Full title: The neural mechanisms of active removal from working memory

Short title: Active removal from working memory

Authors:

Jiangang Shan[1] and Bradley R. Postle[1,2]

Affiliations:

[1]Department of Psychology, University of Wisconsin–Madison, Madison, WI 53706, USA

[2]Department of Psychiatry, University of Wisconsin–Madison, Madison, WI 53719, USA

***Correspondence should be addressed to:**

Bradley R. Postle

Department of Psychology

University of Wisconsin-Madison

Madison, WI 53706, USA

Email: postle@wisc.edu

Abstract

The ability to frequently update the contents working memory (WM) is vital for the flexible control of behavior. Whether there even exists a mechanism for the active removal of information from working memory, however, remains uncertain. In this Preregistered Research Article we will test the predictions of models for three different mechanisms of active removal: hijacked adaptation, context breaking, and mental-context shifting. We will collect functional magnetic resonance imaging (fMRI) data while subjects perform a novel “ABC-retrocuing” task designed to elicit two modes of removal, active or passive. The hijacked-adaptation model posits an adaptation-like modification of perceptual circuits combined with a weak activation of the to-be-removed item. Its predictions will be assessed by using multivariate inverted encoding modeling (IEM) and photic “pings” to assay the state of feature-selective encoding channels and of putative activity-silent representations under active-removal versus passive-removal conditions. The context-breaking model posits a breaking of the stimulus-to-context association posited to be the basis of holding information ‘in’ working memory, and it predicts different patterns of representational dynamics, including different responses to the ping. Finally, the mental-context shifting model posits that interference from no-longer-relevant information is minimized by making the mental context associated with new information dissimilar from that associated with the to-be-“removed” information. This will be tested by using representational similarity analysis (RSA) to compare the rate of contextual shift under active-removal versus passive-removal conditions.

Introduction

A hallmark of working memory (WM) is that it is rapidly updateable, such that information that was relevant in the recent past can be easily replaced once circumstances change and different information has become of primary importance. One way that this is operationalized in the laboratory is with a block of stand-alone trials (e.g., of delayed recognition): Once trial *n* has been completed, subjects have little difficulty encoding a new memory set for trial *n + 1*. Because the set of items is randomly selected for each trial, the memory items for each trial lose their relevance at the end of that trial, and the common intuition is that they should be removed from WM. Despite this intuition, however, the phenomenon of proactive interference indicates that the assumed removal of no-longer-relevant information is often not complete. This is particularly notable for trials featuring "recent-negative" recognition probes that were not in the memory set of the current trial, but were in the memory set on the previous trial -- these lead to an increased false-alarm rate, and to longer reaction times (RTs), for correct rejections [1]. For visual WM tasks that test recall, the imperfect nature of removal manifests itself as serial dependence. For example, when the orientation of a Gabor patch is the feature to memorize and then recall, the reported orientation for trial *n* is commonly found to be biased toward the orientation of the item that had been shown on trial *n – 1* (e.g., [2, 3]). (This effect is commonly referred to as an "attractive bias," because it's as though the response on trial *n* is attracted toward the orientation from *n – 1*.)

There is also a growing body of neural evidence for the incomplete removal of information from WM. In an electroencephalography (EEG) study of delayed recall of orientation, Bae and Luck [4] were able to decode the orientation of the previous trial's sample after the onset of the current trial's sample. For delayed recall of location, Barbosa, Stein et al. [5] were able to decode the previous trial's sample location (from activity in the prefrontal cortex (PFC) of nonhuman primates) from late in the intertrial interval (ITI), just prior to the start of the next trial. Additionally, they observed a similar pattern of reactivation in whole-scalp EEG in humans. Simulations with a bump-attractor network model suggested that the reactivation of no-longer-relevant information may be due to "nonspecific" activation

of a residual neural trace that is “imprinted in neuronal synapses as a latent activity-silent trace” [5]. Tellingly, this model did not include an explicit mechanism for removal of no-longer-relevant information; rather, when an item was no longer relevant, activation was simply withdrawn from it, and the bump of activity representing it receded to baseline. Similarly, in a study using a different formal model of WM performance, the Prefrontal Basal Ganglia Working Memory (PBWM) model, the replacement of a no-longer-relevant item with a new one was accomplished via the “reallocation” of resources away from the former [6]. Thus, many frameworks assume that a default strategy for updating the contents of WM is to employ what we will refer to as the “passive removal” of no-longer-relevant information.

In addition to passive removal, there is also considerable evidence for an active removal mechanism, particularly during tasks that require the simultaneous maintenance of multiple items in WM. One example comes from dual serial retrocuing (DSR) tasks, in which subjects are first shown two stimuli to memorize, then a retrocue indicates which will be tested first; after the first test, a second retrocue is shown to indicate (with equal probability) which of the two items will be tested in a second test. The first retrocue designates one of the two as a “prioritized memory item” (PMI), and the uncued item, by default, becomes an “unprioritized memory item” (UMI). Critically, the UMI can’t be removed from WM, because it may be needed for the second test. After the second retrocue, the newly cued item takes on the status of PMI, and the uncued becomes irrelevant for the remainder of the task (i.e., an “irrelevant memory item,” IMI). Thus, the DSR task creates three operationally different states for a memory representation: PMI; UMI; and IMI. In functional magnetic resonance imaging (fMRI) and EEG studies, the ability to decode the identity of a PMI during the delay period of a WM task is a hallmark of its active state. In contrast to this, in some studies using the DSR procedure, multivariate evidence for an active trace of the UMI can drop to baseline (e.g., [7, 8]). Despite this, the fact that the UMI has not been removed from WM is inferred from that fact that a pulse of transcranial magnetic stimulation (TMS) has two effects: physiologically, it produces a transient reactivation of the active trace of the UMI [9]; behaviorally, it produces an increase in false alarm responding when the UMI is used as an invalid memory probe (reminiscent of proactive

interference effects; [9, 10]). Turning at last to the IMI, evidence for its active removal is inferred from the absence of evidence for either TMS reactivation [9] or a TMS-related false alarm effect [9, 10]. It is important to emphasize that the labels "PMI," "UMI," and "IMI" refer to an item's state of operational relevance for the cognitive system, not to its presumed physiological state. Consider, for example, the fate of an item at the end of a trial. Although it is an IMI, the fact that its identity can be decoded from the response of a pulse of TMS delivered late in the ITI is interpreted as evidence for a residual activity-silent trace of that item [5]. In the DSR, in contrast, the absence of such a TMS-reactivation effect for the IMI is taken as evidence that an active removal mechanism has removed any activity-silent trace of the IMI [9, 10].

Three models of active removal of information from WM

In this Preregistered Research Article we propose to test two current models of active removal -- "context breaking" and "context shifting" – plus a novel model that we are introducing with this study – "hijacked adaptation." Although these three hypothesized mechanisms aren't mutually exclusive, and so could in principle co-exist in the cognitive system, such a state of affairs would seem to be redundant, because each of the three has been proposed to accomplish the same function. (It is also important to note that each of these models is grounded in a different theoretical framework, and we will highlight places where important assumptions are not shared.) This Preregistered Research Article is designed to detect positive evidence for each of these three hypothesized mechanisms, and although it's our expectation that we will find evidence for only one of them, the design allows for the detection of evidence for multiple mechanisms, should it exist. Because context breaking and context shifting have been previously described, we will review them only briefly before introducing hijacked adaptation.

Context breaking. A fundamental tenet of the interference model of visual WM is that the binding between an item's representation (i.e., it's "*content*") and the representation of the task-specific *context* in which that item is

encountered is fundamental to that item being held "in WM" [11]. Thus, in this framework active removal is accomplished by breaking the association between an item's *content* (e.g., the orientation of a Gabor patch) and its *context* (e.g., where this patch had appeared on the screen, or the ordinal position within the series in which it had appeared; [12]). From here forward, we will refer to this as the "context-breaking" model (Figure 1a). Although this model has been described in most detail in theoretical and computational terms, in this fMRI study we infer predictions that it would make at the level of neural systems.

Figure 1. Illustrations of three models of active removal of information from WM. Each ring represents a bank of orientation-tuned perceptual channels, with the level of activity of each channel (small circles that make up the ring) represented by its color. (Thus, a reddish circle flanked by two yellowish ones corresponds to a "bump" of activity centered on the actively represented orientation. The quadrilateral surfaces below each orientation ring in panels *a* and *b* correspond to a spatial priority map, and the line between the ring and the priority map to the binding between a memory item's orientation and its location context. (a) Context breaking. This panel illustrates the same item at two points in time: Breaking the content-to-context binding (scissors) removes the item from WM, and consequently its activity returns to baseline levels. (b) Context shifting. These panels illustrate two items that are processed sequentially in the ABC-retrocuing task, a previously presented item that is no longer relevant, and so has become an IMI, and a "new item" that takes on the status of PMI. The top panel illustrates a "*no-overlap*" trial (see Fig. 2), in which the different location of the new item relative to the IMI results in minimal interference between the two, and mental context drifts at its default steady rate (illustrated by the smooth transition of color saturation). The bottom panel illustrates an "*overlap*" trial (see Fig. 2), in which the new item's appearance at the same location as the IMI would elevate the level of interference between the two. Cognitive control compensates for this overlap in location context by abruptly shifting mental context during the interstimulus interval, such that the mental context associated with the new item will be markedly different from the mental context associated with the IMI. (c) Hijacked adaptation. This panel illustrates the same item at two points in time. The gray scale-colored ring below each orientation ring represents the gain of each corresponding orientation channel. Left side of image represents the moment at which the to-be-removed item receives an intermediate level of activation (top-down signal, not shown) which, because of the orientation-specific pattern of gain modulation, produces an inverted pattern of activity relative to when the item was a PMI (i.e., lower activity in the center channel relative to flanking channels). Right side of image corresponds to a few seconds later during the trial -- although level of activity has returned to baseline, the pattern of reduced gain persists, which will generate a repulsive serial bias in the next trial (see Figure 3).

Context shifting. The recently articulated "Working Memory Episodic Memory" model takes the position that many functions and properties that have traditionally been associated with working memory may be "nothing more" (our quotes) than mechanisms that also contribute to episodic long-term memory (LTM). In particular, Beukers et al [13] challenge the

utility of positing an activity-silent state of WM (as codified by Stokes [14] and assumed by, e.g., Rose et al. [9] and Barbosa, Stein, et al. [5]). It's much more parsimonious, they argue, to simply construe instances of 'representation without activity' (e.g., [7, 8]) as evidence that an LTM representation is created every time a sample is presented in a WM task. From this perspective, transforming a UMI back into a PMI is an instance of retrieval from LTM. Of principal relevance for this Preregistered Research Article is the question of how this Episodic Memory Working Memory model accounts for active removal from WM. Beukers et al. [13] propose that "flexible forgetting" from WM is accomplished by shifting the mental context with which more recent information is encoded, thereby making the retrieval of the IMI less likely (from here forward, we will refer to this as the "context-shifting" model, Figure 1b).

Hijacked adaptation. This is a hypothesized top-down mechanism that works by combining an adaptation-like modification of perceptual circuits with a weak activation of the to-be-removed information. Its core function is two-fold: remove the active trace of the IMI and erase the activity-silent trace of the IMI (Figure 1c and Figure 4). Because this is the first a priori test of this idea (which draws heavily on [15]), the remainder of this introduction will be devoted to the empirical and theoretical contexts that motivate it.

Results from "ABC-retrocuing" provide behavioral evidence for an active removal mechanism

In a recent behavioral study, Shan and Postle [16] designed a novel "ABC-retrocuing" task intended to engage active or passive removal of an IMI from WM. Each trial began with the simultaneous presentation of two sample oriented gratings (items *"A"* and *"B"*) in two of six possible locations. After a brief delay, a circle appearing at one of the two locations indicated that the corresponding item (for this example we'll say *A*) might be tested at the end of the trial, thereby designating *A* a PMI and *B* an IMI. After another brief delay, a third item ("*C*") was presented, and at the end of the trial recall of the orientation of either *A* or *C* was tested with a response dial appearing at the location of the to-be-recalled item. The critical manipulation that was intended to encourage active versus passive removal was the location at which

item *C* would be presented: In the *overlap* condition, item *C*'s location was always the same as that of the IMI (i.e., item *B*); and in the *no-overlap* condition item *C* always appeared at one of the locations that had not been occupied by either item *A* or *B*. Trials were blocked by condition, and subjects were explicitly informed about the condition prior to each block. The logic was that the *no-overlap* condition might encourage passive removal, just because this seems to be the default for many working memory tasks, as evidenced by the proactive interference and serial dependence effects reviewed above. For the *overlap* condition, however, subjects might be motivated to actively remove the IMI from WM, because otherwise its shared location with item *C* could lead to retrieval conflict when the response dial appeared at this shared location (i.e., "cue conflict"; [11]). A final element of the procedure is that each ABC-retrocuing trial was followed by a trial of simple 1-item delayed recall of orientation, with serial dependence of 1-item recall on the immediately preceding ABC-retrocuing trial used to index the fate of the IMI. (The elements of the ABC-retrocuing task are illustrated in Figure 2, although the study by Shan and Postle [16] differed from Figure 2 in two respects: Shan and Postle [16] did not include the "ping" illustrated in Figure 2; and each trial of ABC-retrocuing in [16] was followed by a trial of 1-item delayed recall.)

Figure 2. The ABC-retrocuing task, as designed for this Preregistered Research Article. The top row illustrates a trial in the *no-overlap* condition, the bottom row a trial in the *overlap* condition. See text for details. Note that the design for the behavioral study from Shan and Postle [16] was similar, with the exceptions that delay periods were shorter, there was no ping, and each trial of ABC-retrocuing was followed by a trial of 1-item delayed recall. The digits below each panel of the overlap trial correspond to elapsed seconds, and each TR in the timeline at the bottom of the figure corresponds to 2 sec.

Preliminary results from Shan and Postle [16] revealed a striking difference between the *overlap* and *no-overlap* conditions. In the *no-overlap* condition, item *B* had an attractive serial bias on 1-item recall, consistent with the attractive serial bias observed in several previous studies that we assume were characterized by passive removal of no-longer-relevant stimulus information (e.g., [2, 3, 17]). In the *overlap* condition, in contrast, item *B* had a repulsive serial bias on 1-item recall. That is, whereas in the *no-overlap* condition the responses on 1-item recall trials were biased toward the orientation of the IMI from the preceding trial of ABC-retrocuing, in the *overlap* condition the responses on 1-item recall trials were biased away from (hence, "repulsed by") the orientation of the IMI from the preceding trial of ABC-retrocuing

(Figure 3). The fact that the serial bias from the IMI was flipped depending on condition suggested that the IMI was processed in a very different way during *overlap* versus *no-overlap* trials. The interpretation that the critical difference between the two conditions was active versus passive removal of the IMI was reinforced by the fact that the serial bias exerted by item *A* was attractive in both conditions.

Figure 3. Preliminary data from Shan and Postle [16]. In the *no-overlap* condition the IMI had an attractive serial bias on recall on the subsequent 1-item delayed-recall trial, as estimated by a derivative of gaussian fit (peak-to-peak distance = 2.148°, $p = 0.049$). In the *overlap* condition, in contrast, the IMI had a repulsive serial bias on the subsequent 1-item recall trial (peak-to-peak distance = -2.516°, $p = 0.011$), and these two effects differed significantly from each other ($p = 0.01$). The interpretation that the critical difference between the two conditions was active versus passive removal of the IMI was reinforced by the fact that in both conditions the retrocued item (i.e., item "*A*" Figure 2) exerted comparable levels attractive serial bias on 1-item recall on the subsequent trial (*no-overlap* peak-to-peak distance = 1.865°, $p = 0.061$; *overlap* peak-to-peak distance = 2.463°, $p = 0.019$).

The logic underlying the hijacked-adaptation hypothesis: Common mechanisms may underlie repulsive serial dependence and active removal from WM

First, it is helpful to review some characteristics of serial dependence. Recent work from two independent groups has converged on the view that, in vision, serial bias effects arise from two different levels of processing, each producing an influence opposite in sign to the other (i.e., attractive versus repulsive). At the level of perception, adaptation to recent perceptual events produces repulsion from previous stimuli, whereas at the level of decision making, perceptual decisions are attracted toward previous decisions [17, 18]. These opposing effects also differ with regard to strength of influence on behavior, and to time course. Decisional biases have a stronger influence on behavior, which explains why the serial bias that is most often reported in the literature is attractive. The influence of perceptual adaptation, however, is longer lasting. This accounts for the fact that whereas the dependency on an item from one or two trials previous is typically attractive, this effect flips for longer lags, such that, for example, the influence of the item from five trials previous is repulsive [17]. Of critical relevance for the hijacked-adaptation model, one condition in the ABC-retrocuing results from Shan and Postle [16] was at odds with this pattern: For the IMI in the *overlap* condition, the serial bias from the previous trial was repulsive.

This raises a possibility that is at the heart of the hijacked-adaptation model: active removal of an item in WM may be accomplished via a top-down mechanism that mimics the circuit-level adjustments that are the basis of perceptual adaptation (e.g., [19-22]), but in a manner that is faster (effectively instantaneous) and more pronounced (such that its influence on the subsequent trial overcomes the attractive influence of the decision-making stage). It is important to note that we assume that the effects of perceptual adaptation can be modeled as reductions of gain in the perceptual circuits where this adaptation is taking place. If we start with the assumption that the perception of orientation is accomplished by passing visual signals through a bank of orientation-tuned filters, perceptual adaptation to, say, a 90° grating can be modeled as a decrease of the gain setting of the 90° filter and a smaller decrease of gain at adjacent filters (e.g., those centered on 60° and 120°). (In the framework of multivariate inverted encoding modeling (IEM), which plays a prominent role in this Preregistered Research Article, this putative bank of orientation-tuned filters is operationalized with a basis set of orientation-tuned "perceptual channels.") Next we consider evidence from a WM task [15] that is consistent with this idea.

Lorenc, Vandenbroucke et al. [15] carried out an fMRI study of DSR of oriented-grating stimuli, and one finding was that multivariate decoding evidence for an active trace of the IMI dropped to a level significantly below baseline. When the same data were analyzed with a multivariate inverted encoding model (IEM) the IMI reconstruction of the IMI was "flipped" relative to its reconstruction as a PMI. (For other examples of priority-related "flipping" of IEM reconstructions, see [23-25]). To better understand this effect, the authors carried out computational simulations comparing the effects of modifying the gain, the width, or the spacing (i.e. shifts in tuning profiles) of orientation-tuned perceptual channels, combined with varying "memory strength," a factor that can be understood as the top-down attentional signal that maintains a WM representation in an active state. The empirical "flipping" effect was best modeled by a suppression of the gain of perceptual feature channels corresponding to the value of the IMI, combined with an intermediate level of memory strength.

Integrating across the findings from the perceptual decision-making [17, 18] and WM [15, 16] literatures that we have reviewed here has given rise to the idea that is the principal motivation for this Preregistered Research Article: The active removal of information from WM may be implemented via a top-down "hijacking" of an adaptation-like modification of perceptual circuits, paired with a weak pulse of (top-down) activation. More specifically, this model posits that active removal from WM is accomplished by the co-occurrence of two events. The first is the adaptation-like modulation of the gain of the perceptual channels that were engaged by the encoding of the to-be-removed item. (This is illustrated by the dip in the "level of gain" in Figure 4.) This putative operation is "adaptation-like" because it is triggered by the onset of the retrocue (not by the perceptual processing of the to-be removed item, which occurred at the beginning of the trial) and because it is greater in magnitude than is typical of perceptual adaptation (the repulsive effects of perceptual adaptation are typically weaker than the attractive influence of recent decisions). The second event, which is hypothesized to occur concurrently, is the brief, weak activation of this item (illustrated by the lower level of top-down "activation," relative to the PMI, in Figure 4). (It is important to note that this hypothesized mechanism differs in important details from a different hypothesized mechanism that is not being investigated here, the nonmonotonic plasticity hypothesis, which predicts weakening and forgetting of memories as a direct consequence of moderate reactivation [e.g., 26, 27, 28]. In highjacked adaptation, the construct of "weak activation" corresponds to the "memory strength" parameter in the simulation of Lorenc, Vandenbroucke et al. [15], which combines with a decrease in a distinct "gain" parameter in the model [15]. In our conceptualization of hijacked adaptation, these two effects are caused by two distinct top-down control signals, although a direct test of this possibility is outside the scope of the present work.)

We plan to assess this hijacked-adaptation model by collecting fMRI data while subjects perform an ABC-retrocuing task (Figure 2) while high-contrast task-irrelevant visual stimuli are flashed to "ping" the visual system, so as to assay predicted consequences of this hypothesized mechanism for active removal. In the final subsection of the Introduction, we provide a narrative overview of the logic of our proposed experiment, and how it will operationalize tests of the three

models of active removal that we have reviewed here: hijacked adaptation; context breaking; and context shifting. This will provide context for understanding our *Preregistered Hypotheses*.

Operationalizing tests of three models of active removal from WM

Based on Shan & Postle [16], we assume that the IMI undergoes active removal in the *overlap* condition of the ABC-retrocuing task, but passive removal in the *no-overlap* condition.

Hijacked adaptation. As diagrammed in Figure 4, in the *overlap* condition, the hypothesized hijacked-adaptation operation is expected to produce a phasic "flipping" of the active representation of the IMI (operationalized as an IEM reconstruction of the IMI with a negative slope) during the first several seconds following the retrocue (i.e., early Delay 2.1; note that this would constitute a replication of Lorenc, Vandenbroucke et al. [15]), followed by a disappearance of a detectable active trace (i.e., an IEM reconstruction slope not different from 0). This will correspond to successful removal of the IMI. A longer-lasting consequence of active removal, however, will be the residual adaptation-like change to the gain of perceptual feature channels that correspond to the orientation of the IMI. This will be revealed in the filtering of the ping-evoked response (at TRs 15+16), which will also produce a transient flipped IEM reconstruction of the IMI. Note that the delay period after the retrocue and before the ping (i.e., Delay 2.1) will be relatively long (15.25s), so as to be able to dissociate the endogenously generated flipped reconstruction of the IMI that is triggered by the retrocue (i.e., during early Delay 2.1) from the flipped reconstruction predicted to be evoked by the ping (at TRs 15+16). (Note that this is a novel prediction, in that, e.g., Lorenc, Vandenbroucke et al. [15] did not assess the state of representation of the IMI several seconds after the retrocue, as we will do here.) In the *no-overlap* condition, we predict that the withdrawal of attention will result in the disappearance of evidence for an active representation of the IMI during the first several seconds following the retrocue (i.e., early Delay 2.1). However, because the activity-silent trace of the IMI will not have been removed, the ping-evoked response will produce a conventional (i.e., not flipped) IEM reconstruction of the IMI (at TRs 15+16; c.f., [5]).

This pattern of results (summarized in Figure 4 and formalized in the statement of *Preregistered hypotheses* in *Methods*) would provide neural evidence that the active removal of information from WM can be accomplished via a mechanism of hijacked adaptation. It would also provide evidence relevant for accounts of the repulsive serial bias that is sometimes observed with perceptual discrimination tasks (e.g., [17, 18]). (We note that the viability of the hypothesis tests described here, as well as those that pertain to the context-breaking model (described next), depends on the ability to track, with IEM, the simultaneous representation of two separate items held in WM. Our group has done this successfully in fMRI studies of DSR-with-orientations [25] and DSR-with-direction-of-motion [23], and of in an EEG study 2-back WM for orientations [24].)

(We note that the phenomenon of a flipped IEM reconstruction has also been described in studies that manipulate the priority of items held in WM [24, 25]. For example, in the DSR task when a retrocue designates an item a UMI, the IEM reconstruction of its orientation flips in early visual cortex (but not in IPS), and the IEM reconstruction of its location flips in IPS (but not in early visual cortex). When considered from the perspective of the levels-of-analysis framework of Marr and Poggio [29], however, prioritization and active removal differ in fundamental ways. At the computational level, there are two discrete problems to be solved: holding an item in WM in a deprioritized state [24, 25] versus actively removing an item from WM (the focus of this *Preregistered Research Article*). At the algorithmic level, we belive it is also likely that the two differ profoundly. We have argued elsewhere that deprioritization is accomplished via a mechanism of "rotational remapping"[30], whereas here the mechanism that we are proposing for active removal is hijacked adaptation (the simultaneous suppression of the gain of perceptual feature channels corresponding to the value of the IMI, combined with an intermediate level of activation of that representation). Thus, it is only at the implementation level that priority-based remapping and hijacked-adaptation may both produce "flipped" IEM reconstructions. (For an in-depth consideration of caveats when inferring underlying physiological processes from IEM reconstructions, see [31-34].) Additionally, hypothesis 4 will address predicted differences for IEM reconstructions associated with hijacked adaptation relative to priority-based

remapping (i.e., [24, 25].)

Context-breaking. This model predicts a failure to reconstruct the IMI during the during early Delay 2.1 (TR 7) in the *overlap* condition. This is because the hypothesized unbinding operation, whereby the association between the content of the memory item and its context is actively broken, has the effect of removing the item from WM [12], and thus removing the active trace needed to successfully reconstruct it with IEM. To our knowledge, the context-breaking model does not make an explicit assumption about the possible existence of residual activity-silent traces of removed items. Thus, in the *overlap* condition, either a failure to reconstruct the IMI following the ping (at TRs 15+16; consistent with the absence of a residual activity-silent trace), or a reconstruction with a positive slope (consistent with a residual activity-silent trace of the item in its PMI format), could both be compatible with this model.

Context-shifting. Because this model makes very different assumptions about how WM is organized and controlled, its predictions will be tested with a fundamentally different set of analyses. Rather than 'active removal' per se, the context shifting model predicts larger mental context shifts on *overlap* versus *no-overlap* trials. We will evaluate this prediction by assessing pattern similarity between Delay 1 (at TR 4) and the late portion of Delay 2.1 (at TR 12), in early visual cortex and in entorhinal cortex. In the *overlap* condition, the context-shifting model predicts a larger shift of mental context such that this discrepancy between mental contexts can compensate for the elevated level of cue competition.

(Note that, although the observation of repulsive versus attractive serial biases in previous datasets was important for developing the model of hijacked adaptation (e.g., [16-18]), we consider these effects to be consequences of how a stimulus removed from WM (i.e., actively or passively), rather than of central relevance to the mechanism of removal. Because none of our tests of the three models of removal that are the focus of this Preregistered Research Article involve serial bias as a dependent measure, it is a deliberate choice that our design will not lend itself to analyses of serial dependence effects.)

Figure 4. The mechanism of hijacked-adaptation, illustrated via the hypothesized states of perceptual circuits that encode

and maintain stimulus information in WM during different epochs of the trial. The rows of panels above and below the timelines correspond to four elements of the model: 1) the colored bars represent the activation levels of hypothetical orientation-tuned perceptual channels; 2) the red lines represent the level of top-down activation allocated to each of the two memory items; 3) the black lines represent the level of gain of the perceptual channels, with the default value of each being 1.0, and hijacked-adaptation resulting in channel-specific decreases in this value; 4) and the cartooned networks (Note that the level-of-gain in this figure only reflects the influence of top-down hijacked adaptation; the effects of true perceptual adaptation on channel gain are assumed to be too subtle to be detectable at the scale of the processes being illustrated here.) In the *no-overlap* trial (top five rows), each sample item is represented during Delay 1 with 1) elevated activity in the orientation channels corresponding to their value; 2) comparable levels of top-down activation; 3) baseline levels of gain at each channel; and 4) an activity-silent representation. During Delay 2.1, the representation of the retrocued item (i.e, *A*) remains elevated because its level of top-down activation remains unchanged. The active representation of the IMI (*B*), however, drops to baseline, because it is no longer receiving top-down activation. Importantly, however, the activity-silent representation of *B* remains. Early in Delay 2.2, the ping nonspecifically raises the activity level in every orientation channel. This produces a reactivation of *B*, because the activity from the ping is filtered through the activity-silent representation of *B.* In the *overlap* trial, the cuing of *A* prompts the active removal of *B* via highjacked adaptation: a coordinated decrease in the gain of the channels corresponding to *B* (illustrated by the orientation-specific dip in the gain field) plus a weak phasic activation of these channels (illustrated by the lower level of top-down activation, relative to the retrocued item) that occur during the early portion of Delay 2.1. The effect of these events is effectively instantaneous, and is two-fold: at the level of channel activity they produce an activity-based representation of *B* that is "flipped" (and labeled "IMI"); and at the level of activity-silent representation, the representation of *B* is removed (illustrated with dotted lines) due to synaptic weakening produced by the weak activation paired with the orientation-specific reduction of gain. Later during Delay 2.1, the modified gain field persists but this is not evident in the activity of the perceptual channels with only baseline levels of activity corresponding to the value of the IMI. Finally, early in Delay 2.2, responses to the ping, filtered through the modified gain field, produce a transient pattern of activity that is also a "flipped" version of *B*.

Methods

Preregistered hypotheses

We propose to test 7 primary hypotheses in this Preregistered Research Article:

<u>*Hypothesis 1a*</u>: In the *overlap* condition, the reconstruction of the orientation of the IMI during early Delay 2.1 (TR 7), with an IEM trained on the retrocued item at TR 7, will have a significantly negative slope. *(Rationale: This pattern of a "flipped" IEM reconstruction, replicating [15], is hypothesized to be a consequence of hijacked adaptation. A failure to confirm this hypothesis, in contrast, would be consistent with the context-breaking model, because the active*

representation would simply no longer be in WM. The context-shifting model does not make a prediction either way, but would need to be modified to accommodate confirmation of this hypothesis.)

Hypothesis 1b: In the *no-overlap* condition, the reconstruction of the orientation of the IMI during early Delay 2.1 (TR 7), with an IEM trained on the retrocued item at this time (i.e., TR 7), will have either a small positive slope (smaller than the retrocued item) or a slope not different from 0. (*Rationale: Because this condition is assumed to involve passive removal, from the perspective of the hijacked adaptation model no correlate of active removal is expected. Thus, for this model, the critical prediction is that the reconstruction of the IMI will not have a negative slope. For all three models, passive removal should manifest as a gradual decline in the strength of the active trace as it "fades away;" none of them makes explicit predictions about the time course of passive removal.*

Hypothesis 1c: The slopes from 1a and 1b will differ. *(Rationale: If the "flipped" IEM reconstruction is specific to active removal (predicted by hijacked-adaptation model), the slopes of IEM reconstructions from the two conditions should differ. Confirmation of this hypothesis would provide quantitative evidence that the two conditions differ in terms of the processing of the IMI (active vs. passive removal). This outcome is a necessary precondition for the subsequent hypotheses about predicted consequences of hijacked activation to be valid. Failure to confirm this hypothesis would be consistent with both the context-breaking and context-shifting models, because neither predicts "flipped" IEM reconstruction of the IMI.)*

Hypothesis 2a: In the *overlap* condition, if an IEM can be successfully trained to reconstruct the retrocued item during late Delay 2.1 (i.e., at TR 12), the reconstruction of the orientation of the IMI at this same time point, with that same IEM, will be unsuccessful (i.e., slope not different from 0). *(Rationale: This is a sanity check for both the hijacked-adaptation and context-breaking model, which predict that active removal will have removed any active trace of the IMI. For this hypothesis, bootstrap testing will be supplemented with calculation of Bayes Factors. The context-shift model, in*

contrast, does not make a strong prediction, because a change of context between early- vs. late-Delay 2.1 might only weaken, but not obliterate, cross-condition testing with IEM [c.f., [35].])

Hypothesis 2a' (if needed): In the *overlap* condition, if an IEM cannot be successfully trained to reconstruct the retrocued item during late Delay 2.1 (i.e., at TR 12), the reconstruction of the orientation of the IMI at this same time point with an IEM trained on the retrocued item during early Delay 2.1 (i.e., at TR 7) will be unsuccessful (i.e., slope not different from 0). *(Rationale: This is an alternative way to carry out the same sanity check from* Hypothesis 2a*, if* 2a *cannot be tested.)*

Hypothesis 2b: In the *no-overlap* condition, if an IEM can be successfully trained to reconstruct the retrocued item during late Delay 2.1 (i.e., at TR 12), the reconstruction of the orientation of the IMI at that time point, with that same IEM will be unsuccessful (i.e., slope not different from 0). *(Rationale: This is not a strong test of any of the three hypotheses, merely a statement of the expectation that there will no longer be a detectable active trace of no-longer-relevant item (the IMI) at the end of Delay 2.1 (i.e., 14 sec after the retrocue designated it the IMI).)*

Hypothesis 2b' (if needed): In the *no-overlap* condition, if an IEM cannot be successfully trained to reconstruct the retrocued item during late Delay 2.1 (i.e., at TR 12), the reconstruction of the orientation of the IMI during late Delay 2.1 with an IEM trained on the retrocued item from early Delay 2.1 (i.e., TR 7) will be unsuccessful (i.e., slope not different from 0). *(Rationale: This is an alternative way to confirm the same expectation as described for* Hypothesis 2b*, if* 2b *cannot be tested.)*

Hypothesis 3a: In the *overlap* condition, if an IEM can be successfully trained to reconstruct the retrocued item from the ping-evoked response (i.e., at TR 15+16), the reconstruction of the orientation of the IMI from the ping-evoked response (i.e., at TRs 15+16), with that same IEM, will have a significantly negative slope. *(Rationale: This is a key prediction of the hijacked-adaptation model, which is that the persistence of the pattern of channel-specific gain modification (resultant from the application of this mechanism to effect active removal of the IMI) – the same phenomenon hypothesized to be*

responsible for the repulsive serial bias effect [16] – will be revealed when signals from the ping are filtered through these perceptual channels. Neither the context-breaking nor the context-shifting model would be able to account for this predicted outcome.)

Hypothesis 3a' (if needed): In the *overlap* condition, if an IEM cannot be successfully trained to reconstruct the retrocued item from the ping-evoked response (i.e., from TRs 15+16), the reconstruction of the orientation of the IMI from the ping-evoked response (i.e., from TRs 15+16) with an IEM trained on the retrocued item at TR 7 will have a significantly negative slope. *(Rationale: This is an alternative way to test the prediction of* Hypothesis 3a*, if* 3a *cannot be tested.)*

Hypothesis 3b: In the *no-overlap* condition, if an IEM can be successfully trained to reconstruct the retrocued item from the ping-evoked response (i.e., from TRs 15+16), the reconstruction of the orientation of the IMI from the ping-evoked response (i.e., from TRs 15+16) with that same IEM will have a significantly positive slope. *(Rationale: Because passive removal is assumed to leave the (putative) activity-silent representation of the IMI intact (c.f. [4, 5]), when signals from the ping interact with this activity-silent representation of the IMI (e.g., are filtered through it [9] or, in the sonar metaphor, "bounce off it" [36, 37]) the ping-evoked response will reveal this residual activity-silent representation of the IMI (c.f., [36, 37].)*

Hypothesis 3b' (if needed): In the *no-overlap* condition, if an IEM cannot be successfully trained to reconstruct the retrocued item from the ping-evoked response (i.e., from TRs 15+16), the reconstruction of the orientation of the IMI from the ping-evoked response (i.e., from TRs 15+16), with an IEM trained on the retrocued item from early Delay 2.1 (i.e., from TR 7) will have a significantly positive slope. *(Rationale: This is an alternative way to test the prediction of* Hypothesis 3b*, if* 3b *cannot be tested.)*

Hypothesis 3c: The slopes from *3a* and *3b* will differ. *(Rationale: If the persistence of the pattern of channel-specific gain modification is specific to active removal (predicted by hijacked-adaptation model), the slopes of IEM reconstructions*

from the two conditions should differ. Confirmation of this hypothesis would provide quantitative evidence that the persistent effects of active vs. passive removal differ in the way predicted by the hijacked-adaptation model.)

Hypothesis 3c' (if needed): The slopes from 3a' and 3b' will differ. *(Rationale: This is an alternative way to test the prediction of* Hypothesis 3c*, if* 3c *cannot be tested. Neither the context-breaking nor the context-shifting model would be able to account for this predicted outcome.)*

Hypothesis 4: In the *overlap* condition, in the intraparietal sulcus (IPS) ROI, the IEM reconstruction of the location of the IMI during early Delay 2.1 (TR 7), with an IEM trained on the retrocued item at TR 7, will not have a significantly negative slope. *(Rationale: A "flipped" IEM reconstruction of the location, but not the orientation, of the unattended memory item (UMI) in IPS has been reported elsewhere [25], and interpreted as reflecting the deprioritization of the context of the UMI by a mechanism of "priority-based remapping" [25]. In the ABC-retrocuing task, in contrast, applying the mechanism of hijacked-adaptation to the representation of the location of the IMI could harm performance, because "active removal" of this location might impair the ability to encode item* C*, which will be presented at this same location. Such an outcome would illustrate an important difference between the mechanisms underlying flipped IEM reconstructions observed with priority-based remapping [16, 24, 25] versus with hijacked adaptation [15].)*

Hypothesis 5: In the overlap condition, with an IEM trained on the retrocued item at TR 7, the baseline of the reconstruction of the orientation of the IMI from the ping-evoked response (i.e., from TRs 15+16, Delay 2.2; baseline estimated from the fit of a exponentiated cosine function [c.f., 23, 38]) will be higher than the baseline of the reconstruction of the orientation of the IMI during early Delay 2.1 (TR 7). (*Rationale: This is not a strong test of any of the three models, but a statement of the expectation that the visual ping will produce a reconstruction of the orientation of*

436 *the IMI with a higher baseline, because it will be embedded in an evoked response (Fig. 4; c.f. the IEM reconstruction*
437 *from the mask-evoked response in [30]).*)
438
439 *Hypothesis 6a*: In the anterior lateral entorhinal cortex (alEC) ROI, the multivoxel pattern similarity between late Delay 1
440 (TR 4, just prior to the retrocue) and late Delay 2.1 (TR 12, just prior to the ping) will be higher in the *no-overlap*
441 condition than in the *overlap* condition. *(Rationale: Using multivoxel pattern similarity between timepoint A and*
442 *timepoint B within the same trial as a neural correlate for drift in mental context, the context-shifting model predicts a*
443 *greater multivoxel pattern* dissimilarity *in conditions requiring "active removal" due to high spatial overlap. Previous*
444 *work by Bellmund et al. [39] has implicated alEC as an important neural substrate for the representation of mental*
445 *context.)*
446 *Hypothesis 6b*: In the early visual ROI, the multivoxel pattern similarity between Delay 1 (TR 4, just prior to the retrocue)
447 and Delay 2.1 (TR 12, just prior to the ping) will be higher in the *no-overlap* condition than in the *overlap* condition.
448 *(Rationale: Because little is known about the neural correlates of mental context, it seems reasonable to speculate that a*
449 *strategic shift of mental context may also manifest itself in a region associated with the representation of the stimuli*
450 *being held in WM.)*
451 *Hypothesis 6c*: In the posterior medial EC (pmEC) ROI, the pattern similarity between TR 4 and TR 12 will not differ across
452 conditions. (*Rationale: The same previous study by Bellmund et al. [39] did not find evidence for the encoding of mental*
453 *context in pmEC, and so this region offers an a priori test for the specificity of Hypothesis 6a, should 6a be confirmed.)*
454
455 **Summary Table**

Note: Each of the three models being assessed here can be construed as a research question: the hijacked-activation (HA) model, the context-breaking (CB) model, and the context-shifting (CS) model; the sampling plan is the same for all hypotheses (n = 30).

Hypothesis (model)	**Analysis**	**Pre-specified outcomes**
1a (HA and CB)	Bootstrapping	significantly negative slope = consistent with HA; slope not different from 0 = consistent with CB
1b (HA and CB)	Bootstrapping	slope not different from 0 = consistent with HA and CB; slope different from 0 (either direction) means procedure failed to elicit expected initial signature of passive removal
1c (HA , CB, CS)	Bootstrapping	*1a* and *1b* differ = consistent with qualitative difference between active and passive removal; *1a* and *1b* do not differ significantly suggests no mechanistic difference between "active" and "passive" removal (= inconsistent with HA and CB); = consistent with CS
2a (HA and CB)	Bootstrapping plus Bayes Factors	Slope not different from 0 = evidence for an active trace of the IMI is absent, a necessary condition for interpreting the effect of the ping; slope different from 0 may complicate interpretation of the effect of the ping.
2a' (HA and CB)	Bootstrapping plus Bayes Factors	This is an alternative way to assess *2a*, in the event that an IEM can NOT be successfully trained to reconstruct the retrocued item at TR 12; interpretation of outcomes is the same as *2a*
2b (HA)	Bootstrapping	Same as *2a*, except for *no-overlap* condition
2b' (HA)	Bootstrapping	Same as *2a'*, except for *no-overlap* condition
3a (HA and CB)	Bootstrapping	Negative slope = evidence for a key prediction of HA, the adaptation-like gain modulation of orientation-tuned sensory channels; positive slope =

		disconfirmation of HA; not different from 0 = failure to find evidence for HA, but consistent with the breaking of content-context association of the IMI [12]
3a' (HA and CB)	Bootstrapping	This is an alternative way to assess 3*a*, in the event that an IEM can NOT be successfully trained to reconstruct the retrocued item at TRs 14+15; interpretation of outcomes is the same as 3*a*
3b (HA)	Bootstrapping	Positive slope = evidence for a residual activity-silent representation of the IMI, an assumed consequence of passive removal; not different from 0 = failure to replicate Bae & Luck [4] and Barbosa, Stein et al. [5]
3b' (HA)	Bootstrapping	This is an alternative way to assess 3b, in the event that an IEM can NOT be successfully trained to reconstruct the retrocued item at TRs 14+15; interpretation of outcomes is the same as 3b
3c (HA and CB)	Bootstrapping	*3a* and 3*b* differ = consistent with qualitative difference between active and passive removal; *3a* and 3*b* do not differ would suggest no mechanistic difference between "active" and "passive" removal, which would be inconsistent with HA and CB, but would not pose a problem for CS.
3c' (HA and CB)	Bootstrapping	This is an alternative way to assess *3c*, in the event that either *3a'* or *3b'* were needed
4 (HA)	Bootstrapping	Absence of a negative slope will offer evidence that the mechanisms implementing HA are different from those implementing priority-based remapping *[16, 24, 25], because retrocue-triggered priority-based remapping does produce IEM reconstructions of stimulus location with negative slopes*

5 (Not applicable)	Bootstrapping	As illustrated in Figure 4, although the IEM reconstruction of the IMI is expected to have a negative slope during both Delay 2.1 (i.e., following the retrocue) and Delay 2.2 (i.e., following the ping), the ping-evoked response is expected to nonspecifically drive a higher response in every channel in the basis set of the IEM (*c.f. the IEM reconstruction from the mask-evoked response in [30]*), resulting in a higher baseline for the IEM reconstruction of the IMI during Delay 2.2 than during Delay 2.1
6a (CS)	Paired t-test	Pattern similarity greater for *no-overlap* than for *overlap* in alEC ROI = support CS model; no difference between conditions in alEC = no support for CS model in alEC.
6b (CS)	Paired t-test	Pattern similarity greater for *no-overlap* than for *overlap* in the early visual ROI = support CS model; no difference between conditions = no support for CS model in the early visual ROI
6c (CS)	Paired t-test	Pattern similarity greater for *no-overlap* than for *overlap* in pmEC ROI = a failure to find specificity for the effect in alEC predicted by Hypothesis 6a; no difference between conditions in pmEC = evidence for the specificity of the effect in alEC predicted by Hypothesis 6a.

Subjects

30 subjects who are 18-35 years in age with normal or corrected-to-normal vision and report no history of neurological disease will be recruited from the University of Wisconsin–Madison community. Informed consent will be obtained. All experimental procedures for the Preregistered Research Article have been approved by the University of Wisconsin–Madison Health Sciences Institutional Review Board (protocol ID 2017-0344).

Power analysis. Using data from Yu, Teng, and Postle [25], in which a negative slope of the IEM reconstruction of the UMI in a DSR-of-orientation task has been observed, power analysis of the 2-tailed one sample *t*-test shows we will need data from 30 subjects to achieve 90% power to detect a significantly negative slope for the reconstruction of orientation of the UMI (Cohen's d = 0.617), and data from 26 subjects to detect a significantly positive slope for the reconstruction of orientation of the PMI (Cohen's d =0.675).

To the best of our knowledge, there is no established way to perform power analysis for bootstrapping, which we will use in the current study to test for the predicted positive and negative slopes of reconstructions. We used data from Yu, Teng, and Postle [25] to simulate the *p*-values obtained from *t*-tests versus from bootstrapping with different sample sizes. Because this sample had data from 13 subjects, we generated estimates ranging from *N* = 8 to *N* =12, by randomly drawing N subjects from the sample, without replacement, and conducting a *t*-test and a bootstrap analysis on these data. For the *t*-tests, we collapsed over channel responses on both sides of the target channel, averaged them, and calculated the slope of the averaged UMI reconstruction of each subject with linear regression. The slopes were then compared to 0 with a 2-tailed one sample *t*-test. For bootstrapping, the method was the same as specified in the *Statistical Analyses* subsection of fMRI Analyses section of the *Methods*. This process was repeated 10 times at each N to get 10 (different) sets of subjects and 10 *p*-values for each test. For *N*=13, one *p*-value was obtained from each test. Across sample sizes, the bootstrapping was generally more sensitive than the *t*-test (Figure 5). It has been shown by other researchers that the bootstrap consistently outperformed the t-test in a more systematic way [40]. Based on this, we reason that the sample size estimated by the power analysis for a *t*-test provides a conservative estimation of the sample size required in the current study (because we will be using the more sensitive bootstrapping procedure). In the current study we will use a sample size of 30 subjects.

Figure 5. The *p*-values obtained from bootstrapping tests and *t*-tests of subsets of subjects from Yu, Teng, and Postle [25].

The darker stars overlaid on the dots represent the mean and the error bars show the standard deviation of each set of *p*-values.

Stimuli and procedure

The stimulus presentation and response collection will be implemented with MATLAB (MathWorks, Natick, MA, USA) with the Psychtoolbox-3 extensions [41, 42]. The display will be projected into the scanner and onto a mirror mounted on the head coil at 60-Hz (Avotec Silent Vision 6011 projector; Avotec, Stuart, FL, USA). The viewing distance will be roughly 69 cm and the screen width will be 33.02 cm. The sample stimuli will be grayscale sinusoidal gratings (radius = 3°; contrast = 0.6; spatial frequency = 1 cycles/°; random phase angle) presented on gray background (L= 52, a = 0, and b = 0 in CIEL*ab space). There will be six possible sample orientations: 20°, 50°, 80°, 110°, 140°, 170°; with a random jitter of ±0°- 3° added with each presentation. These and all ensuing stimuli will appear at any of six possible locations on an imaginary circle centered on fixation (radius of 8°, locations centered at each of these polar angles: 30°, 90°, 150°, 210°, 270°, 330°). The retrocue will be a white circle (thickness=0.08°) with the same radius as the sample stimuli. Ping stimuli will be high contrast concentric circles with the same radius and spatial frequency as the gratings (contrast=1). The response dial will be a black bar (thickness=0.08°) corresponding to the diameter of a black circle with the same radius as the gratings (thickness=0.08°). Subjects will be instructed to adjust the orientation of the bar using an MR-compatible trackball (Current Designs, Philadelphia, PA, USA) and to report their response by pressing a button on the trackball when the orientation of the bar matches their memory for the probed sample. After the button is pressed by the subject, the black bar of the response dial will become thicker (thickness=0.16°) to indict the response has been made and cannot be changed. A white fixation dot will be present throughout each block (i.e., also during the ITI).

Each trial of ABC retrocuing will start with the simultaneous presentation of two samples (*A* and *B*; 1 s) followed by *Delay 1* (7 s). Next the retrocue will appear for 0.75 s at the location that had been occupied by either *A* or *B*, thereby

designating a PMI (which might be tested at the end of the trial) and, by implication, the IMI (no longer relevant for that trial). The retrocue will be followed by *Delay 2.1* (15.25 s), which will be followed by the simultaneous presentation (0.25 s) of ping stimuli at each of the six locations, then *Delay 2.2* (7.75 s), then sample item *C* (1 s), then *Delay 3* (1 s). Finally, the response dial will appear at the location that had been occupied by the retrocued item or by item *C*, prompting the recall of the orientation of that item (4-s response window). The inter-trial interval ITI will vary randomly between 6, 8, and 10 s.

On each trial the orientation of items *A* and *B* will be randomly selected, with replacement, from the pool of six possible values. The locations of item *A* and *B* will be randomly selected from the six possible locations. To fully cross the orientations of item *A* and *B*, 21 unique trials are required. 252 trials (12 repetitions per unique trial) will be used for each condition. The retrocuing of *A* or *B* will be randomly determined on every trial. The orientation of item *C* will be randomly selected from the pool of six possible values (i.e., independent of *A* and *B*), and its location will depend on condition: in the *overlap* condition it will appear at the location that had been occupied by the uncued item; in the *no-overlap* condition it will appear in a location randomly selected from the four that had not been occupied by *A* or *B*. The retrocued item or the item *C* will be probed for recall equiprobably.

Trials will be blocked by condition (*overlap*, *no-overlap* condition), and subjects will be explicitly informed of the condition before the start of each block. Each subject will participate in 4 scanning sessions. The first scanning session will consist of 6 runs, each run corresponding to a 14-trial block. The three remaining scanning sessions will each consist of 10 runs (each run corresponding to a 14-trial block). There are fewer runs in the first session due to acquisition of structural images. To facilitate the consistent use of active removal and passive removal, within each session the first 3 blocks (for the first session) or 5 blocks (for the last three sessions) will be of one condition and the remaining blocks will be of the other condition. The order of conditions within a session will be counterbalanced across sessions and across subjects. In the first session, each subject will first do two practice blocks (one block for each condition) outside the

scanner and another practice block (with the same condition as the first real block) inside the scanner. An Avotec RE-5700 eye-tracking system (Avotec) will be used to track eye position throughout each scanning session, and to assure that subjects' eyes are open during the ping.

Behavioral Data Analysis

The mean absolute error of recall across subjects will be calculated for each condition separately. The performance across the two conditions will be compared with a paired *t*-test.

fMRI Data Acquisition

Whole-brain images will be acquired at the Lane Neuroimaging Laboratory at the University of Wisconsin–Madison HealthEmotions Research Institute (Department of Psychiatry) using a 3 Tesla GE MR scanner (Discovery MR750; GE Healthcare, Chicago, IL, USA). A high-resolution T1 image will be acquired with a fast spoiled gradient recalled echo sequence (8.2 ms TR, 3.2 ms TE, 12˚flip angle, 176 axial slices, 256 × 256 in-plane, 1.0 mm isotropic) for each session. Functional data will be acquired with a gradient-echo echo-planar sequence (2 s repetition time [TR], 22 ms echo time [TE], 60˚flip angle) within a 64 × 64 matrix (42 axial slices, 3 mm isotropic).

fMRI Data Preprocessing

fMRI data will be preprocessed with the Analysis of Functional Neuroimages (AFNI) package (https://afni.nimh.nih.gov). To achieve a steady state of tissue magnetization, the first four TRs of each run will be discarded. The data will then be registered to the final volume of each scan and then to the anatomical images from the first session. Volumes will be motion corrected with six nuisance regressors to account for mead motion artifacts. Linear,

quadratic, and cubic trends will be removed for each run and the z-scores of fMRI time series data will be calculated within each run.

fMRI Analyses

Task-related activity. The fMRI data will be fitted to a general linear model (GLM) with regressors for each epoch of the task -- *Encoding A&B* (2 s), *Delay 1* (6 s), *Delay 2.1* (16 s), *Delay 2.2* (8 s), *Encoding C* (2 s), *Recall* (4 s) – each convolved with a canonical hemodynamic response function, as well as nuisance covariates for between-trial and between-scan drift, and head motion.

ROI creation. Hypothesis tests will be carried out in an early visual cortex ROI , an IPS ROI and two entorhinal cortex (EC) ROIs – anterior lateral (al)EC and posterior medial (pm)EC . First, an anatomically defined ROI of early visual cortex will be created from masks corresponding to V1 and V2 (merged, both hemispheres), and an anatomically defined ROI of IPS (comprising IPS0–5; merged, both hemispheres), both based on the probabilistic atlas of Wang and colleagues [43] and warped to each subject's structural scan in native space. Hypothesis testing will be carried out in the 500 voxels within the anatomical early visual cortex ROI with have the strongest weights on the Encoding A&B regressor, which we refer to as the early visual ROI. For the IPS, hypothesis testing will be carried out in the 500 voxels within the anatomical IPS ROI with have the strongest weights on the Delay 2.1 regressor. For the temporal ROIs, following the practice of Bellmund et al. [39], we will co-register masks from Navarro Schröder et al. (2015; [44]) from standard MNI space (1 mm) to each participant's functional space. The subregion masks from Navarro Schröder et al. (2015) will be each intersected with participant-specific EC masks obtained from their structural scan using the automated segmentation in Freesurfer (version 5.3) to improve anatomical precision for the masks.

Inverted Encoding modeling. IEM analyses will be performed with custom functions in MATLAB. In IEM, the responses of each voxel are assumed to be a weighted sum of responses of several hypothetical tuning channels. Six

tuning channels of orientation (or location for Hypothesis 4) will be used and the tuning curve of each channel will be defined as a half-wave–rectified sinusoid raised to the eighth power. We will first compute the weight matrix W (v voxels × k channels) that projects the hypothesized channel responses C1 (k channels × n trials) to the measured voxel responses B1 (v voxels × n trials) with the training dataset to get the estimate of the weight matrix $\widehat{W}$. Then we use $\widehat{W}$ to reconstruct the channels responses $\widehat{C_2}$ from the voxel activities B2 of the testing dataset. The relationship between B1, W and C1 will be characterized by

$$B_1 = WC_1$$

The least-squared estimate of the weight matrix ($\widehat{W}$) will be calculated using linear regression:

$$\widehat{W} = B_1{C_1}^T(C_1{C_1}^T)^{-1}$$

The channels responses $\widehat{C_2}$ of the testing dataset will then be calculated with the weight matrix ($\widehat{W}$) and the BOLD data (B_2):

$$\widehat{C_2} = (\widehat{W}^T\widehat{W})^{-1}\widehat{W}^T B_2$$

The IEMs will be trained with the orientation (or location for Hypothesis 4) of the retrocued item and test on the orientation (or location) of the retrocued item or the IMI at each TR after the offset of item A and B and before the onset of item C. We will use a leave-one-run-out procedure in which the model will be trained with data of all but one run and tested on the left-out run. This process will be repeated until the reconstruction of all runs is acquired. The estimated channels responses will be centered on the orientation (or location) of the tested item. The reconstruction will be generated on all TRs but our pre-registered hypotheses will focus on specific TRs: *Hypothesis 1a, 1b, 1c and 4* will focus on TR 7; *Hypothesis 2a and 2b* will focus on TR 12. For *Hypothesis 3a, 3b and 3c*, the averaged BOLD of TR 15 and TR 16 will be used to train and test the model. In case we cannot get a reliable reconstruction of the retrocued item at TR 12 and/or TRs 15+16 due to the representation of the retrocued item shifts to an activity-silent state after a long time span

since the presentation of the item, the retrocued item at TR 7 will be used to train the model and TR 12 (for *Hypothesis 2a and 2b*) and/or TRs 15+16 (for *Hypothesis 3a, 3b and 3c*) will be tested.

For all analyses of orientation (i.e., Hypotheses 1-3 & Hypothesis 5) IEM training will collapse across the location at which items appeared on the screen. This choice is justified for two reasons. First, in a previous study that presented orientations at only four different locations (and therefore collected enough data to train location-specific IEMs of orientation), location-specific IEMs were found to be only subtly numerically superior (i.e., higher reconstruction slopes) to location-nonspecific IEMs trained with the same number of trials [35]. Second, in a previous study that presented orientations at eight different locations (and therefore did not collect enough data to train location-specific IEMs of orientation), location-nonspecific IEMs or orientation were robust [25]. (Indeed, it is with data from Yu, Teng, and Postle [25] that we performed power calculations for this study.)

Statistical Analyses for Hypotheses 1 - 4. The strength of IEM reconstructions of memory items will be operationalized by their slope. We will collapse over channel responses on both sides of the target channel, average them, and calculate the slope of the reconstruction with linear regression for each subject separately.

We will use bootstrapping to test the statistical significance of the group-average slope of each reconstruction [38, 45]. For each hypothesis test, we will randomly sample 30 reconstructions from the pool of 30 (one per subject), with replacement, and calculate the average of the channel responses. This process will be repeated 10,000 times to get 10,000 resampled group-average reconstructions, and the slopes of these reconstructions will be calculated. Two-tailed *p*-values will be computed as the proportion of positive or negative slopes, whichever is smaller, multiplied by 2. Furthermore, for Hypothesis 2a or 2a' (if needed), we will test whether the slopes differ from 0 with Bayes factor using the Bayesian one sample t-test implemented in JASP (version 0.11.1.0). To test the difference between slopes, we will calculate the difference between the 2 slopes of interest for each one of the 10,000 resampled data sets. Two-tailed *p*-values will be the proportion of positive or negative differences, whichever is smaller, multiplied by 2.

Statistical Analyses for Hypotheses 5. To test whether the baselines of IMI-orientation reconstructions differ between early Delay 2.1 and Delay 2.2 under the *overlap* condition, the IEM will be trained with the orientation of the retrocued item on TR 3 (Delay 1) and this IEM will be used to reconstruct the IMI orientation on TR7 (Delay 2.1) and on TRs 15+16 (Delay 2.2). To create a smooth reconstruction with 180 data points, the training and reconstructing will then be repeated 29 times and the centers of the hypothetical tuning channels will be shifted by 1° on each iteration.

We will use bootstrapping to test the difference between these two baselines. We will randomly sample 30 reconstructions from the pool of 30 (one per subject), with replacement, and calculate the average of the reconstructions. Each average reconstruction will then be fit with an exponentiated cosine function:

$$f(x) = \alpha\left(e^{k(\cos(\mu - x) - 1)}\right) + \beta$$

where x ranges from 1 to 180, $f(x)$ is the reconstruction. μ, k, and α control the center, concentration and amplitude of the function, respectively. β is the baseline of the function. Following a previous study in which this analysis was conducted, the fitting will be conducted by combining a general linear model with a grid search procedure and the resampling and fitting will be repeated only 2,500 times [38]. We will calculate the difference between the 2 baselines for each one of the 2,500 resampled data sets. Two-tailed p-values will be the proportion of positive or negative differences, whichever is smaller, multiplied by 2.

Statistical Analyses for Hypothesis 6. In these analyses we will quantify whether the multivoxel pattern similarity between late Delay 1 (TR 4) and late Delay 2.1 (TR 12) will be lower in the *overlap* condition than in the no-*overlap* condition. For each ROI, we will calculate Pearson correlation coefficients between patterns in these 2 TRs for each trial to measure cross-temporal pattern similarity. The correlation coefficients will be Fisher z-transformed and averaged for each combination of subject and condition. To test the prediction of the WMEM model [13] that the neural representation of context shift will be greater in the *overlap* condition (i.e., in the condition that requires "active removal"), we will conduct a two-way paired t-tests of the cross-temporal pattern similarities for each ROI.

Timeline

We anticipate the data collection will take about 1 year. We will carry out data processing and analysis in parallel to data collection as new data are collected. The analyses of data and write-up and submission of the Stage 2 report are expected to be completed within 4 months after all data are collected. If there are pandemic-related interruptions, the project will be delayed accordingly.

References

1. Monsell S. Recency, immediate recognition memory, and reaction time. Cognitive Psychology. 1978;10(4):465-501.
2. Fischer J, Whitney D. Serial dependence in visual perception. Nature Neuroscience. 2014;17:738–43 doi: doi.org/10.1038/nn.3689.
3. Bliss D, Sun JJ, D'Esposito M. Serial dependence is absent at the time of perception but increases in visual working memory. Scientific Reports,. 2017;7:1-13.
4. Bae G-Y, Luck SJ. Reactivation of previous experiences in a working memory task. Psychological Sci. 2019;30(4):587-95.
5. Barbosa J, Stein H, Martinez RL, Galan-Gadea A, Li S, Dalmau J, et al. Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. Nature Neuroscience. 2020;23:1016-24. doi: doi.org/10.1038/s41593-020-0644-4.
6. Chatham CH, Badre D. Working memory management and predicted utility. Frontiers in Behavioral Neurocience. 2013;7:article 83. doi: 10.3389/fnbeh.2013.00083.
7. Lewis-Peacock JA, Drysdale AT, Oberauer K, Postle BR. Neural evidence for a distinction between short-term memory and the focus of attention. J Cog Neuroscience. 2012;24:61-79. PubMed Central PMCID: PMC3222712.
8. LaRocque JJ, Lewis-Peacock JA, Drysdale A, Oberauer K, Postle BR. Decoding attended information in short-term memory: An EEG study. J Cog Neuroscience. 2013;25:127-42. PubMed Central PMCID: PMC3775605.
9. Rose N, Larocque JJ, Riggall AC, Gosseries O, Starrett MJ, Meyering E, et al. Reactivation of latent working memories with transcranial magnetic stimulation. Science. 2016;354:1136-9. PubMed Central PMCID: PMC5221753.
10. Fulvio JM, Postle BR. Cognitive control, not time, determines the status of items in working memory. Journal of Cognition. 2020;3:1-8. doi: https://doi.org/10.5334/joc.98.
11. Oberauer K, Lin H-Y. An interference model of visual working memory. Psychological Review. 2017;124:21-59.
12. Lewis-Peacock JA, Kessler Y, Oberauer K. The removal of information from working memory. Annals of the New York Academy of Science. 2018;1424:33-44.
13. Beukers AO, Buschman TJ, Cohen JD, Norman KA. Is activity silent working memory simply episodic memory? Trends in Cognitive Sciences. 2021;25:284-93. doi: https://doi.org/10.1016/j.tics.2021.01.003.
14. Stokes MG. 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. Trends in Cognitive Sciences. 2015;19:394-405. doi: 10.1016/j.neuron.2013.01.039. PubMed PMID: 23562541.
15. Lorenc ES, Vandenbroucke ARE, Nee DE, de Lange FP, D'Esposito M. Dissociable neural mechanisms underlie currently-relevant,

674 future-relevant, and discarded working memory representations. Scientific reports. 2020;10(1):1-17. doi: doi.org/10.1038/s41598-
675 020-67634-x.
676 16. Shan J, Postle BR. The influence of active removal from working memory on serial dependence. Journal of Cognition. January
677 2021;accepted Stage 1 Registered Report.
678 17. Fritsche M, Spaak E, de Lange FP. A Bayesian and efficient observer model explains concurrent attractive and repulsive history
679 biases in visual perception. eLife. 2020;9:e55389. doi: doi.org/10.7554/eLife.55389.
680 18. Trapp S, Pascucci D, Chelazzi L. Predictive brain: Addressing the level of representation by reviewing perceptual hysteresis.
681 Cortex. in press. doi: https://doi.org/10.1016/j.cortex.2021.04.011.
682 19. Barlow HB. Possible principles underlying the transformation of sensory messages. In: W WR, Sensory Communication.
683 Cambridge MMITPpDhdom, 9780262518420.003.0013, editors. Sensory Communication. Cambridge, MA: MIT Press; 1961.
684 20. Clifford CWG, Wenderoth P, Spehar B. A functional angle on some after-effects in cortical vision. Proceedings of the Royal
685 Society of London Series B: Biological Sciences 2000;267:1705–10. doi: doi.org/10.1098/rspb.2000.1198.
686 21. Gibson JJ, Radner M. Adaptation, after-effect and contrast in the perception of tilted lines. I. quantitative studies. J Exp Psychol.
687 1937;20:453–67. doi: https://doi.org/10.1037/h0059826.
688 22. Jin DZ, Dragoi V, Sur M, Seung HS. Tilt aftereffect and adaptation-induced changes in orientation tuning in visual cortex. Journal
689 of Neurophysiology. 2005; 94:4038–50. doi: https://doi.org/10.1152/jn.00571.2004,.
690 23. Sahan MI, Sheldon AD, Postle BR. The neural consequences of attentional prioritization of internal representations in visual
691 working memory. Journal of Cognitive Neuroscience. 2020;32:917-44. doi: doi.org/10.1162/jocn_a_01517.
692 24. Wan Q, Cai Y, Samaha J, Postle BR. Tracking stimulus representation across a 2-back visual working memory task. Royal Society
693 Open Science. 2020;7:190228 doi: doi.org/10.1098/rsos.190228. PubMed Central PMCID: PMCPMC7481691.
694 25. Yu Q, Teng C, Postle BR. Different states of priority recruit different neural codes in visual working memory. PLoS Biology.
695 2020;18:e3000769. doi: https://doi.org/10.1371/journal.pbio.3000769. PubMed Central PMCID: PMCPMC7500688.
696 26. Norman KA, Newman EL, Detre G. A neural network model of retrieval-induced forgetting. Psychological Review. 2007;114:887–
697 953.
698 27. Lewis-Peacock JA, Norman KA. Competition between items in working memory leads to forgetting. Nature Communications.
699 2014;5:5768. doi: 10.1038/ncomms6768.
700 28. Wang TH, Placek K, Lewis-Peacock JA. More Is less: Increased processing of unwanted memories facilitates forgetting. The
701 Journal of Neuroscience. 2019;39:3551–60.
702 29. Marr D, Poggio T. From understanding computation to understanding neural circuitry. Artificial Intelligence Laboratory AI Memo
703 Massachusetts Institute of Technology. 1976:hdl:1721.1/5782. AIM-357.
704 30. Wan Q, Menendez JA, Postle BR. Rotational remapping between differently prioritized representations in visual working
705 memory. bioRxiv. under review. doi: doi.org/10.1101/2021.05.13.443973.
706 31. Liu T, Cable D, Gardner JL. Inverted encoding models of human population response conflate noise and neural tuning width. J
707 Neurosci. 2018;38:398–408.
708 32. Gardner JL, Liu T. Inverted encoding models reconstruct an arbitrary model response, not the stimulus. eNeuro. 2019;6.
709 33. Sprague TC, Adam KCS, Foster JJ, Rahmati M, Sutterer DW, Vo VA. Inverted encoding models assay population-level stimulus
710 representations, not single-unit neural tuning. eNeuro. 2018;5.
711 34. Sprague TC, Boynton GM, Serences JT. The importance of considering model choices when interpreting results in computational
712 modeling. eNeuro. 2019;6:ENEURO.0196-19.2019. doi: doi.org/10.1523/ENEURO.0196-19.201.
713 35. Cai Y, Sheldon AD, Yu Q, Postle BR. Overlapping and distinct contributions of stimulus location and of spatial context to
714 nonspatial visual short-term memory. Journal of Neurophysiology. 2019;121:1222-31. doi: 10.1152/jn.00062.2019. PubMed Central
715 PMCID: PMC6485733.
716 36. Wolff MJ, Jochim J, Akyürek EG, Stokes MG. Dynamic hidden states underlying working-memory-guided behavior. Nature

Neuroscience. 2017. doi: doi:10.1038/nn.4546.
37. Wolff MJ, Ding J, Myers NE, Stokes MG. Revealing hidden states in visual working memory using electroencephalography. Frontiers in systems neuroscience. 2015;9. doi: 10.3389/fnsys.2015.00123. PubMed PMID: 23562541.
38. Ester EF, Sprague TC, Serences JT. Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory. Neuron. 2015;87:893-905.
39. Bellmund JL, Deuker L, Doeller CF. Mapping sequence structure in the human lateral entorhinal cortex. Elife. 2019;8:e45333.
40. Ahad NA, Abdullah S, Lai CH, Mohd Ali N. Relative power performance of t-test and bootstrap procedure for two-sample. Pertanika Journal of Science & Technology. 2012;20(1):43-52.
41. Brainard DH. The Psychophysics Toolbox. Spatial Vision. 1997;10:433-6.
42. Pelli DG. The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spatial Vision. 1997;10:437-42.
43. Wang L, Mruczek REB, Arcaro MJ, Kastner S. Probabilistic maps of visual topography in human cortex. Cerebral Cortex. 2015;25:3911-31.
44. Schröder TN, Haak KV, Jimenez NIZ, Beckmann CF, Doeller CF. Functional topography of the human entorhinal cortex. Elife. 2015;4:e06738.
45. Ester EF, Sutterer DW, Serences JT, Awh E. Feature-selective attentional modulations in human frontoparietal cortex. J Neurosci. 2016;36:8188-99.