Ann. N.Y. Acad. Sci. ISSN 0077-89

Check for updates

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Special Issue: *Attention in Working Memory*
REVIEW

The removal of information from working memory

Jarrod A. Lewis-Peacock,[1] Yoav Kessler,[2] and Klaus Oberauer[3]

[1]Department of Psychology and Institute for Neuroscience, University of Texas at Austin, Austin, Texas. [2]Department of Psychology and Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Beer-Sheva, Israel. [3]Department of Psychology, University of Zurich, Zurich, Switzerland

Address for correspondence: Jarrod A. Lewis-Peacock, Ph.D., Department of Psychology, Institute for Neuroscience, University of Texas at Austin, Seay 2.222, Austin, TX 78712. jalewpea@utexas.edu

What happens to goal-relevant information in working memory after it is no longer needed? Here, we review evidence for a selective *removal* process that operates on outdated information to limit working memory load and hence facilitates the maintenance of goal-relevant information. Removal alters the representations of irrelevant content so as to reduce access to it, thereby improving access to the remaining relevant content and also facilitating the encoding of new information. Both behavioral and neural evidence support the existence of a removal process that is separate from forgetting due to decay or interference. We discuss the potential mechanisms involved in removal and characterize the time course and duration of the process. In doing so, we propose the existence of two forms of removal: one is *temporary*, and reversible, which modifies working memory content without impacting content-to-context bindings, and another is *permanent*, which unbinds the content from its context in working memory (without necessarily impacting long-term forgetting). Finally, we discuss limitations on removal and prescribe conditions for evaluating evidence for or against this process.

Keywords: working memory; attention; forgetting; inhibition

What is removal?

Removal is *the exclusion of information from working memory in service of the current goal.* It is one of several processes that control the contents of working memory and enable an efficient use of its limited capacity. Some of these control processes support the maintenance of information in working memory by strengthening the information or establishing temporary bindings between items and their context. These include rehearsal, consolidation, and attentional refreshing. Other processes control the flow of information into working memory (input gating), manipulate and modify information in working memory (updating), and select the appropriate items in working memory that need to affect performance (output gating). Removal helps support all these major functions.

This definition of removal has two important implications. First, a necessary condition for demonstrating removal is showing that the information was encoded into working memory before removal took place, and it is no longer in working memory subsequently. This could be achieved by (but is not limited to) demonstrating a diminished set-size effect,[1] or a reduced neural trace,[2–4] following a retrocue during a memory delay informing the person that some of their current working memory contents are no longer relevant. A second implication of our definition is that removal is goal-directed, namely it is an *adaptive* process that supports the current task goal. This characteristic is necessary in order to exclude forms of maladaptive forgetting, such as time-based decay[5] or interference,[6] from our definition. Conceptually, these processes may coexist with removal. At the same time, it can be challenging to distinguish them empirically, and therefore they could also be invoked as alternatives to removal for explaining the same findings. In the final section of this manuscript, we consider experimental approaches to differentiate these putatively independent processes.

Removing information from working memory is necessary to keep working memory up to date with our thinking and acting: The contents of our

doi: 10.1111/nyas.13714

thoughts and our intended actions change at a rapid pace, and so does the environment we attend to. Therefore, information once relevant rapidly becomes irrelevant and needs to be removed to avoid clutter in working memory.[7] In the lab, removal is studied in three scenarios. The first is when a subset of the information in working memory is marked as relevant for the present goal, as in the case of a retrocue,[8] while the rest of working memory contents is marked as irrelevant.[5,6,8,9] In this case, removing the irrelevant items from working memory reduces the goal-relevant memory load, and thereby enhances the maintenance of the remaining items (e.g., by reducing interference).[10] A second scenario that calls for removal is the inadvertent encoding of irrelevant items into working memory, despite being previously marked as distractors, due to an imperfect input selection.[11] The situations in which input selection is done proactively (i.e., through selective gating[12]) versus reactively (through post-encoding removal) still need to be specified, along with the possible modulation of these two strategies by individual differences (cf. Braver *et al.*[13]). A third scenario that involves removal is (partial or complete) updating of working memory contents with new information. Updating paradigms typically require participants to maintain several items in working memory, each related to a specific spatial or temporal position (context), and to update each of them upon demand.[14–18] Removing outdated associations between items and their context is needed in order to enable the creation of new ones, which also helps to reduce proactive interference from previous items that were associated with each contextual cue. Notably, removal can take place in advance, in preparation for updating, even before the new information is available.[15] This implies that removal is not merely a by-product of substituting old items with newer ones, but a separate process[19] that can take place in a proactive manner. Common to all these cases is the need for a removal process to effectively limit the amount of information held in working memory in order to overcome its capacity limitations in the support of flexible, goal-directed behavior.

Evidence for removal

To evaluate whether removal should be considered a separate and unique process for working memory, or whether it may be redundant with other processes, we will review empirical evidence for and against removal. Evidence for removal comes in three forms: (1) an improvement of performance in working memory tasks after irrelevant information has been removed; (2) reduced access to the removed information; and (3) reduced neural activity correlated with the removed information. In this section, we will review each kind of evidence in turn. We will address possible evidence against removal in the final section.

Removal facilitates access to the remaining contents of working memory

When part of the current contents of working memory are removed, the load on the limited capacity of working memory decreases, and this should lead to faster and more accurate access to the remaining, still relevant information. This improvement has been observed in studies in which part of the current memory set is cued as irrelevant after encoding. For instance, in the Modified Sternberg paradigm, participants encode two subsets of items, distinguished by their color or their location on the screen. Subsequently, one of the subsets is cued as relevant and the other as irrelevant for the upcoming recognition test. About 1 s after this cue, the set size of the irrelevant list ceases to affect the reaction time (RT) to the recognition probe.[1,4,20] This effect is reversible: when a subset cued to be irrelevant for a first memory test is subsequently cued to be relevant for a second test, people can do the second test without problems, and their RTs for the second test are again affected by the size of the previously irrelevant but now relevant subset.[20]

Recent experiments have revealed boundary conditions for removal of subsets from working memory: a subset of the current contents of working memory can be removed only if the to-be-remembered and the to-be-forgotten subsets have already been encoded as two distinct groups; removing a subset of items selected at random after encoding is difficult, and perhaps impossible.[21] Yet, when a single item selected at random from the current memory set is cued to be relevant (inviting removal of all other items), access to that item is facilitated, and the number of irrelevant items again ceases to affect RTs about 1–2 s after the cue.[22,23] Moreover, when a single item is cued as relevant within a first memory set, it is easier to add a second

memory set to working memory afterwards. This observation provides evidence that removing all but one item from the first set frees up working memory capacity for the second set.[23]

Experiments with a working memory–updating paradigm provide further evidence that removal of old information facilitates updating. In the updating paradigm,[17,24] participants need to update their working memory contents over several steps, in which they replace one or several of the current memory items by a new stimulus. For instance, they initially encode a list of letters presented across a row of frames. During each subsequent updating step, they see a new letter in one or several of the frames to replace the letter currently remembered for that frame. If the to-be-updated frames are marked ahead of presentation of the new letters, the time participants take for updating working memory (i.e., their RT to each new set of letters) is markedly reduced. This time saving is explained by the preemptive removal of the old letter in each marked frame from working memory before the new letter appears.[14,15]

Removed contents of working memory become less accessible

The perhaps most obvious prediction from the assumption that some information is removed from working memory is that this information should be less accessible afterwards. Testing this prediction is complicated by the fact that when participants are asked repeatedly to report or process the information they had supposedly removed, they will learn quickly that removing that information from working memory may not be necessary, or even worthwhile, if the unloading and reloading of working memory content is more costly than simply maintaining it. Some experimenters have mitigated that problem by inviting participants to drop some information from working memory, and asking them to report that information on only a small subset of trials. For instance, Muter[25] had participants work on a random mixture of primarily two kinds of tasks: when they saw a letter trigram followed by the instruction "LETTERS," they had to recall the letters. When instead the trigram was followed by a three-digit number, they had to count down in threes from that number instead. In about 2% of the trials—toward the end of the experiment—the trigram was followed by a three-digit number, inviting removal of the trigram, but then the instruction "LETTERS" asked for recall of the trigram. Memory for the trigram in these trials was extremely poor already after a retention interval of 2 seconds.[25,26] Williams *et al.*[27] demonstrated an analogous effect in working memory for visual materials: after encoding an array of two colors, one of them was cued to be relevant and tested in 95% of all trials. In 5% of trials, the other color was tested. Memory for the noncued item was barely better than chance merely 1.5 s after the cue, again indicating rapid and thorough erasure of the supposedly irrelevant information from working memory. These examples demonstrate instances in which removed information had reduced accessibility, even to the point of no accessibility, that is, the information was completely forgotten. Removal from working memory has also been linked with longer term forgetting of that information.[28] Not all information that is removed is forgotten, however, as shown by experiments in which recently removed information can be reloaded into working memory without sacrificing memory performance.[2,20]

Other experiments demonstrate more indirectly that access to removed information is reduced. In the working memory updating paradigm, replacing an item by a new item that is identical (or even similar) to the old one facilitates updating as shown by faster updating times.[29] This effect suggests that encoding of the new stimulus commences partially in parallel with removal of the old, so that the representation of the new stimulus in working memory can build on the residual traces of the old one. However, when the to-be-updated item is cued beforehand and enough time is provided between the cue and the replacement stimulus, this facilitating effect is much reduced.[14,15] This is what would be expected if, in response to the cue, the participant removed the old item from working memory before even seeing the new stimulus, so that the old memory trace could no longer facilitate updating when the new stimulus was presented. This situation highlights potential performance costs associated with removal: maintaining irrelevant information in working memory might be preferred if this facilitates its eventual replacement with related information, but only if this facilitation outweighs the costs associated with not removing that information (e.g., maintenance efforts and the potential for interference with relevant information).[30]

Evaluating this tradeoff and responding adaptively would likely be challenging for participants, and this could be a fruitful avenue of future research.

The relevance of efficient removal for working memory can be appreciated by looking at populations who struggle with disengaging from information in working memory. Clinical research on depression has provided evidence that self-reported rumination (i.e., prolonged dwelling on negative thoughts) is correlated with difficulties in removing irrelevant information from working memory,[31–34] but not for people with social anxiety, who also tend to dwell on negative thoughts.[32,33] Understanding the underlying mechanisms of removal may lead to more effective and focused interventions.

Removed contents of working memory become neurally silent

Recent developments in cognitive neuroscience enable researchers to decode—in a coarse manner—the content of representations implemented by ongoing neural activity.[35,36] In this way, electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) signals can be used to monitor which information is represented in a neurally active manner during the retention interval of a working memory task. This method has been applied to the Modified Sternberg task, in which one of two subsets in working memory is cued as relevant for the upcoming recognition probe. Shortly after such a cue, the content of the irrelevant subset can no longer be decoded from neural activity. When after the first memory test a second cue indicates that set to be relevant for a second memory test, the neural activity pattern correlating with the previously irrelevant, now relevant, set comes back.[2,4,37] These results show that neurally active representations in working memory can be temporarily removed, and this removal is reversible, implying that there must also be a neurally silent representation[38] that maintains the information in a state that is not decodable from neural activity with current methods (cf. Schneegans and Bayes[39] and Christophel *et al.*[40] for recent challenges to this interpretation).

The neurally silent representations can be reactivated intentionally by the participant when they are expected to be relevant, but they can also be reactivated by exogenous manipulations of neural activity.[41,42] Rose *et al.*[41] used the two-test Modified Sternberg task and applied transcranial magnetic stimulation (TMS) in the interval following a retrocue, during which the currently irrelevant content was not decodable. The TMS pulse briefly rendered the irrelevant content decodable again. However, this worked only after the first of two successive cues—at a time when the currently irrelevant subset could become relevant again for the second test. After the second test—when the now irrelevant subset was known never to become relevant again—TMS did not render it decodable again. This finding hints at two different kinds of removal, one temporary, and the other permanent. We will address this possible distinction further as we turn our discussion now to the likely cognitive and neural mechanisms underlying removal.

The mechanisms of removal

Removal involves the exclusion of information from working memory in service of the current goal. We propose that this exclusion can be either *temporary* or *permanent* depending on whether the information might be relevant later (Fig. 1).[2,20] These anticipated demands influence how the removal process unfolds in time, and which aspects of the working memory representation it affects. According to many models of working memory,[43,44] encoding requires a representation of the to-be-remembered information that is temporarily bound to a representation of the context in which it was encountered (e.g., the serial position in a list, or the spatial position in an array). Removal could therefore operate on any or all three of these parts: the content, the context, and the binding between them. Removal from working memory does not necessarily imply forgetting of that information from long-term memory (cf. Williams *et al.*[27]), because information removed from working memory can have a separate representation in long-term memory.

The evidence reviewed in the previous section about removal leading to the neural silencing of the active representations of those items, which can be restored if cued as relevant shortly after,[2,4,37] suggests that *temporary* removal does not abolish any information from memory, but reversibly alters the state of this information to improve goal-relevant processing. Specifically, removal may transform items from activation-based storage to weight-based (synaptic) storage.[45] There are two possible interpretations of this transformation.

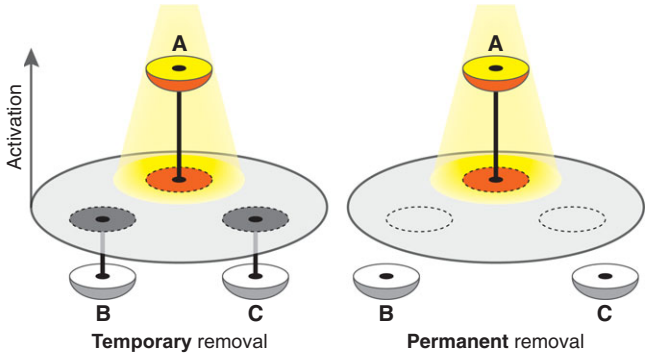

Figure 1. Removal from working memory. Diagrams depicting the hypothesized states of working memory representations following their temporary removal (left) and permanent removal (right). In both scenarios, item A is goal relevant and inside the focus of attention (golden halo). It is neurally active and bound (black line) to a specific context in working memory (orange disc). Items B and C are irrelevant and have been removed. They have become neurally inactive, and their context bindings either remain intact (temporary removal) or have been severed (permanent removal). Only temporarily removed items can be retrieved and reactivated by refocusing attention on their intact context bindings.

One is that the contents of working memory are outsourced into long-term memory (sometimes referred to as activated long-term memory[46–48]). This idea is supported by the observation of limitations that long-term memory imposes on removal from working memory: recent evidence suggests that prior knowledge and familiarity of items in working memory may impede the selective updating[49] or removal[50] of those items.

Another possibility is that information that is temporarily irrelevant is transferred into an unattended and hidden state of representation within working memory, maintained in the brain not by active neuronal firing but by short-term plasticity at the synapses of neurons coding for the information.[38,51–53] This hidden state contains items that have been removed from the focus of attention but are still in working memory.[52] This idea coincides with theoretical states of intermediate-term memory outside the focus of attention such as the region of direct access[54] or accessory memory.[55] This information can be reactivated into a neurally detectable (and cognitively attended) state through cues that leverage intact content-to-context bindings.

Temporary removal contrasts with a more permanent form of removal that acts to completely and irreversibly remove irrelevant information from working memory. Working memory contents become permanently irrelevant when participants learn that the information will never again be needed to support task performance. This can occur during working memory updating tasks in which existing items need to be modified or replaced[15] and in retrocueing tasks after the final cue in a given trial identifies that information that will no longer be needed.[1,9,20] When an item becomes permanently irrelevant, it neurally deactivates and no longer casts the same shadow on ongoing maintenance or task performance. That is, information that becomes completely irrelevant (and thus a candidate for permanent removal) is no longer neurally detectable with exogenous stimulation as is temporarily irrelevant information,[41] it is no longer behaviorally detectable with attentional biases on visual search tasks,[56] and it imposes less intrusion cost on memory performance as did temporarily removed information.[20,57]

The mechanisms supporting removal from working memory are not yet clear. We propose that temporary removal from working memory operates on content representations while sparing the context and the content-to-context bindings. One possible mechanism for this is the withdrawal of attention from an irrelevant representation in working memory, which leads to a temporary and *reversible* silencing of its neural representation[3,4] and of its influence on concurrent processing.[20,56] Using context cues to retrieve this content back into the focus of attention (thereby reinstating its neural and behavioral consequences) requires intact content-to-context bindings.

In contrast, we propose that permanent removal operates primarily on the binding between content and context, and in some situations, may operate on all three components. Permanent removal involves the active unbinding of contents from their contexts in working memory. This is an adaptive process that can reduce interference during subsequent encoding. For example, consider a situation in which item A (e.g., a word) is bound to context 1 (e.g., serial position 1) in working memory. When A becomes task-irrelevant, its activation may dissipate due to the withdrawal of attention (temporary removal), but its representation will remain intact and still bound to that context (Fig. 1). If item B is then bound to context 1, this will create a situation of cue overload[58] and the attempt to retrieve B (given context 1 as a retrieval cue) would be hindered by retrieval competition with A.[59] This highlights why permanent removal (unbinding an item from its context) can provide additional performance benefit over temporary removal. This form of removal is implemented in the SOB-CS model of working memory.[60] SOB-CS is a model of concurrent maintenance and processing in working memory tasks such as the complex-span paradigm,[61] in which encoding of memory items alternates with processing of distractors. According to SOB-CS, both items and distractors are encoded into working memory through temporary bindings of content representations to their contexts, that is, their current serial position in the list. These bindings are created through rapid Hebbian association learning. When processing of a distractor is finished, it is no longer needed, and therefore removed from working memory by a gradual unbinding process. This unbinding process is implemented as a process called Hebbian antilearning,[62] which does the opposite of Hebbian learning: It removes, rather than creates, item–context bindings. Note that this is a deliberate, goal-directed removal process affecting only goal-irrelevant information. As such, it is conceptually distinct from time-based decay that gradually weakens all item–context bindings in working memory. We address the challenge of distinguishing experimentally these alternative accounts in the concluding section of this article.

Much research has been devoted to the voluntary or involuntary inhibition of memory representations.[63,64] How does removal relate to inhibition? Inhibition is the deactivation of representations—sometimes below baseline—that makes accessing these representations harder. An important characteristic of inhibition in one prominent theory is that it affects the item itself, not its binding to a retrieval cue. Therefore, the effect of inhibition on the retrievability of a memory representation is cue independent.[65] As such, inhibition could play a role in temporary removal as a mechanism for rapidly silencing currently unneeded neural representations. At the same time, inhibition is different from the unbinding mechanism that we envision for permanent removal.

Attentional engagement during removal

To date, very little is known about the attentional demands of removal. As discussed above, if sustained attention is indeed required to support working memory,[66–68] then temporary removal may be a by-product of the withdrawal of attention from the remembered information. Alternative accounts,[53,69] however, argue against the necessity of sustained attention for working memory retention. According to this view, the removal of information would require something in addition to the mere withdrawal of attention. Instead, removal might require active unbinding of items from their context in working memory, as we have suggested may underlie the permanent form of removal. Such a process could be implemented through Hebbian antilearning, which requires an active representation of the content and the context whose binding needs to be untied.[60] Therefore, removal of content–context bindings would arguably entail attentional focus toward (not away from) the to-be-removed item and its context. To date, there is no evidence that removal involves attending to the to-be-removed information.

Another question related to the role of attention in removal is whether the removal process requires central attentional capacity, sometimes referred to as a central bottleneck. The assumption that removal requires central attention is supported by evidence showing that directed forgetting of an item involves an attention-demanding removal process.[70,71] One attempt to test whether removal of irrelevant information from working memory relies on the central bottleneck resulted in ambiguous evidence.[11] We conclude that the available evidence is insufficient to determine whether or not removal relies on central attention.

The time course of removal

Early experiments on removal of irrelevant information from working memory pointed to a gradual removal process that takes about 1–2 s to completely remove a set of three items.[1] This time course was corroborated by a study tracking the EEG correlate of the to-be-removed information over time.[4] A gradual removal process at about this rate is also assumed in the SOB-CS model.[60] The assumption of gradual removal of distractors over about 1–2 s explains why memory performance in complex-span tasks improves with increasing free time following each distractor,[72] although a recent complex-span study suggests that distractor removal might be much slower than assumed in SOB-CS so far.[73]

More recent experiments point to different time courses of removal under different conditions.[21] In these experiments, participants initially encoded six words, of which three were subsequently cued to be irrelevant. Memory was tested by a recognition probe, to be compared to one specific word in the remaining relevant subset. The effectiveness of removing the irrelevant words was measured by comparing recognition RTs and accuracies in the removal condition to two baselines, one in which all six words had to be remembered, and one in which only three words were encoded and remembered. When the cue identifying the relevant words was presented after encoding of the entire list, recognition RTs (but not accuracies) in the removal condition gradually approached those in the set-size 3 baseline as the interval between cue and probe increased from 0.1 to 1.5 seconds. This finding suggests that the irrelevant subset was removed from working memory gradually, in line with Oberauer.[1] In contrast, when presentation of each word was immediately followed by a cue telling the participant whether or not to remember it, RTs and accuracies in the removal condition were equivalent to the set-size 3 baseline regardless of the postcue time, as if the to-be-forgotten words had never existed. Yet, these words must have been encoded into memory at least briefly because the forget cues were fully effective even when presented 1 s after offset of the preceding word. This brief maintenance is presumably accomplished by working memory (cf. Kessler[74] for an alternative explanation involving direct encoding into long-term memory). If the individual words are initially maintained in working memory, it follows that they can be removed very rapidly and effectively from working memory right after encoding, before any further processes intervene. Very fast removal of a just-encoded representation challenges the SOB-CS model, because it would imply that each distractor representation can be removed from working memory very rapidly once it is no longer needed for the processing task. If that is the case, slow and gradual removal of distractors cannot explain the beneficial effect of longer free time in between distractors.

The two observed time courses of removal suggest two kinds of removal, reminiscent of the distinction between item-wise- and list-wise-directed forgetting in the long-term memory literature.[75] Item-wise removal, applied right after encoding of a stimulus, is very fast and improves both speed and accuracy of access to the remaining contents of working memory. In contrast, subset-wise removal, applied to one of several subsets after encoding the entire memory set, is slower, taking about 1–2 s, and primarily improves speed of access to the remaining working memory contents. One possible explanation of these different time courses is in terms of the distinction proposed above: temporary versus permanent removal. Item-wise removal, immediately after encoding a stimulus, reflects the complete and permanent removal of the new working memory content, including its temporary bindings to its context, as implemented in SOB-CS. In contrast, subset-wise removal consists of the—potentially temporary—cessation of persistent neural activation of content representations,[2–4] leaving content–context bindings intact. Subset-wise removal of content–context bindings may be possible but difficult after several other processes have intervened between encoding and removal (see the next section, and Oberauer[21]).

Limitations on removal

Under some circumstances, removal is a slow and perhaps laborious process, suggesting that it is not efficiently applied in all situations. Here, we discuss under which circumstances, and why, removal is limited. The effectiveness of (permanent) removal of item–context bindings is limited by its prerequisites: To remove the item–context bindings of one or several items within a memory set, these bindings must first be selectively accessed. This can be

difficult, sometimes so difficult that removal does not take place at all. The first evidence for this difficulty comes from a study by Ecker *et al.*[15] They used a working memory updating paradigm in which the to-be-updated subset of the items was precued in each trial prior to the presentation of the new items, enabling the participants to remove the outdated items in advance. Indeed, when one item (out of a set of three) was precued for removal, updating times were faster when the new items were presented. Preemptive removal of the old item produced an encoding benefit for the new one. This benefit, however, did not scale with the number of items precued for removal: The updating times were not any faster when two (of the three) items were cued. This suggests some form of partial removal: Either both cued items were each only partially removed before the new items appeared, or only one of the two cued items was selected and fully removed. A possible reason for the removal of only one of two items is the time-consuming nature of serially scanning throughout the list and switching between removed and maintained items.[18,76] Performance was the fastest when the entire current memory set was updated (see also Kessler and Meiran[17]). In this condition, little benefit for precuing was observed, suggesting that removing the entire set is not done sequentially, item by item, but rather in a rapid "wipe out" manner.

Another series of experiments also revealed circumstances under which removal is difficult.[21] Participants were asked to encode six words into working memory and then three of them were indicated to be irrelevant, and thus could be removed. When the three to-be-removed words were selected at random, people did not remove them at all. Only when the words formed an already predefined subset at encoding was there any evidence for removal. Removal of a random subset of three out of six items might be forbiddingly difficult for the same reasons as those identified with the updating task: participants would have to scan the entire set of six words, switching back and forth between maintenance and removal, and this might be too time consuming or challenging to allow for efficient removal.

In addition to the above time constraints, the strength of removal is also limited in its precision. A perfect removal process would completely remove items from working memory, without leaving residues and without oversuppressing them. However, evidence for both types of imperfect removal strength is available. For example, recent neural evidence suggests that cueing the relevant dimension (e.g., phonological, semantic, or visual) of a single item in working memory results in a biasing[77] of neural representation toward the relevant dimension while representations of the irrelevant dimensions are reduced but not completely removed.[50] Behaviorally, item-position repetition benefits are observed in the working memory updating paradigm[78] (see also Lendínez *et al.*[29] for a generalization to semantic similarity). This benefit is reduced, but not eliminated, following a removal cue,[14] showing that removal was incomplete. Based on neural readouts from fMRI data, removal success can be highly variable from trial to trial, and incomplete removal of irrelevant items has been linked to the subsequent forgetting of those items (versus items that were completely removed) in recognition tests of long-term memory.[79]

Conversely, evidence for excess removal comes from repetition costs that have been observed using the reference-back paradigm.[80] Kessler[74] recently demonstrated n-2 repetition costs (backward inhibition[81]) within a series of working memory updating steps, but not within a series of trials that did not involve updating. Such costs are typically observed in the task switching domain (cf. Gade *et al.*[82] for a demonstration in declarative working memory). They provide evidence for an overshooting of removal. Namely, changing the item that is associated with a specific context (e.g., location) from item A to B involves removal of the binding between A and that context. When the effect of removal lingers during the following trials, associating A again in that context is harder, and hence takes more time, than associating another item. The finding of n-2 repetition costs in working memory updating, implicating overly strong removal, is the opposite of the above evidence for a reduced (but not abolished) repetition gain following a removal cue. Notably, in Kessler's experiments only one item had to be stored in working memory, while the study of Ecker *et al.*[14] required the storage of three items. Paradoxically, the single-item context might be conceived as harder due to a higher cue overload compared to the 3-item context, leading participants to apply removal more strongly in that context (see Jost *et al.*[83] for an analogous example of variable

inhibition in task switching). Further empirical work is required to examine this idea.

What counts as evidence against removal?

We reviewed evidence (see above) that (1) removal facilitates processing of the remaining contents of working memory, and that (2) the removed contents become less accessible and (3) neurally silent. Now we consider two kinds of evidence against removal. First, any evidence that supports an alternative explanation of these three kinds of findings strengthens the case against removal. Second, failure to find these three effects when removal is predicted to happen also counts as evidence against removal.

One potential alternative explanation of the empirical signatures of removal is time-based decay, together with selective maintenance of the relevant information (e.g., through rehearsal). For this explanation to work, decay would have to occur very rapidly, eliminating working memory representations within 1–2 s, or even faster (see above). This is an unlikely assumption, in particular for verbal materials, for which the balance of evidence implies that they do not decay at all[84,85] (but see Ricker *et al.*[86] for a different view). One way to adjudicate between decay and removal is to combine a short-term directed-forgetting paradigm[87] with a secondary task that impedes rehearsal and refreshing. For instance, after reading a short list of words, participants might receive a remember or forget cue, followed by a brief period that is either unfilled or filled with a secondary task. After that, a second list of words is encoded that always needs to be remembered. The decay assumption predicts a beneficial effect of the secondary task in the remember condition: preventing rehearsal should allow more decay, resulting in a reduced load on working memory from the first list. As a consequence, memory for the second list should be better. Conversely, the removal assumption predicts a detrimental effect of the secondary task in the forget condition: preventing removal should leave a larger load on working memory from the first list, and thereby lead to worse memory for the second list.

A variant of this alternative explanation that does not require decay states that, rather than removing irrelevant information, the working memory system strengthens relevant information. As a consequence, the irrelevant information, although still in working memory, becomes less accessible due to the much stronger competition from relevant information. This explanation faces empirical and conceptual problems. Empirically, it does not agree with the finding that the decodability of irrelevant information from neural signals drops to baseline.[2,4,41] It also cannot account for evidence of successful directed forgetting in the absence of any to-be-remembered information in working memory,[88] and it cannot explain why, after cueing all but one item of a first memory set as irrelevant, a subsequently encoded second memory set is remembered better.[23] This should not happen in a limited-capacity memory system if all that happens in response to a cue is strengthening of the cued item. Conceptually, this explanation implies that with every shift of relevance the overall strength of memory traces in working memory increases. Moreover, every new content encoded into working memory—for instance, in an updating task—adds further memory traces that must have more strength than the previous ones they are to replace. Over a lifetime, this would increase the strength of working memory representations to dizzying heights.

Another approach to find evidence against removal could be to demonstrate that, after an opportunity to remove irrelevant material from working memory, accessibility of that material is undiminished. For this test to be informative, care must be taken that the test of accessibility actually measures the information that is assumed to be removed. A recent study by Dagry and Barrouillet[89] illustrates the potential problems in this endeavor. These authors set out to test the assumption in the SOB-CS model[60] that, in a complex-span task,[61] memory traces of distractors are removed from working memory. In SOB-CS, distractors are inadvertently encoded by binding them to the list positions of adjacent memory items, and they are removed from working memory by unbinding them from these list positions. The prediction to be tested, therefore, is that distractor–position bindings are less accessible after an opportunity to remove them. This could be accomplished, for instance, by asking participants (perhaps on a tiny subset of trials, as in Muter[25]) to recall the distractors in serial order, or to recall the distractors following a given memory item. Dagry and Barrouillet,[89] however, tested memory for distractors independent of their bindings to list positions through a free-recall test and through short-term repetition priming. Both

forms of memory test gauge the strength of some form of memory about the recent occurrence of the distractor stimuli, not about their bindings to specific list positions. On the assumption that short-term repetition priming[90] reflects the strength of persistent neural activation of a recently processed stimulus, the strength of short-term repetition priming could be used to gauge the (temporary) removal of content representations—but not the (permanent) removal of content–context bindings.

Summary

We reviewed evidence for removal as a distinct control process that excludes irrelevant information from working memory in service of goal-directed behavior. Removal can be temporary, operating only on the contents of working memory without impacting content–context bindings necessary for reversing this process. Or it can be permanent, wiping out the content and contextual bindings of outdated information. Removal can be proactive to facilitate new encoding, and it can be reactive to improve access to subsets of relevant information already in working memory. The speed of removal depends on whether an item or a subset of items is being removed, ranging from a very fast and effortless (permanent) removal of a just-encoded item, to a more gradual, laborious (temporary) removal of a subset of items. The process may be applied too weakly, leading to partial removal, or too strongly, leading to lingering inhibition, both of which can impose behavioral costs. The mechanisms supporting removal are currently underspecified, and this review is meant to motivate future research on key unresolved issues that will help us understand how the brain reduces, reuses, and recycles information.

Acknowledgments

This research was funded in part by a grant from the National Institute of Mental Health (1R21MH108848-01A1) to J.A. Lewis-Peacock, a grant from the Israel Science Foundation (#458/14) to Y. Kessler, and a grant from the Swiss National Science Foundation to K. Oberauer (project 149193).

Competing interests

The authors declare no competing interests.

References

1. Oberauer, K. 2001. Removing irrelevant information from working memory: a cognitive aging study with the modified Sternberg task. *J. Exp. Psychol. Learn. Mem. Cogn.* **27:** 948–957.
2. Lewis-Peacock, J.A., A.T. Drysdale, K. Oberauer, *et al.* 2012. Neural evidence for a distinction between short-term memory and the focus of attention. *J. Cogn. Neurosci.* **24:** 61–79.
3. Lewis-Peacock, J.A. & B.R. Postle. 2012. Decoding the internal focus of attention. *Neuropsychologia* **50:** 470–478.
4. LaRocque, J.J., J.A. Lewis-Peacock, A.T. Drysdale, *et al.* 2013. Decoding attended information in short-term memory: an EEG study. *J. Cogn. Neurosci.* **25:** 127–142.
5. Mueller, S.T., T.L. Seymour, D.E. Kieras, *et al.* 2003. Theoretical implications of articulatory duration, phonological similarity, and phonological complexity in verbal working memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **29:** 1353–1380.
6. Oberauer, K. & S. Lewandowsky. 2008. Forgetting in immediate serial recall: decay, temporal distinctiveness, or interference? *Psychol. Rev.* **115:** 544–576.
7. May, C.P., L. Hasher & M.J. Kane. 1999. The role of interference in memory span. *Mem. Cognit.* **27:** 759–767.
8. Souza, A.S. & K. Oberauer. 2016. In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Atten. Percept. Psychophys.* **78:** 1839–1860.
9. Griffin, I.C. & A.C. Nobre. 2003. Orienting attention to locations in internal representations. *J. Cogn. Neurosci.* **15:** 1176–1194.
10. Cowan, N. & C.C. Morey. 2007. How can dual-task working memory retention limits be investigated? *Psychol. Sci.* **18:** 686–688.
11. Oberauer, K. & S. Lewandowsky. 2016. Control of information in working memory: encoding and removal of distractors in the complex-span paradigm. *Cognition* **156:** 106–128.
12. Nee, D.E. & J. Jonides. 2009. Common and distinct neural correlates of perceptual and memorial selection. *NeuroImage* **45:** 963–975.
13. Braver, T.S., J.R. Gray & G.C. Burgess. 2008. Explaining the many varieties of working memory variation: dual mechanisms of cognitive control. In *Variation in Working Memory*. A. Conway, C. Jarrold, M. Kane, *et al.*, Eds.: 76–106. Oxford University Press.
14. Ecker, U.K.H., S. Lewandowsky & K. Oberauer. 2014. Removal of information from working memory: a specific updating process. *J. Mem. Lang.* **74:** 77–90.
15. Ecker, U.K., K. Oberauer & S. Lewandowsky. 2014. Working memory updating involves item-specific removal. *J. Mem. Lang.* **74:** 1–15.
16. Kessler, Y. & N. Meiran. 2006. All updateable objects in working memory are updated whenever any of them are modified: evidence from the memory updating paradigm. *J. Exp. Psychol. Learn. Mem. Cogn.* **32:** 570–585.
17. Kessler, Y. & N. Meiran. 2008. Two dissociable updating processes in working memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **34:** 1339–1348.
18. Kessler, Y. & K. Oberauer. 2014. Working memory updating latency reflects the cost of switching between maintenance and updating modes of operation. *J. Exp. Psychol. Learn. Mem. Cogn.* **40:** 738–754.

19. Banich, M.T., K.L. Mackiewicz Seghete, B.E. Depue, *et al.* 2015. Multiple modes of clearing one's mind of current thoughts: overlapping and distinct neural systems. *Neuropsychologia* **69:** 105–117.
20. Oberauer, K. 2005. Control of the contents of working memory—a comparison of two paradigms and two age groups. *J. Exp. Psychol. Learn. Mem. Cogn.* **31:** 714–728.
21. Oberauer, K. 2018. Removal of irrelevant information from working memory: sometimes fast, sometimes slow, and sometimes not at all. *Ann. N.Y. Acad. Sci.* **XXXX:** XX–XX.
22. Shepherdson, P., K. Oberauer & A.S. Souza. 2018. Working memory load and the retro-cue effect: a diffusion model account. *J. Exp. Psychol. Hum. Percept. Perform.* **44:** 286–310.
23. Souza, A.S., L. Rerko & K. Oberauer. 2014. Unloading and reloading working memory: attending to one item frees capacity. *J. Exp. Psychol. Hum. Percept. Perform.* **40:** 1237–1256.
24. Yntema, D.B. & G.E. Mueser. 1962. Keeping track of variables that have few or many states. *J. Exp. Psychol.* **63:** 391–395.
25. Muter, P. 1980. Very rapid forgetting. *Mem. Cognit.* **8:** 174–179.
26. Marsh, R.L., M.M. Sebrechts, J.L. Hicks, *et al.* 1997. Processing strategies and secondary memory in very rapid forgetting. *Mem. Cognit.* **25:** 173–181.
27. Williams, M., S.W. Hong, M.-S.S. Kang, *et al.* 2013. The benefit of forgetting. *Psychon. Bull. Rev.* **20:** 348–355.
28. Festini, S.B. & P.A. Reuter-Lorenz. 2013. The short- and long-term consequences of directed forgetting in a working memory task. *Memory* **21:** 763–777.
29. Lendínez, C., S. Pelegrina & T. Lechuga. 2011. The distance effect in numerical memory-updating tasks. *Mem. Cognit.* **39:** 675–685.
30. Shenhav, A., S. Musslick, F. Lieder, *et al.* 2017. Toward a rational and mechanistic account of mental effort. *Annu. Rev. Neurosci.* **40:** 99–124.
31. Chang, E.P., U.K.H. Ecker & A.C. Page. 2017. Impaired memory updating associated with impaired recall of negative words in dysphoric rumination—evidence for a removal deficit. *Behav. Res. Ther.* **93:** 22–28.
32. Joormann, J. & I.H. Gotlib. 2008. Updating the contents of working memory in depression: interference from irrelevant negative material. *J. Abnorm. Psychol.* **117:** 182–192.
33. Yoon, K.L., J. LeMoult & J. Joormann. 2014. Updating emotional content in working memory: a depression-specific deficit? *J. Behav. Ther. Exp. Psychiatry* **45:** 368–374.
34. Zetsche, U., C. D'Avanzato & J. Joormann. 2012. Depression and rumination: relation to components of inhibition. *Cogn. Emot.* **26:** 758–767.
35. Lewis-Peacock, J.A. & K.A. Norman. 2014. Multivoxel pattern analysis of functional MRI data. In *The Cognitive Neurosciences.* 5th ed. M.S. Gazzaniga & G.R. Mangun, Eds.: 911–920. MIT Press.
36. Haxby, J.V., A.C. Connolly & J.S. Guntupalli. 2014. Decoding neural representational spaces using multivariate pattern analysis. *Annu. Rev. Neurosci.* **37:** 435–456.
37. Sprague, T.C., E.F. Ester & J.T. Serences. 2016. Restoring latent visual working memory representations in human cortex. *Neuron* **91:** 694–707.
38. Stokes, M.G. 2015. "Activity–silent" working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn. Sci.* **19:** 394–405.
39. Schneegans, S. & P.M. Bays. 2017. Restoration of fMRI decodability does not imply latent working memory states. *J. Cogn. Neurosci.* **29:** 1977–1994.
40. Christophel, T.B., P. Iamshchinina, C. Yan, *et al.* 2018. Cortical specialization for attended versus unattended working memory. *Nat. Neurosci.* **21:** 494–496.
41. Rose, N.S., J.J. LaRocque, A.C. Riggall, *et al.* 2016. Reactivation of latent working memories with transcranial magnetic stimulation. *Science* **354:** 1136–1139.
42. Wolff, M.J., J. Jochim, E.G. Akyürek, *et al.* 2017. Dynamic hidden states underlying working-memory-guided behavior. *Nat. Neurosci.* **20:** 864–871.
43. Burgess, N. & G.J. Hitch. 2006. A revised model of short-term memory and long-term learning of verbal sequences. *J. Mem. Lang.* **55:** 627–652.
44. Lewandowsky, S. & S. Farrell. 2008. Short-term memory: new data and a model. *Psychol. Learn. Motiv. Adv. Res. Theory* **49:** 1–48.
45. O'Reilly, R.C. & Y. Munakata. 2000. *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain.* MIT Press.
46. Anderson, J.R. 1983. *The Architecture of Cognition.* Cambridge, MA: Harvard University Press.
47. Cowan, N. 1995. *Attention and Memory: An Integrated Framework.* New York: Oxford University Press.
48. Ruchkin, D.S., J. Grafman, K. Cameron, *et al.* 2003. Working memory retention systems: a state of activated long-term memory. *Behav. Brain Sci.* **26:** 709–777.
49. Artuso, C. & P. Palladino. 2017. How sublexical association strength modulates updating: cognitive and strategic effects. *Mem. Cognit.* **46:** 285–297.
50. Lewis-Peacock, J.A., A.T. Drysdale & B.R. Postle. 2015. Neural evidence for the flexible control of mental representations. *Cereb. Cortex* **25:** 3303–3313.
51. Mongillo, G., O. Barak & M. Tsodyks. 2008. Synaptic theory of working memory. *Science* **319:** 1543–1546.
52. Postle, B.R. 2016. How does the brain keep information "in mind"? *Curr. Dir. Psychol. Sci.* **25:** 151–156.
53. Sreenivasan, K.K., C.E. Curtis & M. D'Esposito. 2014. Revisiting the role of persistent neural activity during working memory. *Trends Cogn. Sci.* **18:** 82–89.
54. Oberauer, K. 2002. Access to information in working memory: exploring the focus of attention. *J. Exp. Psychol. Learn. Mem. Cogn.* **28:** 411–421.
55. Olivers, C.N.L., J. Peters, R. Houtkamp, *et al.* 2011. Different states in visual working memory: when it guides attention and when it does not. *Trends Cogn. Sci.* **15:** 327–334.
56. Mallett, R. & J. Lewis-Peacock. 2018. Behavioral decoding of working memory items inside and outside the focus of attention. *Ann. N.Y. Acad. Sci.* **XXXX:** XX–XX.
57. Festini, S.B. & P.A. Reuter-Lorenz. 2014. Cognitive control of familiarity: directed forgetting reduces proactive interference in working memory. *Cogn. Affect. Behav. Neurosci.* **14:** 78–89.

58. Watkins, O.C. & M.J. Watkins. 1975. Buildup of proactive inhibition as a cue-overload effect. *J. Exp. Psychol. Hum. Learn. Mem.* **1:** 442–452.
59. Unsworth, N. & R.W. Engle. 2007. The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychol. Rev.* **114:** 104–132.
60. Oberauer, K., S. Lewandowsky, S. Farrell, *et al.* 2012. Modeling working memory: an interference model of complex span. *Psychon. Bull. Rev.* **19:** 779–819.
61. Daneman, M. & P.A. Carpenter. 1980. Individual differences in working memory and reading. *J. Verbal Learn. Verbal Behav.* **19:** 450–466.
62. Koch, G., V. Ponzo, F.D. Lorenzo, *et al.* 2013. Hebbian and Anti-Hebbian spike-timing-dependent plasticity of human cortico-cortical connections. *J. Neurosci.* **33:** 9725–9733.
63. Anderson, M.C. & S. Hanslmayr. 2014. Neural mechanisms of motivated forgetting. *Trends Cogn. Sci.* **18:** 279–292.
64. Jonides, J., E.E. Smith, C. Marshuetz, *et al.* 1998. Inhibition of verbal working memory revealed by brain activation. *Proc. Natl. Acad. Sci. USA* **95:** 8410–8413.
65. Anderson, M. 2003. Rethinking interference theory: executive control and the mechanisms of forgetting. *J. Mem. Lang.* **49:** 415–445.
66. Jonides, J., R.L. Lewis, D.E. Nee, *et al.* 2008. The mind and brain of short-term memory. *Annu. Rev. Psychol.* **59:** 193–224.
67. Chun, M.M. 2011. Visual working memory as visual attention sustained internally over time. *Neuropsychologia* **49:** 1407–1409.
68. Chun, M.M., J.D. Golomb & N.B. Turk-Browne. 2011. A taxonomy of external and internal attention. *Annu. Rev. Psychol.* **62:** 73–101.
69. Myers, N.E., M.G. Stokes & A.C. Nobre. 2017. Prioritizing information during working memory: beyond sustained internal attention. *Trends Cogn. Sci.* **21:** 449–461.
70. Fawcett, J.M. & T.L. Taylor. 2008. Forgetting is effortful: evidence from reaction time probes in an item-method directed forgetting task. *Mem. Cognit.* **36:** 1168–1181.
71. Fawcett, J.M. & T.L. Taylor. 2012. The control of working memory resources in intentional forgetting: evidence from incidental probe word recognition. *Acta Psychol. (Amst).* **139:** 84–90.
72. Barrouillet, P., S. Bernardin, S. Portrat, *et al.* 2007. Time and cognitive load in working memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **33:** 570–585.
73. Hoareau, V., S. Portrat, K. Oberauer, *et al.* 2017. Computational and behavioral investigations of the SOB-CS removal mechanism in working memory. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*. G. Gunzelmann, A. Howes, T. Tenbrink, *et al.*, Eds.: 532–537. London: Cognitve Science Society.
74. Kessler, Y. 2018. *N*-2 repetition leads to a cost within working memory and a benefit outside it: specifying the role of removal in memory updating. *Ann. N.Y. Acad. Sci.* **XXXX:** XX–XX.
75. Macleod, C.M. 1999. The item and list methods of directed forgetting: test differences and the role of demand characteristics. *Psychon. Bull. Rev.* **6:** 123–129.
76. Kessler, Y. & K. Oberauer. 2015. Forward scanning in verbal working memory updating. *Psychon. Bull. Rev.* **22:** 1770–1776.
77. Kuo, B.C., M.G. Stokes & A.C. Nobre. 2012. Attention modulates maintenance of representations in visual short-term memory. *J. Cogn. Neurosci.* **24:** 51–60.
78. Oberauer, K. & K. Vockenberg. 2009. Updating of working memory: lingering bindings. *Q. J. Exp. Psychol.* **62:** 967–987.
79. Lewis-Peacock, J.A. & K.A. Norman. 2014. Competition between items in working memory leads to forgetting. *Nat. Commun.* **5:** 5768.
80. Rac-Lubashevsky, R. & Y. Kessler. 2016. Dissociating working memory updating and automatic updating: the reference-back paradigm. *J. Exp. Psychol. Learn. Mem. Cogn.* **42:** 951–969.
81. Mayr, U. & S.W. Keele. 2000. Changing internal constraints on action: the role of backward inhibition. *J. Exp. Psychol. Gen.* **129:** 4–26.
82. Gade, M., A.S. Souza, M.D. Druey, *et al.* 2017. Analogous selection processes in declarative and procedural working memory: N-2 list-repetition and task-repetition costs. *Mem. Cognit.* **45:** 26–39.
83. Jost, K., V. Hennecke & I. Koch. 2017. Task dominance determines backward inhibition in task switching. *Front. Psychol.* **8:** 755.
84. Oberauer, K. & S. Lewandowsky. 2013. Evidence against decay in verbal working memory. *J. Exp. Psychol. Gen.* **142:** 380–411.
85. Oberauer, K. & S. Lewandowsky. 2014. Further evidence against decay in working memory. *J. Mem. Lang.* **73:** 15–30.
86. Ricker, T.J., E. Vergauwe & N. Cowan. 2016. Decay theory of immediate memory: from Brown (1958) to today (2014). *Q. J. Exp. Psychol.* **69:** 1969–1995.
87. Tehan, G. & M.S. Humphreys. 1998. Creating proactive interference in immediate recall: building a DOG from a DART, a mop, and a FIG. *Mem. Cognit.* **26:** 477–489.
88. Festini, S.B. & P.A. Reuter-Lorenz. 2017. Rehearsal of to-be-remembered items is unnecessary to perform directed forgetting within working memory: support for an active control mechanism. *J. Exp. Psychol. Learn. Mem. Cogn.* **43:** 94–108.
89. Dagry, I. & P. Barrouillet. 2017. The fate of distractors in working memory: no evidence for their active removal. *Cognition* **169:** 129–138.
90. McKone, E. 1998. The decay of short-term implicit memory: unpacking lag. *Mem. Cognit.* **26:** 1173–1186.