs in our ROIs. Those individual RDMs were computed using the RSA Toolbox (27) as follows: for each pair of items, the activity patterns in a given ROI were compared using spatial correlation and the dissimilarity was then given by 1 minus the correlation. Individual RDMs of the left fusiform cortex were then used to train and test computational models of memory suppression based on the simulation of a virtual grid of activity.

ACKNOWLEDGMENTS. We thank our volunteers for their participation; the radiographers for help in collecting fMRI data; Ian Charest, who wrote the script for individual ROI definition and provided technical help in the RSA analysis; as well as Nikolaus Kriegeskorte, Bernard Staresina, Taylor Schmitz, and Maria Wimber for their comments and feedback. This work was supported by UK Medical Research Council MC_A060 5PR00 (to M.C.A.) and MC_A060_5PR10 (to R.N.H.).

1. Anderson MC, Green C (2001) Suppressing unwanted memories by executive control. *Nature* 410(6826):366–369.
2. Anderson MC, Huddleston E (2012) Towards a cognitive and neurobiological model of motivated forgetting. *True and False Recovered Memories: Toward a Reconciliation of the Debate*, Nebraska Symposium on Motivation, ed Belli RF (Springer, New York), Vol 58, pp 53–120.
3. Anderson MC, et al. (2004) Neural systems underlying the suppression of unwanted memories. *Science* 303(5655):232–235.
4. Depue BE, Curran T, Banich MT (2007) Prefrontal regions orchestrate suppression of emotional memories via a two-phase process. *Science* 317(5835):215–219.
5. Benoit RG, Anderson MC (2012) Opposing mechanisms support the voluntary forgetting of unwanted memories. *Neuron* 76(2):450–460.
6. Levy BJ, Anderson MC (2012) Purging of memories from conscious awareness tracked in the human brain. *J Neurosci* 32(47):16785–16794.
7. Schacter DL (1987) Implicit memory: History and current status. *J Exp Psychol Learn Mem Cogn* 13(3):501–518.
8. Hamann SB, Squire LR (1997) Intact perceptual memory in the absence of conscious memory. *Behav Neurosci* 111(4):850–854.
9. Bechara A, et al. (1995) Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science* 269(5227): 1115–1118.
10. Brewin CR, Gregory JD, Lipton M, Burgess N (2010) Intrusive images in psychological disorders: Characteristics, neural mechanisms, and treatment implications. *Psychol Rev* 117(1):210–232.
11. Freud S, Breuer J (1966) *Studies on Hysteria*, trans Strachey J (Avon Books, New York).
12. Park H, Rugg MD (2008) The relationship between study processing and the effects of cue congruency at retrieval: fMRI support for transfer appropriate processing. *Cereb Cortex* 18(4):868–875.
13. Danker JF, Anderson JR (2010) The ghosts of brain states past: Remembering reactivates the brain regions engaged during encoding. *Psychol Bull* 136(1):87–102.
14. Wheeler ME, Petersen SE, Buckner RL (2000) Memory's echo: Vivid remembering reactivates sensory-specific cortex. *Proc Natl Acad Sci USA* 97(20):11125–11129.
15. Moscovitch M (2008) The hippocampus as a "stupid," domain-specific module: Implications for theories of recent and remote memory, and of imagination. *Can J Exp Psychol* 62(1):62–79.
16. Hirsch CR, Holmes EA (2007) Mental imagery in anxiety disorders. *Psychiatry* 6(4): 161–165.
17. Reynolds M, Brewin CR (1999) Intrusive memories in depression and posttraumatic stress disorder. *Behav Res Ther* 37(3):201–215.
18. Anderson MC, Spellman BA (1995) On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychol Rev* 102(1):68–100.
19. Anderson MC (2003) Rethinking interference theory: Executive control and the mechanisms of forgetting. *J Mem Lang* 49(4):415–445.
20. Kim K, Yi D (2013) Out of mind out of sight: Perceptual consequences of memory suppression. *Psychol Science* 24(4):569–574.
21. Tulving E, Schacter DL (1990) Priming and human memory systems. *Science* 247(4940): 301–306.
22. Schacter DL, Chiu CYP, Ochsner KN (1993) Implicit memory: A selective review. *Annu Rev Neurosci* 16:159–182.
23. Henson RN (2003) Neuroimaging studies of priming. *Prog Neurobiol* 70(1):53–81.
24. Grill-Spector K, Henson RN, Martin A (2006) Repetition and the brain: Neural models of stimulus-specific effects. *Trends Cogn Sci* 10(1):14–23.
25. Schott BH, et al. (2005) Redefining implicit and explicit memory: The functional neuroanatomy of priming, remembering, and control of retrieval. *Proc Natl Acad Sci USA* 102(4):1257–1262.
26. Dehaene S, et al. (2001) Cerebral mechanisms of word masking and unconscious repetition priming. *Nat Neurosci* 4(7):752–758.
27. Kriegeskorte N (2011) Pattern-information analysis: From stimulus decoding to computational-model testing. *Neuroimage* 56(2):411–421.
28. Tzourio-Mazoyer N, et al. (2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15(1):273–289.
29. Grill-Spector K, Kushnir T, Hendler T, Malach R (2000) The dynamics of object-selective activation correlate with recognition performance in humans. *Nat Neurosci* 3(8): 837–843.
30. Bar M, et al. (2001) Cortical mechanisms specific to explicit visual object recognition. *Neuron* 29(2):529–535.
31. Vinckier F, et al. (2007) Hierarchical coding of letter strings in the ventral stream: Dissecting the inner organization of the visual word-form system. *Neuron* 55(1): 143–156.
32. Woodruff CC, Johnson JD, Uncapher MR, Rugg MD (2005) Content-specificity of the neural correlates of recollection. *Neuropsychologia* 43(7):1022–1032.
33. Ishai A, Ungerleider LG, Haxby JV (2000) Distributed neural systems for the generation of visual images. *Neuron* 28(3):979–990.
34. Stokes M, Thompson R, Cusack R, Duncan J (2009) Top-down activation of shape-specific population codes in visual cortex during mental imagery. *J Neurosci* 29(5): 1565–1572.
35. Miller EK, Desimone R (1994) Parallel neuronal mechanisms for short-term memory. *Science* 263(5146):520–522.
36. Reber PJ, Gitelman DR, Parrish TB, Mesulam MM (2005) Priming effects in the fusiform gyrus: Changes in neural activity beyond the second presentation. *Cereb Cortex* 15(6): 787–795.
37. Friston KJ, Harrison L, Penny WD (2003) Dynamic causal modelling. *Neuroimage* 19(4): 1273–1302.
38. Penny WD, et al. (2010) Comparing families of dynamic causal models. *PLOS Comput Biol* 6(3):e1000709.
39. Martino J, Brogna C, Robles SG, Vergani F, Duffau H (2010) Anatomic dissection of the inferior fronto-occipital fasciculus revisited in the lights of brain stimulation data. *Cortex* 46(5):691–699.
40. Wakana S, Jiang H, Nagae-Poetscher LM, van Zijl PC, Mori S (2004) Fiber tract-based atlas of human white matter anatomy. *Radiology* 230(1):77–87.
41. Pernet CR, Wilcox R, Rousselet GA (2013) Robust correlation analyses: False positive and power validation using a new open source Matlab toolbox. *Front Psychol* 3:a606.
42. Kriegeskorte N, Mur M, Bandettini PA (2008) Representational similarity analysis - connecting the branches of systems neuroscience. *Front Syst Neurosci* 2:4.
43. 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.
44. Hornberger M, Rugg MD, Henson RN (2006) fMRI correlates of retrieval orientation. *Neuropsychologia* 44(8):1425–1436.
45. Gottfried JA, Smith AP, Rugg MD, Dolan RJ (2004) Remembrance of odors past: Human olfactory cortex in cross-modal recognition memory. *Neuron* 42(4):687–695.
46. Gabrieli JDE (1998) Cognitive neuroscience of human memory. *Annu Rev Psychol* 49: 87–115.
47. Squire LR (1992) Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychol Rev* 99(2):195–231.
48. Hannula DE, Ranganath C (2009) The eyes have it: Hippocampal activity predicts expression of memory in eye movements. *Neuron* 63(5):592–599.
49. Wimmer GE, Shohamy D (2012) Preference by association: How memory mechanisms in the hippocampus bias decisions. *Science* 338(6104):270–273.
50. Sheldon SA, Moscovitch M (2010) Recollective performance advantages for implicit memory tasks. *Memory* 18(7):681–697.
51. Hannula DE, Greene AJ (2012) The hippocampus reevaluated in unconscious learning and memory: At a tipping point? *Front Hum Neurosci* 6:80.
52. Henson RN, Gagnepain P (2010) Predictive, interactive multiple memory systems. *Hippocampus* 20(11):1315–1326.
53. Gagnepain P, et al. (2011) Is neocortical-hippocampal connectivity a better predictor of subsequent recollection than local increases in hippocampal activity? New insights on the role of priming. *J Cogn Neurosci* 23(2):391–403.
54. Konkle T, Brady TF, Alvarez GA, Oliva A (2010) Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *J Exp Psychol Gen* 139(3): 558–578.
55. Wager TD, Nichols TE (2003) Optimization of experimental design in fMRI: A general framework using a genetic algorithm. *Neuroimage* 18(2):293–309.
56. Dilks DD, Julian JB, Kubilius J, Spelke ES, Kanwisher N (2011) Mirror-image sensitivity and invariance in object and scene processing pathways. *J Neurosci* 31(31):11305–11312.

Supporting Information

Gagnepain et al. 10.1073/pnas.1311468111

SI Data

This section reports a note on the outlier exclusion procedure during region of interest analyses.

Outliers were defined as 3 SDs above or below the mean of the difference between a priori contrasts of conditions of interest. One outlier participant was identified during the analysis of the think/no-think phase. This participant was an outlier in three regions of interest: the left and right hippocampus and the right fusiform. During the think/no-think phase, we observed increased activity for think relative to no-think items in both left [$t(22) = 2.75$, $P < 0.01$] and right [$t(22) = 1.79$, $P < 0.05$] hippocampus when this outlier was excluded from the analysis; when the outlier was included, this effect was marginal in the left hippocampus [$t(23) = 1.41$, $P = 0.086$] and absent in the right [$t(23) = 0.61$, $P = 0.27$]. In the right fusiform gyrus, we observed increased activity for think relative to no-think items [$t(22) = 1.95$, $P < 0.05$] when excluding this outlier; when included, this effect was no longer significant [$t(23) = 0.92$, $P = 0.18$]. Note that we did not exclude this outlier participant from the analyses reported in the manuscript. Rather, we report them here for reader consideration.

SI Simulation Methods

This section reports in detail (*i*) the models used to simulate fusiform activity under different suppression accounts, (*ii*) the Markov chain Monte Carlo (MCMC) algorithm to sample the entire space of parameter values for each simulated model and to approximate the posterior distribution of these parameters, and (*iii*) the method to test goodness of fit (GOF) of each model across participants, as well as (*iv*) evidence of MCMC convergence.

General Overview. The goal of these simulations was to compute the theoretical representational dissimilarity matrix (RDM_{-t}) for a simulated fusiform cortex under different assumptions of about how memory is suppressed. A first factor distinguishing the models corresponded to how voxels were selected and modulated (i.e., voxel selection): (*i*) targeted (i.e., activity dependent, in which a subset of voxels is modulated for each item based on their higher degree of initial activation), (*ii*) random (in which a randomly chosen subset of voxels is selected for each item), and (*iii*) fixed (in which a randomly selected set of voxels is consistently modulated across items). A second factor captured how memory suppression was implemented via (*i*) inhibition (in which voxel activity is divided by some factor), (*ii*) truncated activation (in which memory reinstatement is stopped but not directly inhibited, resulting in fewer voxels remaining active), and (*iii*) retrieval alone (in which activity for no-think items is not modulated at all (and only think items were modulated; see *Model Construction*). After fitting model parameters for each of these models and each participant, and generating the corresponding RDM_{-t}, we compared which RDM_{-t} provided the best fit to the RDM observed in the real fusiform gyrus (RDM_{-fus}) for each participant. GOF values were then entered into a second-level analysis treating participants as a random effect variable.

To generate the critical theoretical RDM_{-t} for each account, we constructed a model M (e.g., targeted inhibition) given some parameters $\theta_1, \theta_2, \ldots, \theta_N$, which can be formulated as $RDM_{-t} = M(\theta_1, \theta_2, \ldots, \theta_N)$. For each generative model M, we estimated the values of the parameters that best fit the data. Here, we used a MCMC approach to sample the entire space of parameter values and to approximate the posterior distribution of each parameter. Then, the maximum a posteriori estimate (MAP) (i.e., the mode of the posterior distribution) was taken as the best fit of each parameter and these estimates used to establish the GOF of each model. Note that we repeatedly split the observed RDM_{-fus} into two halves so that one half provided a training set used to fit model parameters, and the other half provided a test set to calculate the GOF of the model (i.e., a cross-validation approach).

In this section, we first detail how RDM_{-t} was generated under the different theoretical accounts of memory suppression. We then present the MCMC algorithm used to fit model parameters. Finally, we report the cross-validation method used to estimate the GOF distribution of each theoretical model. All these simulations were performed in MATLAB (MathWorks).

Model Construction. To perform this simulation, we first created a G_{vi} grid (with rows corresponding to voxels, *v*, and columns to items, *i*) of random values drawn for each voxel from a multivariate normal distribution $x \sim N(\mu, \Sigma)$, where μ was drawn from a standard uniform distribution across the open interval $\{0\ 1\}$ for each item and the off-diagonal elements of $i \times i$ covariance matrix Σ were set to a free parameter *c*. G was divided into 12 think (T) and 12 no-think (NT) items such that $G_{vi} = [T_{vi}, NT_{vi}]$. Note here that we used 12 items in each condition instead of 24 because of the cross-validation procedure, which assigned half of the items in the RDM_{-fus} to a training set and the other half to a test set. The parameter *c* determines the mean correlation across all patterns (i.e., items). Each column in this initial grid represents the initial pattern of activity triggered by a memory cue paired with a stored object. From this initial pattern, activity was then modulated differently for think and no-think items.

Memory suppression type: Inhibition versus truncated activation versus retrieval alone. For both think and no-think trials, a proportion (*x* and *y*, respectively; see voxel selection type) of voxels was first selected. Think trials were enhanced by an enhancement factor (*e*), such that $T_{xi} = T_{xi} \cdot e$ (e.g., doubled when $e = 2$).

- Inhibition: no-think selected voxels were down-scaled by a suppression factor (*s*) such that $NT_{yi} = NT_{yi} \cdot s$ (e.g., halved when $s = 0.5$).
- Truncated activation: the number of selected voxels whose reactivation was truncated in no-think trials corresponded to a ratio, *r*, of the number of selected voxels in the think condition, i.e., a proportion $x \cdot r$ of all voxels (e.g., 25% of voxels when $r = 0.5$ and $x = 0.5$). Activity of this subset of voxels during no-think trials was up-scaled by the same *e* as for think trials.
- Retrieval alone: the initial grid of activity was not modulated for no-think trials.

Voxel selection type: Targeted versus random versus fixed voxel selection.

- Targeted: for both think and no-think trials, a proportion (*x* and *y*, respectively) of voxels that were most highly activated were selected (e.g., the top 30% when $x = 0.3$). This selective mechanism was applied separately for each item.
- Random: for both think and no-think trials, a proportion (*x* and *y*, respectively) of voxels were randomly selected. This random selection was applied separately for each item.
- Fixed: the same proportions of voxels (*x* and *y*) were selected as in other models, in a random fashion (regardless of activity level), but this selection was fixed across items within a condition. Note that under this account, an additional overlapping factor (*o*) was introduced to control for the degree of overlap between voxels selected in the think and no-think conditions,

such that $o = 0.5$ means that half the voxels selected in the no-think condition were the same voxels as selected for the think condition.

Finally, for all models, once activity was modulated for think and no-think items, noise randomly drawn from a standard uniform distribution on the open interval $\{0\ 1^{*}n\}$, with $n < 1$, was added to each voxel and pattern such that $G_{vi} = G_{vi} + R_{vi}$.

The goal of the next MCMC step was then to sample the entire parameter space and to identify which parameter values were most likely to fit to the observed $RDM_{\text{-fus}}$.

MCMC Algorithm. Our goal was to sample from the unknown target (i.e., posterior) distribution $p(\theta_j)$ of each of the $j = 1...N$ parameters presented above. Here we used a Metropolis sampler, which creates a Markov chain that produces a sequence of values

$$\theta_j^{(1)} \rightarrow \theta_j^{(2)} \rightarrow \theta_j^{(3)} \rightarrow \ldots \rightarrow \theta_j^{(t)},$$

where $\theta_j^{(t)}$ represents the state of a Markov chain at iteration t. In the Metropolis procedure, we initialize the first state, $\theta_j^{(1)}$ to some initial random value. For each parameter, we then used a standard uniform (see below) proposal distribution $q(\theta_j)$ to generate new candidate θ_j^{*}. The use of a uniform distribution is convenient as it makes no assumption about the shape of the target distribution, and it satisfies a key requirement of the Metropolis sampler, which is to have a symmetrical proposal distribution such that $q(\theta_j^{*}|\ \theta_j^{(t-1)}) = q(\theta_j^{(t-1)}|\ \theta_j^{*})$. The next step is then to either accept or reject the new proposal θ_j^{*}, with the probability of accepting the new proposal being

$$\alpha = \min\left(1, \frac{p(\theta_1^{*}, \theta_2^{t-1}, \ldots, \theta_N^{t-1})}{p(\theta_1^{t-1}, \theta_2^{t-1}, \ldots, \theta_N^{t-1})}\right),$$

To compute this acceptance probability, we calculated for a given model M (see *General Overview*), $RDM_{\text{-t}}$ with the new proposal, such that new $RDM_{\text{-t}} = M\ (\theta_1^{*}, \theta_2^{(t-1)}, \ldots, \theta_N^{(t-1)})$, as well as $RDM_{\text{-t}}$ at the state of the chain $t - 1$, such that old $RDM_{\text{-t}} = M\ (\theta_1^{(t-1)}, \theta_2^{(t-1)}, \ldots, \theta_N^{(t-1)})$. Then we computed the cost of both the new proposal and old state such that new cost $= 1 - r$(new $RDM_{\text{-t}}$, $RDM_{\text{-fus}}$) and old cost $= 1 - r$ (old $RDM_{\text{-t}}$, $RDM_{\text{-fus}}$), with r being the Spearman rank correlation between the two vectorized RDMs. The probability of accepting the new proposal becomes then:

$$\alpha = \min(1, \exp(-(\text{new cost}/\text{old cost}))).$$

Hence, when a new cost value decreases relative to the old cost after a new proposed parameter (i.e., better fit), α increases toward 1 [i.e., new parameter θ_j^{*} is more likely than the old one $\theta_j^{(t-1)}$]. To make a decision on whether to accept or reject the new proposal, we draw a value, u, from a uniform standard distribution on the open interval $\{0\ 1\}$. If $u < \alpha$ or if the new cost value decreases relative to old cost, we accept the proposal θ_j^{*} and the next state is then set to $\theta_j^{(t)} = \theta_j^{*}$. If $u > \alpha$, we reject the new proposal and the next state is set to be equal to the old state, $\theta_j^{(t)} = \theta_j^{(t-1)}$.

At each iteration t, we generate independently a new proposal for each parameter entering our model M and either accept or reject the proposal. Here is a summary of the steps of the Metropolis sampler:

i) Set $t = 1$.
ii) Generate an initial value drawn from a uniform proposal distribution (see below) for each parameter $\theta_1, \theta_2, \theta_3, \ldots, \theta_N$.
iii) Generate a proposal θ_1^{*}, from $q(\theta_1)$ which is the uniform proposal distribution of θ_1, with $\theta_{1min} < \theta_1 < \theta_{1\ max}$.
iv) Evaluate the acceptance probability $\alpha = \min(1, p(\theta_1^{*}, \theta_2^{t-1}, \ldots, \theta_N^{t-1})/p(\theta_1^{t-1}, \theta_2^{t-1}, \ldots, \theta_N^{t-1}))$, with $\alpha = \min(1, \exp(-(\text{new cost}/\text{old cost})))$.
v) Generate u from a uniform $\{0\ 1\}$. If $u < \alpha$ or new cost < old cost, accept the proposal and set $\theta_1^{(t)} = \theta_1^{*}$; else set $\theta_1^{(t)} = \theta_1^{(t-1)}$. Apply the same process for $\theta_2, \ldots, \theta_N$.
vi) Repeat until $t = T$.

When t reaches the number of iterations specified (here $T = 5{,}000$), we then have an approximation of the posterior distribution of each parameter θ. Because this Metropolis algorithm always accepts a new proposal when it is more likely under the posterior distribution than the old state, the sampler will move toward the regions of the state space where the posterior distribution has high density (in other words, toward parameter values which are more likely to explain the data, i.e., $RDM_{\text{-fus}}$; Fig. S5). However, even if the new proposal provides a worse fit to the data than the current state, it might still be accepted because $u < \alpha$ could arise by chance (if the drawn value is very low). This process of always accepting a new parameter value that improves model fit but occasionally accepting other values to ensure that the sampler explores the whole state space, i.e., samples all parts of the posterior distribution (including the tails).

However, this parameter space is limited by the open interval chosen for the uniform proposal distribution, so it is important that these proposal distributions cover the entire space of possible values, bound by some limits. Here we used the following uniform discrete distributions for the parameters described in the above model construction section:

- Average correlation across all patterns, $c = U(0.1,0.9)$, step = 0.05.
- Number of voxels composing the grid G, $v = U(20,1000)$, step = 20.
- Proportion of noise added to the data, $n = U(0.05, 0.9)$, step = 0.05.
- Suppression factor, $s = U(0.05, 0.9)$, step = 0.05 (inhibition accounts only).
- Retrieval factor, $e = U(1/\max(s), 1/\min(s))$, i.e., $e = U(1.1, 20)$, step = 0.05.
- Proportion of modulated voxels for think items, $x = U(0.05, 0.9)$, step = 0.05.
- Proportion of modulated voxels for no-think items, $y = U(0.05, 0.9)$, step = 0.05.
- Ratio of modulated voxels for no-think items compared with think items, $r = U(0.05, 0.95)$, step = 0.05 (truncated activation accounts only).
- Proportion of overlapping voxels between think and no-think condition, $o = U(0, 1)$, step = 0.05 (fixed voxel selection accounts only).

Note that for the retrieval alone model, s and y were not relevant and hence not sampled by the Metropolis algorithm.

Random-Effect Analysis and Cross-Validation. The MCMC algorithm presented above allows us to sample from the posterior distribution of each parameter and to identify the regions of the state space where the posterior distribution has high density for the RDM of a given participant. Once the initial samples of the MCMC algorithm have been discarded (the burn-in period was set to 250 samples; see *MCMC Convergence*), the mode of this posterior distribution hence reflects a reasonable estimate of the most likely parameter values under the posterior distribution (MAP estimation), i.e., providing the best fit to the data. However, with so many parameters to each model, and relatively few data, there is a danger that the models will overfit the data (i.e., fit the noise in the data, rather than the true signal). To evaluate this, we used cross-validation to select the model that best generalizes from one half of the data (training set) to the other half

of the data (test set). We randomly split the RDM_{-fus} of each participant into two independent halves 100 times, each time fitting the training half using the above MCMC algorithm, and using the posterior mode of each parameter to estimate the GOF for the test half. GOF was defined as $r(RDM_{-t}, RDM_{-fus})$, where r is the Spearman correlation between the two vectorized RDMs. These 100 GOF values for each participant and the models were then averaged, resulting in a 24-participant × 9-model data set. Statistical differences between models were then tested with a bootstrap with replacement approach on the mean difference between pairs or families of models (using 2,000 bootstraps), allowing us to compute the confidence intervals for the differences between models (corresponding to bias-corrected and accelerated percentile method).

Fig. 5*C* in the main text reports the mean GOF across participants. The cross-validation approach used for model fitting and testing is illustrated in Fig. S4.

MCMC Convergence. The first 250 samples of the MCMC chains were discarded and not collected. Different diagnostic tests were performed to check whether the chains have converged to their stationary distributions. Those tests were performed on each sampled parameter for each model and each participant, discarding the first 250 initial samples. One way to assess convergence is to compute the autocorrelations between the draws of the Markov chain. The lag k autocorrelation ρ_k is the correlation between every draw i of the chain x and its k_{th} lag:

$$\rho_k = \frac{\sum_{i=1}^{n-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Fig. S6 illustrates how k_{th} lag autocorrelation is smaller as k increases for a given participant and random split, indicating that the chains have mixed quickly to their stationary distribution. This pattern was true across all participants and random splits.

Another assessment of stationary distribution is the Gelman–Rubin diagnostic which can be performed by running the same Markov chain multiple times (as was done for the cross-validation approach above) and to estimate the variance of the parameter as a weighted sum of the within-chain and between-chain variance. The within variance (W) is the mean of the variance of m chains, such that

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2,$$

where s_j^2 is the variance of the j_{th} chain x with n samples, such that

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2.$$

The between variance (B) is given by

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{x}_j - \bar{\bar{x}})^2, \text{ where } \bar{\bar{x}} = \frac{1}{m}\sum_{j=1}^{m}\bar{x}_j.$$

We can then estimate the variance of the stationary distribution as a weighted average of W and B:

$$\widehat{Var}(x) = \left(1 - \frac{1}{n}\right)W + \frac{1}{n}B.$$

The estimated potential scale reduction factor (EPSR) corresponds then to

$$\hat{R} = \sqrt{\frac{\widehat{Var}(x)}{W}}.$$

EPSR measures the degree to which the posterior variance would decrease if we were to continue sampling to infinity. If EPSR ≈ 1, then that estimate is reliable, meaning the variance between the chains is similar to the variance within each chain, and that the chains have converged to the stationary distribution.

Here EPSR < 1.06 for all parameters of each model tested for each participant and random split, indicating that the MCMC algorithm converged well.

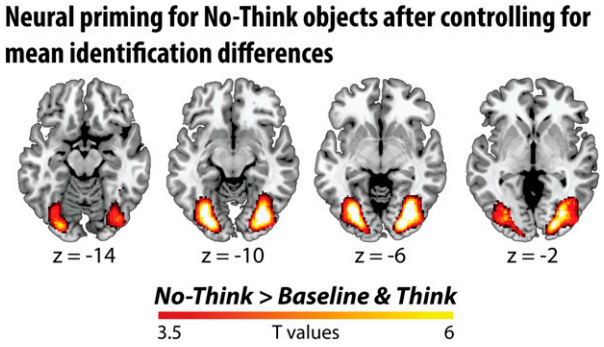

Fig. S1. Memory inhibition effect during the final priming test phase after controlling for mean identification time differences across conditions. Note that, contrary to Fig. 3*B*, we did not mask this effect with the main effect of neural priming.

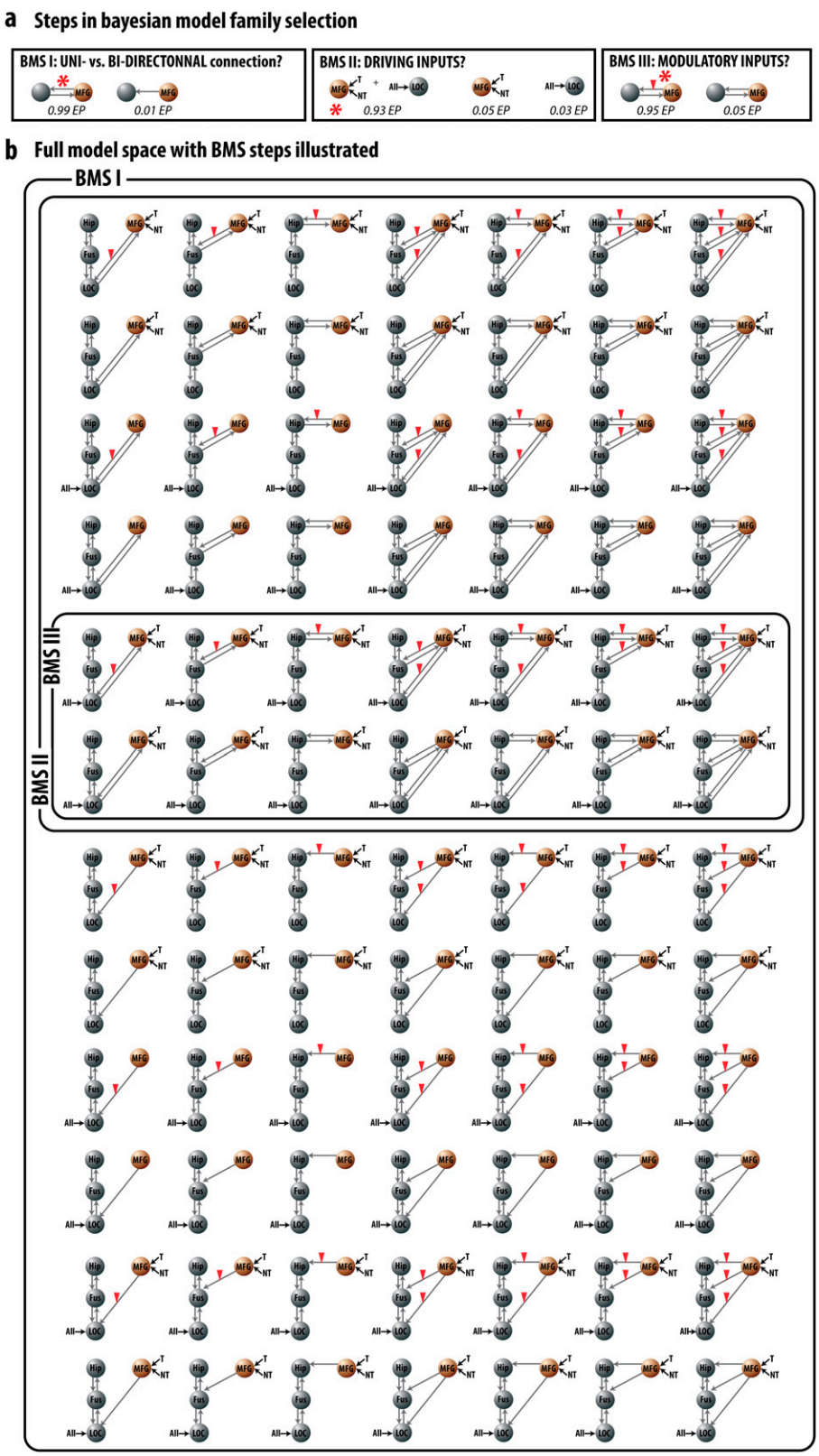

Fig. S2. Dynamic causal modeling (DCM) model space and Bayesian model selection (BMS) procedures. (*A*) BMS was first applied on the direction (bilateral versus unilateral intrinsic connections) family. The bilateral subgroup won (as indicated by red asterisk) against the unilateral subgroup with an exceedance probability of 0.99, and an expected posterior probability of 0.77. Within the bilateral family of models, we then compared which driving input was more likely. Models including a driving input in both the lateral occipital complex (LOC) and the middle frontal gyrus (MFG) won with an exceedance probability of 0.926 (against 0.0532 for the MFG only and 0.0208 for the LOC only), and an expected posterior probability of 0.5538 (against 0.2476 for the MFG only and 0.1986 for the LOC only). Finally, we compared the remaining seven modulatory models (i.e., including a top-down modulation of the coupling between the

Legend continued on following page

MFG and posterior regions during no-think trials) to the seven null models that did not include any modulatory input on top-down connections. This analysis overwhelmingly favored models with modulation over models without modulation (exceedance probability = 0.95, expected posterior probability = 0.73). (*B*) The full space of DCM models. Gray arrows represent connections between nodes. Black arrows correspond to the driving inputs of the models. The red arrowheads illustrate the presence of a modulatory input for no-think items on connection strength.

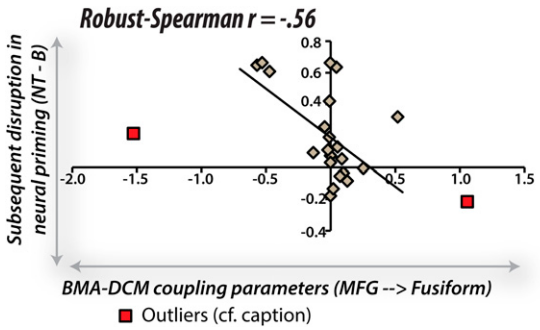

Fig. S3. Relationship between modulatory parameters for the MFG→fusiform DCM model and disrupted neural priming for no-think items (i.e., no-think – baseline). Robust Spearman correlation (1) revealed the presence of two bivariate outliers consistently identified using box-plot rule, the median absolute deviation to the median, and the median of absolute distances (S outlier; cf. ref. 1).

1. Pernet CR, Wilcox R, Rousselet GA (2013) Robust correlation analyses: False positive and power validation using a new open source Matlab toolbox. *Front Psychol* 3:a606.

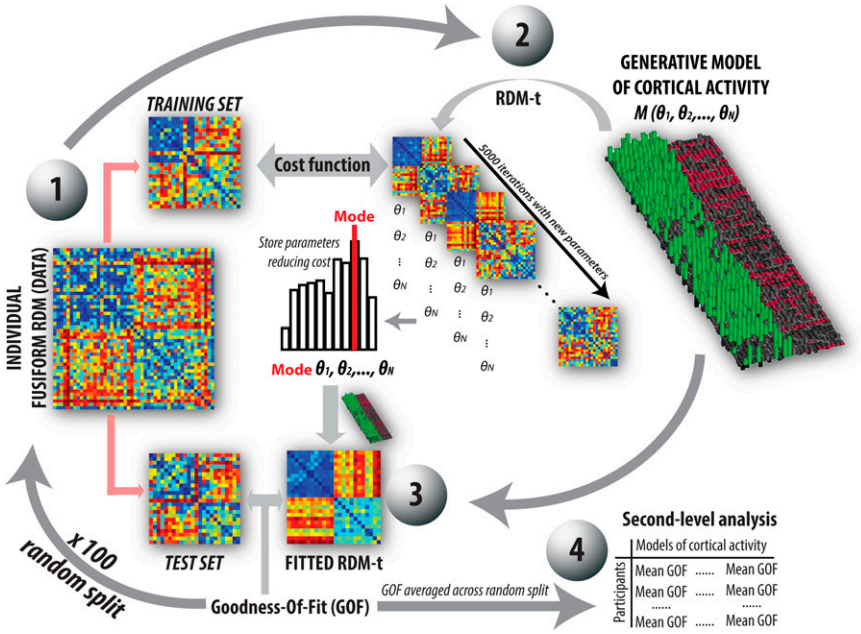

Fig. S4. A visual illustration depicting the model fitting and testing procedures based on MCMC: a cross-validation approach. We randomly split the RDM_{-fus} of each participant into two independent halves 100 times, each time fitting the training half using the MCMC algorithm. The MCMC algorithm allows us to sample from the posterior distribution of each parameter and to identify the regions of the state space where the posterior distribution has high density. Once the initial samples of the MCMC algorithm have been discarded (the burn-in period was set to 250 samples, see *MCMC Convergence*), the mode of this posterior distribution hence reflects a reasonable estimate of the parameter values providing the best fit to the data. We used the posterior mode of each parameter to test how model fitting generalizes to the other half of the data (test set) and to estimate the GOF (i.e., Spearman correlation between the two vectorized RDM_{-fus} and RDM_{-t}). The 100 GOF values computed were then averaged for each model and each participant, and entered into a second-level analysis treating participants as a random effect variable.

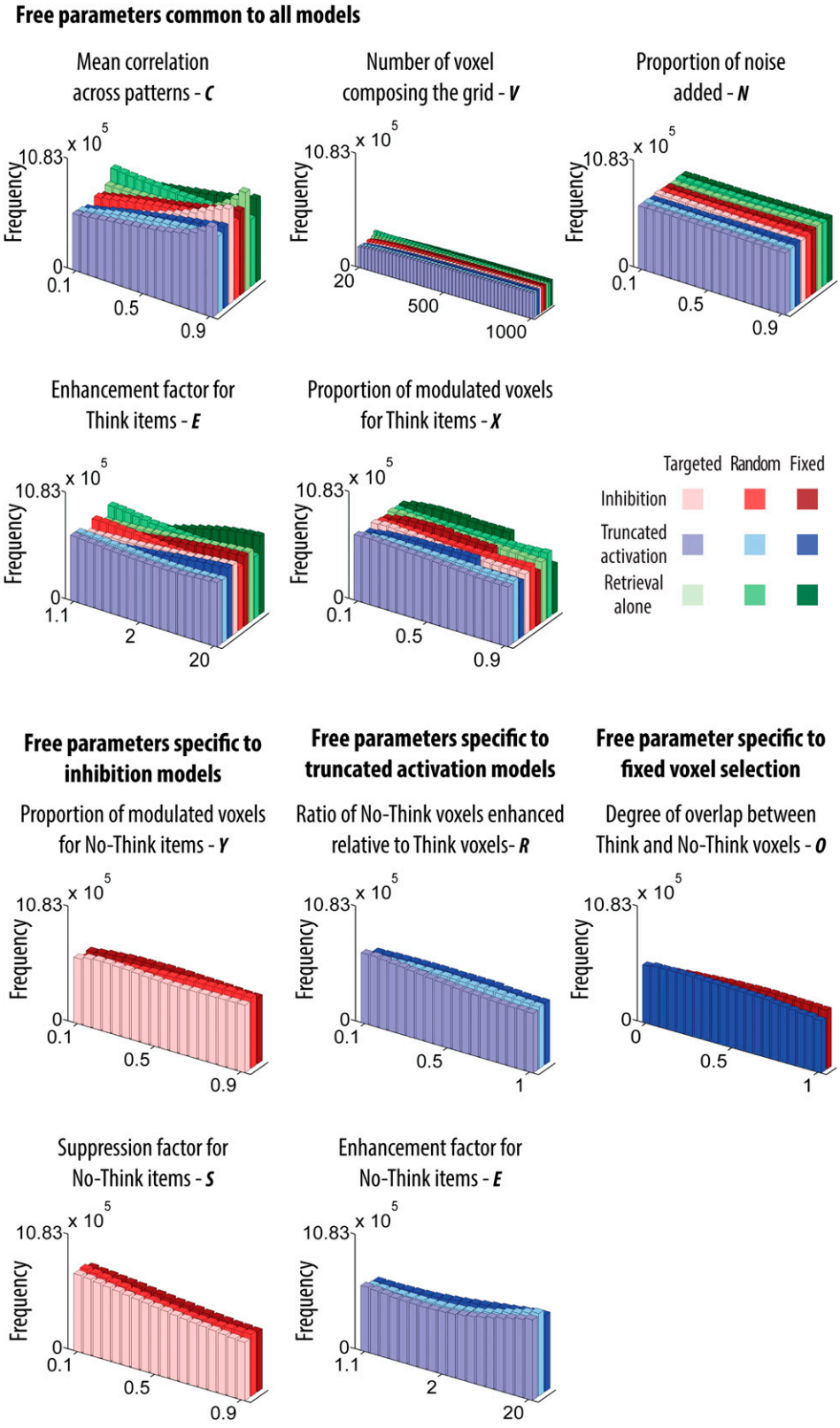

Fig. S5. Histogram of the sample distribution for each model free parameter obtained after MCMC convergence. The MCMC Metropolis algorithm ensures that the whole state space of parameters is sampled, and the sampler will move toward the regions of the state space that provide a better fit to the data. Hence, more frequent values represent a critical feature that is necessary to explain the data under a given model, while distributions that stay largely uniform indicate that these parameters do not have much impact on model fit. For instance, the distribution of the suppression factor (*S*) is skewed toward lower values, indicating that suppression has an impact on model fit. In contrast, the distribution of the number of voxels remains flat showing that our findings generalize well across a range of voxel number values. These histograms were plotted after 100 repetitions of the MCMC Metropolis sampler comprising 4,750 iterations (i.e., discarding the first 250 burn-in samples) across 24 participants.

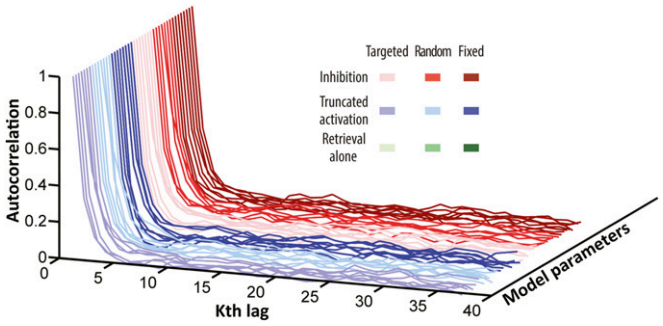

Fig. S6. Autocorrelation between the draws of the Markov chains showing that the MCMC algorithm converged well. Autocorrelation is a cross-correlation of the sample time series with itself as a function of a time separation (i.e., k_{th} lag). A decrease in autocorrelation when k lag increases indicates a fast mixing of the chain and a convergence to a stationary distribution. Here, these autocorrelations were plotted for one participant and for each model parameter after a single random split of the MCMC Metropolis sampler comprising 4,750 iterations (i.e., discarding the first 250 burn-in samples), to illustrate that the MCMC algorithm converged well. Note that the Gelman–Rubin diagnostic test also indicated a convergence of the Markov chains to stationary distribution (*SI Simulation Methods*).

Table S1. Peak coordinates of the regions showing a think versus no-think difference at $P_{FWE} < 0.05$ (whole brain)

Anatomical description	No. of voxels	MNI coordinates, mm			T	P_{FWE}
		x	y	z		
No-think > think						
Right SFG	377	20	16	58	8.96	<0.001
Right IPC	705	44	−46	36	8.5	<0.001
Right MFG	331	44	24	46	8.37	<0.001
Right IFG	254	50	20	8	8.26	<0.01
Left IPC	68	−60	−52	38	6.8	<0.01
Right inferior orbitofrontal gyrus	25	42	46	−8	6.58	<0.05
Right SFG (anterior)	67	22	52	18	6.29	<0.05
Right medial SFG	45	10	36	42	6.23	<0.05
Left LOC	7	−46	−80	−4	6.22	<0.05
Left inferior temporal gyrus	7	−58	−28	−20	6.24	<0.05
Left MFG	5	−40	28	44	6.01	<0.05
Right superior parietal gyrus	7	34	−52	58	5.98	<0.05
Think > no-think						
Left fusiform gyrus	52	−32	−32	−24	7.09	<0.01
Left IFG	49	−42	32	14	7.09	<0.01

The think > no-think difference observed in the hippocampus survived correction when the search volume was restricted to the left [$t(23) = 4.01$, $P_{FWE} < 0.05$, $x = -32$, $y = -26$, $z = -14$] and to the right [$t(23) = 3.65$, $P_{FWE} < 0.05$, $x = 34$, $y = -8$, $z = -26$] hippocampus. IFG, inferior frontal gyrus; IPC, inferior parietal cortex; MNI, Montreal Neurological Institute; P_{FWE}, P family-wise error; SFG, superior frontal gyrus.

Table S2. Peak coordinates of the regions showing neural priming (think + no-think + baseline < unprimed) and memory inhibition (no-think > think + baseline) during the final priming test phase at $P_{FWE} < 0.05$

Anatomical description	No. of voxels	MNI coordinates, mm			T	P_{FWE}
		x	y	z		
Think + no-think + baseline < unprimed (whole-brain correction)						
Right inferior temporal and fusiform gyri	442	46	−54	12	7.3	<0.001
Left LOC	679	−44	−68	−6	7.24	<0.001
Left fusiform gyrus		−40	−54	−10	6.13	<0.01
Left fusiform gyrus		−34	−44	−18	5.74	<0.01
Right inferior temporal gyrus	183	48	8	26	6.36	<0.01
No-think > think + baseline (main effect of neural priming as restricted search volume)						
Left LOC	243	−38	−76	−8	5.28	<0.001
Left fusiform gyrus		−36	−60	−6	4.96	<0.01
Right fusiform gyrus	69	40	−58	−10	4.65	<0.01
Left fusiform gyrus	3	−34	−52	−16	3.42	0.059

An additional whole-brain correction showed a memory inhibition effect (no-think > think + baseline) in the right LOC [$t(23) = 10.88$, $P_{FWE} < 0.05$, $x = 26$, $y = -84$, $z = -6$], although this region did not show an initial neural priming effect (at least when P values were whole-brain corrected). Fig. S1 also reports whole-brain memory inhibition effect (no-think > think + baseline) during the final priming test phase after controlling for mean identification differences across conditions. Regions in italics correspond to submaxima peak coordinates in the cluster.