Trends in Cognitive Sciences

CellPress

Opinion

Prioritizing Information during Working Memory: Beyond Sustained Internal Attention

Nicholas E. Myers,[1,2,*] Mark G. Stokes,[1,2] and Anna C. Nobre[1,2,*]

Working memory (WM) has limited capacity. This leaves attention with the important role of allowing into storage only the most relevant information. It is increasingly evident that attention is equally crucial for prioritizing representations within WM as the importance of individual items changes. Retrospective prioritization has been proposed to result from a focus of internal attention highlighting one of several representations. Here, we suggest an updated model, in which prioritization acts in multiple steps: first orienting towards and selecting a memory, and then reconfiguring its representational state in the service of upcoming task demands. Reconfiguration sets up an optimized perception–action mapping, obviating the need for sustained attention. This view is consistent with recent literature, makes testable predictions, and links WM with task switching and action preparation.

Trends

Recent research has uncovered our remarkable flexibility in prioritizing information in WM, refining the concept of multiple representational states in WM.

Neuroimaging studies have investigated the networks controlling prioritization in WM.

Prioritization activates prefrontal and parietal brain areas associated with the deployment of visual attention, suggesting a parallel between attention to external stimuli and attention to memory contents ('internal attention').

However, additional prefrontal areas are specific to WM prioritization. We propose that they reflect the recruitment of high-priority information for the next action. What can this tell us about the neural basis of different representational states in WM? We speculate that prioritized information is reflected in the task-specific tuning of a neural network important for action selection and preparation.

The Changing Concept of Priority in Working Memory

The subject of this article is the neural basis and behavioral consequence of prioritizing information maintained in visual short-term, 'working' memory (WM). By WM in this context, we refer to the ability to store and manipulate recently acquired information for a period of seconds, independently of continuous sensory stimulation, to guide behavior over the short-term. This ability is central to intelligent behavior [1,2] and, therefore, touches on nearly all domains of cognitive neuroscience (such as fluid intelligence, perceptual decision-making, or model-based learning; e.g., [3–5]). The severe limits to how much can be encoded in WM (conceived as a small number of quantized representations [6–8] or as a limited pool of mnemonic resources [9]) hamper our ability to act optimally when there is too much information to be considered at once. As a consequence of this bottleneck, attention is of central importance to WM [10–15]: Those who cannot select the most important information and keep out irrelevant distraction unnecessarily clutter their WM store [16,17].

Early studies exploring the role of attention in WM manipulated selective encoding (i.e., prioritizing a subset of items during encoding). Later, studies revealed that focusing on the relevant pieces of information even after they have already been encoded also improves memory [18–20]. Such retrospective cueing cannot influence basic sensory processing of the memory items, or their encoding, but rather operates at a pure mnemonic level, prioritizing the contents maintained in WM.

Neurodevelopmental (e.g., [21]) and psychiatric disorders (e.g., [22]), as well as healthy aging [23,24], severely affect WM capacity, making it imperative that we better understand how prioritization within WM can help us make the most of a preciously limited cognitive resource.

[1]Oxford Centre for Human Brain Activity, Oxford University, Oxford, UK
[2]Department of Experimental Psychology, Oxford University, Oxford, UK

*Correspondence:
nicholas.myers@ohba.ox.ac.uk (N.E. Myers) and
kia.nobre@ohba.ox.ac.uk (A.C. Nobre).

Here, we focus on new empirical and theoretical advances that shed light on the role of prioritization in WM, and how this may relate to task preparation. In synthesizing this literature, we suggest that both attentional selection and task preparation have critical roles in prioritizing information in WM to guide optimal performance.

We begin by drawing parallels with the better-understood mechanisms of selective attention for perception. We then build on this model with the aim of explaining more fully how prioritization may operate in WM, and within internal information stores more generally. We propose that, in addition to any benefits brought about by attentional selection of individual items, behavioral benefits also arise in large part because of preparation of the right behavioral policy (for instance, by setting up appropriate contingencies between upcoming stimuli and behavior). Our proposal can account for several otherwise odd findings in the behavioral literature. Moreover, it may help pin down the dual roles of selection and preparation in prioritizing information in mind. Furthermore, our model makes predictions about the possible neural basis of the architecture of WM.

Attention in Perception and Working Memory

WM is famously burdened with severe capacity limits. As in many other domains of cognition that contain a bottleneck [25,26], the preferential selection of pertinent information appears crucial if we are to make the best use of our limited resource. In the domain of perception, the term ‘selective attention’ is invoked to describe such preferential biases towards behaviorally relevant stimuli. In extending this literature, attention has been shown to be influential for selecting information for encoding into WM [27–29], and for preventing distracting information from gaining access to it [30]. The benefits of attention are generally assumed to follow the biased competition principle [31]: gains in processing (e.g., [32–34]) for an attended location or feature are achieved by biasing the receptive fields of neurons in their favor, at the expense of unattended locations or features.

Without question, attention before or during encoding has high utility for behavior. However, the relevance of stimuli is not always obvious while they are still present; sometimes, we need to prioritize information that has already been encoded in WM. For example, you may be looking around your apartment for your car keys and your phone simultaneously, holding templates of both in your WM as you scan your surroundings. Suddenly the phone starts ringing, so you prioritize finding the phone first to get to it in time. This ability had already been noted by William James in his endlessly cited definition of attention as the ‘taking possession by the mind [. . .] of one out of what seem several simultaneously possible objects or trains of thought’ ([35], pp. 403–404). The ability to manipulate WM content flexibly is also a hallmark of classic definitions of WM [1]. As with selective prioritization before and during encoding, prioritizing important items in WM during the retention interval has been shown to lead to a substantial memory boost [18–20]. Experimentally, prioritization within WM is generally induced by presenting a cue during the retention interval that directs focus to one of the items already held in mind. Cues can refresh a previously presented item [36,37], bring a subset of items currently in WM into the **focus of attention** (see Glossary) for immediate recall [38], or retroactively indicate that one item is most likely to be probed at the end of the delay interval. The latter is often referred to as a ‘**retrocue**’ (as opposed to a precue presented before WM encoding, see [19], or a postcue presented together with the probe).

At first glance, the benefit of retrocueing appears paradoxical: memory is seemingly improved out of thin air. After all, the relevant information has already been stored in the brain, so how could providing an orienting cue possibly improve the strength of this information after the fact? Indeed, over 10 years of investigation into the behavioral correlates and neural mechanisms of prioritization in WM have not yielded a conclusive explanation. Most proposals draw parallels

Glossary

Focus of attention: specialized state within WM. As opposed to items that are merely maintained, the single item in the focus of attention [20] is selected and elevated to a separate representational state so that it can be updated, manipulated, or recalled. Representations in the focus of attention are recalled more quickly and with greater accuracy than other WM representations.

Internal attention: goal-directed selection of information that is not currently presented in the environment, such as long- or short-term memory or goals. Internal attention is thought to draw on the same selection mechanism that is deployed to attend to information arriving from the environment.

Latent storage: proposed neurophysiological mechanism for the neural storage of WM memoranda by reconfiguring the state of a memory network through short-term changes in its pattern of connections. After reconfiguration, persistent spiking is no longer necessary because the memory is stored in a latent state, for example in temporarily changed synaptic weights.

Output gating: some computational models of WM emphasize the importance of an input gate that determines which pieces of sensory information are allowed into the limited-capacity WM store. Similarly, more recent computational models propose a second gate determining which of the items that are stored in WM are permitted to drive behavior, or ‘output’. ‘Output-gating’ an item could correspond to moving it into the focus of attention, although the exact relationship is unclear.

Retrocue: a cue presented retrospectively, during the retention interval of a WM task, indicating that a subset of all items already held in memory is most relevant to behavior, for example because it is most likely to be probed.

Task switching: switching tasks (rule-guided responses to a limited set of stimuli) incurs costs in terms of slower reaction times and increased error rates. Switch costs occur because of the sudden need to reconfigure a task set in response-guiding brain networks. Cueing a task switch in advance reduces but does not eliminate switching costs.

between the effects of retrocueing and selective attention to external stimuli [9,13–15,39]. The same cognitive and neural mechanisms (selective attention) are deployed in each case, with the main difference being the substrate on which they operate, yielding a distinction between external and **internal attention**. In sum, these models emphasize that retrocueing benefits depend on a sustained bias of selective attention toward cued locations or features during a memory delay.

Overlap between External and Internal Attention

At a basic level, the behavioral effects of retrocues indeed appear to be similar to the effects of external attention. Responses when probing cued items are faster and more accurate [18,19,40–42]. On invalid cue trials, responses are often slower and less accurate ([40,43–45], although invalidity costs are not consistent across studies, as discussed in the next section). When cue validity is manipulated, more reliable cues lead to a larger benefit [46–48].

At the neural level, similar top-down attention networks are engaged for internal and external attention shifts (such as the frontal eye fields and the superior parietal lobe, [49–59]). Neurophysiological markers of attention shifts, such as desynchronization of alpha-band oscillations in the hemisphere contralateral to where the cued item was presented, appear also to be roughly comparable between external attention [60] and retrocueing [59,61–64]. In parallel, retrocues appear to reduce load-dependent signals, such as the contralateral delay activity [65–67], as if they allowed the removal of uncued items from the memory store. This removal mechanism is reminiscent of the filtering of distractors during encoding [16]. In tandem with the top-down control signals, activity is also modulated in sensory brain regions corresponding to the cued location or feature [57,68–73], which likely contribute to WM representations or their manipulation through sensory recruitment [74–79]. For example, when a visual stimulus category (e.g., faces) is cued, BOLD activity in the corresponding brain area (fusiform gyrus) increases [68,80,81]. This has been interpreted as increased processing of, or focused attention towards, the cued category. However, in many cases, this increase may also reflect anticipation of a probe stimulus from that category at the end of the delay [57].

Open Questions for the Internal Attention Framework

For external information, attentional selection comes with a clear trade-off: attention to one object entails withdrawal from others. Selective attention is, to some extent, a zero-sum operation. By contrast, selecting an internal representation need not have this same constraint. Arguably, the successful encoding and maintenance of individual items within WM already entails a high degree of individuation and orthogonalization of their representations, thus decreasing the amount of potential interference among memoranda relative to what can occur during the encoding phase. Therefore, while selective biasing may still be in operation, the nature of the substrate is such that its consequences may be different. In principle, at least, they could still be selected later, at low or no cost. From a functional perspective, it would be desirable to maintain memories for as long as they are potentially relevant to behavior, and only delete them once they are unlikely to be useful. Therefore, while the biased-competition principle is a good starting point for proposing a mechanism of internally guided attention, its most basic component (selection via biased competition) may operate in crucially different ways on external versus internal information.

The need for a distinction between external and internal attention has been highlighted before [13]: ‘Attention is not unitary’ ([13], p. 76). We welcome such a careful differentiation between attention to perception and attention to WM. We further propose that, while both perception and WM have limited capacity, the nature of the limit may differ considerably between the two. In perception, the challenge is to form cohesive and individuated item representations by bringing together their various attributes and separating these from competing sources of stimulation. In WM, the challenge is to select and use appropriate items to guide behavior.

Nominally, selection of one piece of information from among several in WM must occur to prioritize it (via a retrocue, for example). However, we propose that this process differs at both the mechanistic and implementational levels from selection during perception. Our elaboration of the proposed mechanism follows in the next section.

Prioritizing Information in Working Memory

We propose that, instead of invoking 'internal attention,' the prioritization of WM contents is better described as the attentional selection and, importantly, the reformatting of one out of several currently held memories to guide the next action. We speculate that this is a multistep process. After a cue indicates the increased relevance of a particular item within WM, the first step is to orient toward and select the cued item in the WM store. Orienting and selection can be thought of as the targeting of those neurons (for example, in visual cortex) that are tuned to the location where the cued item was encountered and to the stimulus dimensions that are relevant to the memory (i.e., in a color WM task, activation might increase in color-sensitive visual areas, such as V4). This allows for the effective grouping of all features belonging to the cued object [82], which in turn may reduce noise in the neural population representing it ([39], see also Box 1), potentially leading to increases in the precision of recalling cued items (e.g., [46,59,83,84]). Orienting attention in WM may be almost identical, at the neural level, to the effects of preparatory orienting of attention for perception, and could explain the activation of canonical attention control circuits after a retrocue, as summarized above. Critically, prioritization in WM allows for the immediate selection of memoranda. Selection of the cued representation can be thought of as an increased activation of the neurons coding for the relevant features of the cued object, possibly via reactivation of an ensemble that has encoded the item in its latent state (see Box 2 and [73]). Importantly, selection acts on one out of several objects in memory that have already been individuated and stored separately. Therefore, this selection step differs in terms of prioritization in WM versus in perception: in WM, the relevant information

Box 1. Working Memory As Internal Attention

Undoubtedly, we have gained much from drawing parallels between internal and external attention. Much like selective attention towards perceptual representations is thought to bias competitive processing in favor of one representation over another, internal attention is argued to bias processing towards one mnemonic representation over others in a shared memory store [9,14,15,39]. The shift of resources improves retention of the cued item or the behavior guided by it. An influential review [13] argued that attention shares common principles across the substrates it acts on. What is shared are the purpose of attention (overcoming limited capacity via selection), and its consequence (modulation of the selected information). The core of this process is that 'multiple stimuli [. . .] compete for selection, and the goal of attention is to bias competition in favor of a target object' ([13], p. 75). Therefore, 'selecting a memory from competing memories should be viewed as an attentional operation. The cost is that unattended information may be missed'.

An extension of the internal attention account is that prioritizing a WM representation may equate to transforming it from a latent to an activated neural state [15,73,109,134,135], rather than transforming it to an output-oriented representation (Box 2, main text). Thus, active versus latent storage corresponds at the neural level with attended versus unattended WM states at the cognitive level. In a recent study [73], retrocues improved decoding of the cued item in retinotopic visual areas. The authors argued that this finding was consistent with the reactivation of a latent code in sensory brain regions (the activated population permitting improved decoding, as noted elsewhere, [136]). However, the sluggishness of the fMRI signal may have precluded them from testing whether the reactivation reflects a temporary process. Most importantly, it is unclear how moving from latent to active representation alone could account for improved recall accuracy without simultaneously decreasing memory for uncued representations. For example, increasing the activity of a neural ensemble might give it greater influence over a downstream readout population, compared with a competing ensemble encoding an uncued representation (i.e., biased competition [31]). Alternatively, activation might suppress activity of competing ensembles via lateral inhibition [39]. Either way, the increased activity confers a benefit only by virtue of its suppression of competitors.

Box 2. Latent versus Active Neural States Supporting Working Memory

Our framework predicts two functional states in WM: maintenance of information without a specific action plan, and of a prioritized item in an action-oriented format. We propose that the latter depends on flexible changes in the tuning of an action-oriented network that includes lateral PFC [124,125]. WM has been proposed to rely on changes in the underlying state of a neural ensemble [126], permitting **latent storage** without requiring sustained activity [127]. The latent state change could rely on many physiological mechanisms. What they have in common is that they do not depend on an unbroken chain of sustained spiking. One candidate is short-term synaptic plasticity: the connectivity within a neural ensemble changes such that its response to subsequent input reflects the WM content [128]. Alternatively, memory-specific ensembles could emerge by synchronizing to a common oscillatory rhythm [125].

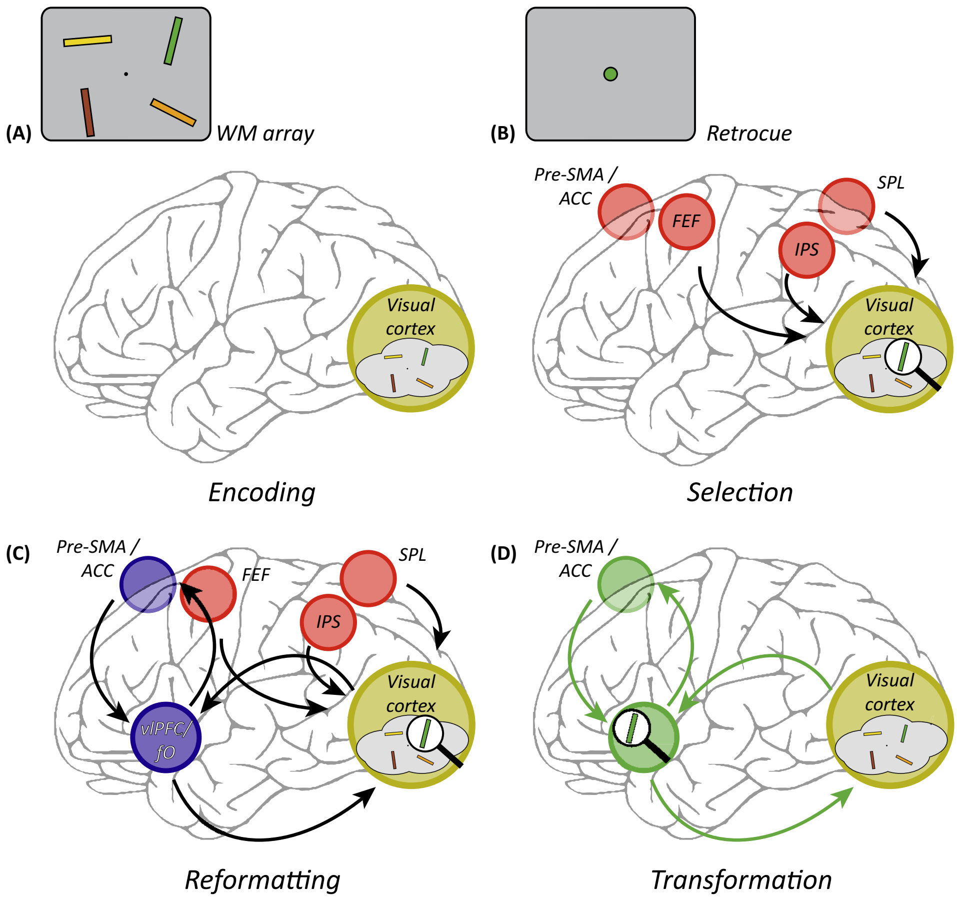
The concept of latent storage has been applied to maintenance of WM items per se, but latent storage may also be of particular relevance for representing prioritized WM items. Prioritization may lead to the task-specific transfer of a latent code stored in visual cortex to a lateral prefrontal network, possibly through temporary changes in the synaptic weights in PFC. Alternatively, unprioritized items could already be represented in PFC activity, in patterns that do not drive downstream motor regions [129] until they are prioritized.

After reconfiguration, sustained firing is unnecessary; thus, both prioritized and unprioritized representations are maintained using latent storage. Crucially, only the former may be represented in a format that is optimized for behavior. By contrast, unprioritized representations might be stored without immediately influencing behavior [130].

Latent storage of WM in changed connection weights invokes comparisons to long-term memory (LTM), which may operate along similar lines. Prominent WM theories propose activated LTM as the basis of WM [8,20,131]. While latent WM and LTM storage may both depend on changing synaptic weights instead of persistent firing, the specifics of how and for how long synaptic weights are changed may differ considerably between WM and LTM [132,133]. For example, any synaptic weight changes subserving WM must be short-lived to avoid interference from traces of recent WM contents. This is not a constraint for LTM. Furthermore, WM prioritization may reconfigure the action-oriented network in PFC so that it can immediately produce a context-appropriate response to the probe, without first needing to recall information from LTM (as proposed by some WM models, e.g., [131]).

is already stored and can be selected immediately. By contrast, during attention to external events, selection cannot take place until the cued event occurs, and selection requires the identification of cued objects and their associated features among all perceptual input. While orienting and selection can be clearly delineated from one another in WM, for succinctness we use ‘selection’ instead of ‘orienting and selection’ throughout the rest of this article.

Selection in memory appears to be a key element to successful prioritization in WM. Most accounts of retrocueing assume that this selection step is sufficient to account for the full range of experimental data. The essential novel aspect of our proposal is the speculation that the behavioral benefit additionally accrues in a further step. Following selection, the cued sensory representation can be transformed into a prospective, action-oriented representation, the better to influence behavior. By contrast, this step cannot take place in preparatory attention because the relevant information has not yet been presented. This transformation allows the current task set to become more specific. For instance, in a typical visual WM change-detection experiment, the task set on a trial without a retrocue might reflect the following rule: ‘press button A if the probe stimulus matches the WM stimulus that was presented at the same location, and button B otherwise’. Now imagine a retrocue trial, where one of the WM stimuli is cued (say it happens to be a green bar; Figure 1). In our framework, the reformatted representation of the cued WM stimulus is now part of the task set. Therefore, the task set has become more specific and much simpler: ‘press button A if the probe stimulus is green, and button B otherwise’. This process can be thought of as a form of cued task-set switching. When a task set is cued, responses are typically more efficient than when the task is not cued [85–88]. Therefore, the improved preparation for the task of responding accurately to the probe may, in part, contribute to the observed reaction time and accuracy benefits. Crucially, once reformatting is complete, it is no longer necessary to sustain selective attention to the sensory representation that initially stored the cued information [89].



Trends in Cognitive Sciences

Figure 1. Proposed Sequence of Prioritization in Working Memory (WM). (A) During encoding, sensory brain areas (in yellow) are recruited and modified to reflect the relevant features of a WM array. In the task illustrated here, sensory brain areas are the most likely substrate of memory maintenance because the latter requires memory for fine sensory details. However, the same mechanism could act on other brain areas if WM contents are stored elsewhere. (B) A cue indicating that one WM item is of particular relevance (the green oriented bar, in this case) leads to orientation toward, and selection of, the relevant representation. This operation recruits the top-down attention network that is also involved in external attention shifts (red circles). (C) In a second step, the identified information is prioritized. This step recruits a prefrontal network (in blue) comprising the anterior insula or ventrolateral prefrontal cortex and frontal operculum (vlPFC/fO), and possibly presupplementary motor area or anterior cingulate cortex (pre-SMA/ACC). (D) Finally, the selected representation is now reformatted to bring the network into an optimal state to respond to the recall demands of the probe stimulus.

Multiple States of Representation in Working Memory

Our framework helps explain findings that are harder to reconcile with the prevailing account. For instance, in apparent contradiction of the sustained attention model, maintaining a constant attentional focus on cued representations is not necessary for retrocues to benefit behavior: after a retrocue has been fully processed, attention can be withdrawn from the cued item towards another task [90,91] or another WM representation [44] without impacting the retrocue benefit. These findings can be readily explained in our framework since, after prioritization and reformatting are complete, a sustained selective bias is no longer strictly necessary.

As we argue, retrocueing benefits arise in large part due to the prospective reformatting of the cued representation for use at the time of the probe. Reformatting appears to be a flexible process, meaning that other stored items could be prioritized at minimal cost. This is consistent with the finding that benefits can occur without costs to uncued items under some circumstances [18,46–48,92]. At least in principle, resource trade-offs are not a necessity for benefits. Arguably, our framework is also consistent with the intuition that items held simultaneously in WM are individually prioritized at different points in time as they become relevant to behavior [93]. For example, previously uncued items can be subsequently refocused by a second retrocue [44,45] or an internally generated change in expectation [94]. In each case, the retrocue effectively increases net WM capacity. In contrast to our proposal, a sustained internal attention mechanism that enhances the target representation and suppresses distractors should entail a trade-off in memory and, therefore, cannot explain such findings as easily.

As mentioned above, invalid trials do appear to create costs in some studies. In our framework, such costs could still occur for a variety of reasons. First, there are scenarios in which the initial selection step could create costs to uncued representations. For example, high cue validity (e.g., when cues correctly predict the probe item on almost 100% of trials) may encourage a strategy of focusing all resources solely on the cued item by dropping uncued items from memory [83]. Such a strategy would be less successful when cues have lower validity, where it pays off to maintain uncued items in case of an invalid trial [46–48,95,96]. Additionally, high-validity cues might be used on a relatively higher proportion of trials, further increasing the retrocueing benefit (see Outstanding Questions).

The selection step might also account for effects of retrocueing on the precision of memory in some studies. Memory errors can be decomposed into errors due to Gaussian noise in the representation of the probed memory (i.e., its precision), errors due to forgetting, and errors due to misreporting the feature of an incorrect item [97,98]. The selection step, by strengthening the association between a cued location and the features of the object presented at that location, and by suppressing some of the residual interference from other items stored in WM, could lead to reductions in noise in the representation, leading to the occasional observation of small increases in memory precision [46,59,83,84]. However, the neural basis of this effect is unknown and difficult to fully explain, even within the sustained attention account. Finally, the selection step may also be used to select an ensemble of multiple items from WM when multi-item cues are used [99]. In this case, it appears that the entire set of cued items may be prioritized as an ensemble, rather than all cued items individually [100], possibly drawing on the representation of WM contents by the visual system at multiple, hierarchical organized spatial levels (from features bound to objects to ensembles of objects [82,101,102]).

Furthermore, invalidity costs may also arise during the second step of reformatting the cued item. When a cued item is reformatted into an action-oriented format and a corresponding probe is anticipated, invalid trials may produce costs because of errors in probe anticipation and task preparation instead of, or in addition to, any memory errors. This appears to fit the general pattern of the data: in single-cue studies, an invalid trial will occur when the incorrect task set is prepared and, therefore, generates switch costs or response conflict (e.g., [19,40,62]), which appear to have a particularly consistent influence on reaction time in addition to reducing memory accuracy somewhat [96]. This would be likely when an incorrect probe is expected, while uncued items are still partially retained in memory. By contrast, studies using a second cue in the delay that can redirect prioritization to a previously uncued item tend to find smaller or no costs on those trials [18,103–105]. Similarly, unanticipated **task switches** are known to incur behavioral costs [88,106,107]. In our framework, probing an uncued item amounts to an unanticipated task switch because the response must now be based on different information. Given that the task-set representation is necessarily limited (since only one given set of rules should determine actions at any one time, especially if other rules would produce conflicting behavior), this could incur costs. Therefore, task-switching costs induced by setting up an inappropriate task set, over and above impaired memory alone, may explain performance decreases for uncued items. Importantly, this need not indicate a competition between the memory representations themselves. As a result, we expect cueing costs to be minimal when probe anticipation is controlled (for example, when a second cue cancels a prior retrocue).

In sum, our framework is consistent with several behavioral findings that are difficult to reconcile with a purely attentional account. Our framework, based on representational reformatting, relates to previous proposals [19,95] arguing that retrocues modulate the representation of a cued stimulus via attentional strengthening without necessarily requiring that other stimuli are deleted to provide more resources. Importantly, our framework makes several testable predictions about the nature of the cueing benefit. First, knowing the form of recall will affect the magnitude of the retrocue benefit because it will allow for more specific task-set preparation. Second, several studies have shown that visual attention is drawn to items in the environment that match an item held in WM (in at least one of its features, e.g., location [12], color [108], etc.). It has been shown that this effect occurs only for items in the focus of attention [109]. Therefore, retrocued items should guide attention more than defocused items. However, this effect of attentional capture should also depend on the format in which they are about to be recalled. This appears to be the case [108], but has not yet been explicitly tested. For example, visual stimuli resembling a retrocued item should show increased attentional capture compared with stimuli resembling unprioritized WM contents. However, our framework predicts that this capture effect should be larger if the WM task requires precise visual information, compared with when prioritized items can be maintained via a verbal strategy.

Neural Evidence for Multiple States in Working Memory

Neuroimaging studies support the existence of a second stage in the prioritization of information in WM. Overall, these studies suggest that cueing a memory leads to reorganization of an output-oriented circuit that can then drive behavior faster and more accurately. These findings fall into two categories. The first is that output-related brain regions respond to retrocues and correlate with behavior, and the second is the additional activation of regions previously associated with task-set switching. We discuss these sets of results in turn.

First, in addition to the well-documented activation of the top-down attention network (Figure 2), numerous studies have found additional activation in (primarily ventrolateral) prefrontal cortex (PFC) [53–57,80,81] and striatum, which are less reliably related to external attention shifts. In one recent fMRI study involving retrocues and precues [110], retrocues led to correlations between the response strength of the caudate (as well as premotor cortex) and improvements in memory (as measured by reaction time). The authors [110] argued that this finding was consistent with the existence of **output gating** in WM. The 'output gate' here can be thought of as a bottleneck forcing the selection of one item from all items currently held in WM, so that it alone can guide the next action. The 'output gate' concept may be analogous to the behavior-guiding representational state proposed here. Other studies have also found striatal activation in response to retrocues on occasion [51], but this structure has generally been overlooked in discussions of the topic.

Second, the ventrolateral PFC, stretching into the frontal operculum or anterior insula [111], is consistently activated in response to retrocues [49,53–57,59,80,81], and has been shown to be uniquely activated during retrocueing compared with external attention shifts (Figure 2) [54]. The role of these areas is still unclear. Several studies have indicated that ventrolateral PFC is involved in the top-down access to, and selection from, sensory cortex of the cued information [80,81]. Consequently, disrupting activity in this area via transcranial magnetic stimulation reduces the benefits of retrocueing [81]. In addition, retrocueing tasks recruit task-switching-related brain circuits in lateral and medial PFC [112]. This is consistent with our interpretation

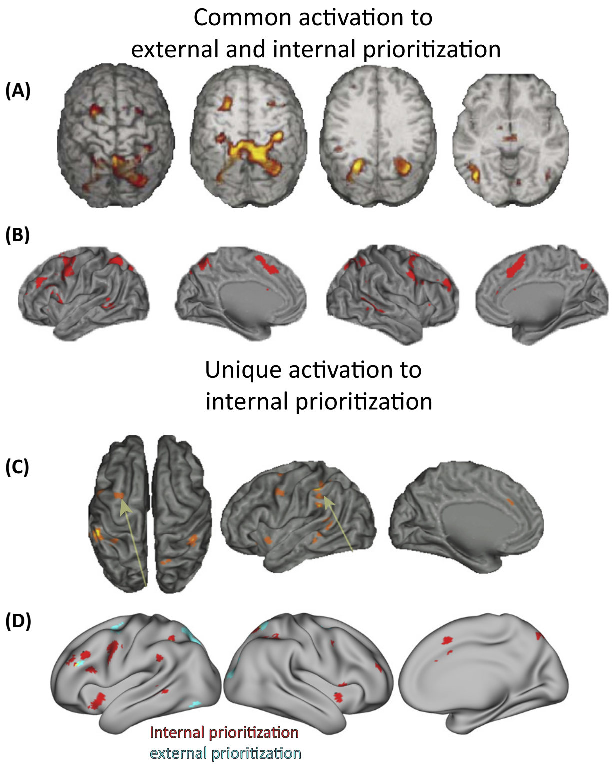


Figure 2. Shared and Unique Networks for Attentional Selection and Prioritization in Working Memory (WM). (A) Spatial cues directing attention to external stimuli or to content in WM activate a network spanning frontal eye fields, the presupplementary motor cortex (pre-SMA) and anterior cingulate cortex (ACC), the intraparietal sulcus, and the superior parietal lobule [49]. (B) This overlap has been confirmed in multiple neuroimaging studies [54] and a meta-analysis [59]. (C) Additional areas respond only to prioritization within WM, including ventrolateral prefrontal cortex stretching into the frontal operculum and anterior insula. In the pre-SMA and ACC, either activation is stronger than during external attention shifts, or additional subregions are recruited [52]. (D) A recent meta-analysis [59] found several, mostly prefrontal, areas responding to retrocues (internal prioritization, red) but not to precues (external prioritization, blue).

that transferring a retrocued item into the behavior-optimized state is akin to implementing a new task set.

These studies imply that additional prefrontal circuits are responsible for the top-down prioritization of items in WM. However, fMRI lacks the temporal resolution to determine whether these additional areas and more traditional attention-related areas co-activate simultaneously, or whether they activate sequentially (as proposed in our framework). We predict that using selection to reactivate a memory representation is a transient process (step 1: selection) that leads to a reconfiguration of the stimulus-response mappings of an action-selection network (step 2: reformatting). After reformatting, sustained attention is no longer necessary (Figure 1). One possibility is that cued information has been transferred to the PFC [2,41,113] via temporary synchronization of the cued region [114] and that, after this process is completed, the attention-related modulation subsides. Studies investigating the neural basis of the focus of attention have found evidence that the lateral parietal cortex, rather than just PFC, is critical for the deployment of this function [115–118]. The extent of network interactions between these areas during prioritization remains to be fully investigated, but a recent study found evidence that frontal and parietal areas are both important for switching a working-memory representation into the focus of attention [119]. Behavioral data suggest that the benefits of retrocues emerge after 300–500 ms [120–122]; that is, the entire process of selection is completed within less than a second, making it difficult to use methods with low temporal resolution (such as BOLD fMRI) to settle the question of whether sustained attention to a cued feature is necessary for prioritized read-out.

Neuroimaging methods with the requisite temporal resolution, such as electroencephalography or magnetoencephalography (MEG), have suggested that orienting and selection are time limited [59,63,123]. A recent MEG study [59] used lateralization of 10-Hz oscillations as a marker of internal attention shifts. The relative power of 10-Hz oscillations in sensory brain areas contralateral to where attention is directed, compared with power in ipsilateral areas, is a reliable indicator of covert attention shifts to external stimuli [60]. Similarly, cueing a location where a current WM item had previously appeared led to reliable lateralization. However, the lateralization of 10-Hz oscillations was transient after a retrocue, subsiding in less than a second. Given that the lateralization was only temporary, it seems unlikely that sustained attention or sustained active processing in a retinotopic representational format is necessary for retrocueing benefits to occur. Instead, this benefit could be conferred by a representational state change. Interestingly, additional activation in the insula that was specific to retrocues appeared only after the top-down attentional signal had peaked, supporting the idea of a two-step process of attentional selection, followed by representational reformatting [59]. The neural basis of representations after prioritization is still being investigated (Box 1). A second study testing the efficacy of retrocues in older participants [63] confirmed the same temporary lateralization. Interestingly, the strength of the retrocueing benefit (the increase in accuracy compared to a neutral cue) correlated negatively with the duration of the 10-Hz lateralization. Participants with the largest benefit showed the most transient lateralization: therefore, faster prioritization may indicate a more accurate reconfiguration of the network. While this is consistent with our hypothesis that selection is only a temporary process, it seems to contradict the idea that sustained internal attention is crucial to improving behavior, because this should result in the exact opposite pattern: improved memory in those participants with more sustained 10-Hz lateralization.

Concluding Remarks

In summarizing the behavioral and neural literature on prioritization of WM contents, we have argued that prioritizing an item arises in multiple stages and across multiple representational states. When the behavioral relevance of one out of several items in memory increases, top-down selection activates the neural subpopulation coding the cued information. Importantly, in a second step, the cued information undergoes a transformation in its representational state, from a task-agnostic mnemonic representation to a task-specific representation that is best suited to guide behavior. This transformation may coincide with a transfer of information from sensory to action-guiding areas of the brain. However, whether it also coincides with a temporary switch from latent to active representation (i.e., involving sustained spiking of memory-encoding neurons) is still unresolved. With recent advances in modeling memory-guided behavior and in multivariate analysis of high-dimensional neural recordings, we are confident that the predictions arising from our framework can be put to the test soon.

Acknowledgments

This article was funded by the Medical Research Council (to M.G.S.), the Wellcome Trust (A.C.N.: Senior Investigator Award 104571/Z/14/Z, and N.E.M.), University College, Oxford (N.E.M.), the James S. McDonnell Foundation (A.C.N.), and the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre based at Oxford University Hospitals NHS Foundation Trust, Oxford University. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health. We are grateful to Nahid Zokaei, Freek van Ede, and two anonymous reviewers for helpful comments on the manuscript, and to George Wallis for help with figures.

Outstanding Questions

What is the representational format and neural substrate of a prioritized WM item? How does this format relate to the representation of task sets or task rules?

What is the relationship between cue validity and the size of the retrocueing effect?

Are similar prioritization mechanisms also important in preparing for more classic forms of WM manipulation (e.g., mental arithmetic)?

In reality, we experience a continuous stream of thoughts passing through WM. How do we extend the concept of flexible prioritization to continuous, temporally extended cognition?

How do we switch between an internal and an external focus?

Do long-term memory and WM share selection and prioritization mechanisms? Furthermore, how does long-term memory influence the interplay between perception and WM?

Retrieval from long-term memory can induce forgetting of associated memories. Does a similar phenomenon exist in WM?

Can we dissociate the short-term representation of task goals or rules from the representation of other kinds of content in WM? Does the neural dissociability of goals and content depend on the task context?

WM is sustained by several representational states. Which of these corresponds to the traditional notion of the attention-guiding template? What other sources of attentional guidance exist, and what can the fractionation of WM tell us about the fractionation of the control of attention?

References

1. Baddeley, A.D. and Hitch, G. (1974) Working memory. *Psychol. Learn. Motiv.* 8, 47–89
2. Miller, E.K. and Cohen, J.D. (2001) An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24, 167–202
3. Engle, R.W. *et al.* (1999) Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *J. Exp. Psychol. Gen.* 128, 309–331
4. Smith, P.L. and Ratcliff, R. (2009) An integrated theory of attention and decision making in visual signal detection. *Psychol. Rev.* 116, 283–317
5. Otto, A.R. *et al.* (2013) Working-memory capacity protects model-based learning from stress. *Proc. Natl. Acad. Sci. U. S. A.* 110, 20941–20946
6. Luck, S.J. and Vogel, E.K. (2013) Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends Cogn. Sci.* 17, 391–400
7. Phillips, W.A. (1974) On the distinction between sensory storage and short-term visual memory. *Percept. Psychophys.* 16, 283–290
8. Cowan, N. (2000) The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.* 24, 87–114 discussion 114–185
9. Ma, W.J. *et al.* (2014) Changing concepts of working memory. *Nat. Neurosci.* 17, 347–356
10. Gazzaley, A. and Nobre, A.C. (2011) Top-down modulation: bridging selective attention and working memory. *Trends Cogn. Sci.* 16, 129–135
11. Awh, E. and Jonides, J. (2001) Overlapping mechanisms of attention and spatial working memory. *Trends Cogn. Sci.* 5, 119–126
12. Awh, E. *et al.* (2000) The role of spatial selective attention in working memory for locations: evidence from event-related potentials. *J. Cogn. Neurosci.* 12, 840–847
13. Chun, M.M. *et al.* (2011) A taxonomy of external and internal attention. *Annu. Rev. Psychol.* 62, 73–101
14. Chun, M.M. (2011) Visual working memory as visual attention sustained internally over time. *Neuropsychologia* 49, 1407–1409
15. Kiyonaga, A. and Egner, T. (2013) Working memory as internal attention: toward an integrative account of internal and external selection processes. *Psychon. Bull. Rev.* 20, 228–242
16. Vogel, E.K. and Machizawa, M.G. (2004) Neural activity predicts individual differences in visual working memory capacity. *Nature* 428, 748–751
17. Cusack, R. *et al.* (2009) Encoding strategy and not visual working memory capacity correlates with intelligence. *Psychon. Bull. Rev.* 16, 641–647
18. Landman, R. *et al.* (2003) Large capacity storage of integrated objects before change blindness. *Vis. Res.* 43, 149–164
19. Griffin, I.C. and Nobre, A.C. (2003) Orienting attention to locations in internal representations. *J. Cogn. Neurosci.* 15, 1176–1194
20. Oberauer, K. and Hein, L. (2012) Attention to information in working memory. *Curr. Dir. Psychol. Sci.* 21, 164–169
21. Holmes, J. *et al.* (2010) Working memory deficits can be overcome: impacts of training and medication on working memory in children with ADHD. *Appl. Cogn. Psychol.* 24, 827–836
22. Gold, J.M. *et al.* (2010) Reduced capacity but spared precision and maintenance of working memory representations in schizophrenia. *Arch. Gen. Psychiatry* 67, 570–577
23. Gazzaley, A. *et al.* (2005) Top-down suppression deficit underlies working memory impairment in normal aging. *Nat. Neurosci.* 8, 1298–1300
24. Reuter-Lorenz, P.A. and Sylvester, C.-Y.C. (2005) The cognitive neuroscience of working memory and aging. In *Cognitive Neuroscience of Aging: Linking Cognitive and Cerebral Aging* (Cabeza, R. *et al.*, eds), pp. 186–217, Oxford University Press
25. Pashler, H. (1984) Processing stages in overlapping tasks: evidence for a central bottleneck. *J. Exp. Psychol. Hum. Percept. Perform.* 10, 358–377
26. Raymond, J.E. *et al.* (1992) Temporary suppression of visual processing in an RSVP task: an attentional blink? *J. Exp. Psychol. Hum. Percept. Perform.* 18, 849–860
27. Zanto, T.P. and Gazzaley, A. (2009) Neural suppression of irrelevant information underlies optimal working memory performance. *J. Neurosci.* 29, 3059–3066
28. Schmidt, B.K. *et al.* (2002) Voluntary and automatic attentional control of visual working memory. *Percept. Psychophys.* 64, 754–763
29. Murray, A.M. *et al.* (2011) Markers of preparatory attention predict visual short-term memory performance. *Neuropsychologia* 49, 1458–1465
30. Gazzaley, A. (2011) Influence of early attentional modulation on working memory. *Neuropsychologia* 49, 1410–1424
31. Desimone, R. and Duncan, J. (1995) Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222
32. Hillyard, S.A. *et al.* (1998) Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. *Philos. Trans. R. Soc. B Biol. Sci.* 353, 1257–1270
33. Nobre, A.C. *et al.* (1998) Modulation of human extrastriate visual processing by selective attention to colours and words. *Brain* 121, 1357–1368
34. Kastner, S. and Ungerleider, L.G. (2000) Mechanisms of visual attention in the human cortex. *Annu. Rev. Neurosci.* 23, 315–341
35. James, W. (1890) *The Principles of Psychology,* Henry Holt
36. Johnson, M.K. *et al.* (2002) Second thoughts versus second looks: an age-related deficit in reflectively refreshing just-activated information. *Psychol. Sci.* 13, 64–67
37. Chun, M.M. and Johnson, M.K. (2011) Memory: enduring traces of perceptual and reflective attention. *Neuron* 72, 520–535
38. Oberauer, K. (2002) Access to information in working memory: exploring the focus of attention. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 411–421
39. Bays, P.M. (2015) Spikes not slots: noise in neural populations limits working memory. *Trends Cogn. Sci.* 19, 431–438
40. Astle, D.E. *et al.* (2012) Orienting attention to locations in mental representations. *Atten. Percept. Psychophys.* 74, 146–162
41. Sligte, I.G. *et al.* (2010) Detailed sensory memory, sloppy working memory. *Front. Psychol.* 1, 175
42. Makovski, T. *et al.* (2008) Orienting attention in visual working memory reduces interference from memory probes. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 369–380
43. Gözenman, F. *et al.* (2014) Invalid retro-cues can eliminate the retro-cue benefit: evidence for a hybridized account. *J. Exp. Psychol. Hum. Percept. Perform.* 40, 1748–1754
44. Rerko, L. *et al.* (2014) Retro-cue benefits in working memory without sustained focal attention. *Mem. Cogn.* 42, 712–728
45. van Moorselaar, D. *et al.* (2015) Forgotten but not gone: retro-cue costs and benefits in a double-cueing paradigm suggest multiple states in visual short-term memory. *J. Exp. Psychol. Learn. Mem. Cogn.* 41, 1755–1763
46. Gunseli, E. *et al.* (2015) The reliability of retro-cues determines the fate of noncued visual working memory representations. *Psychon. Bull. Rev.* 22, 1334–1341
47. Berryhill, M.E. *et al.* (2012) Shifting attention among working memory representations: testing cue type, awareness, and strategic control. *Q. J. Exp. Psychol.* 65, 426–438
48. Shimi, A. *et al.* (2014) Orienting attention within visual short-term memory: development and mechanisms. *Child Dev.* 85, 578–592
49. Nobre, A.C. *et al.* (2004) Orienting attention to locations in perceptual versus mental representations. *J. Cogn. Neurosci.* 16, 363–373
50. Bledowski, C. *et al.* (2009) What “works” in working memory? Separate systems for selection and updating of critical information. *J. Neurosci.* 29, 13735–13741

51. Tamber-Rosenau, B.J. (2011) Cortical mechanisms of cognitive control for shifting attention in vision and working Memory. *J. Cogn. Neurosci.* 23, 2905–2919
52. Nee, D.E. *et al.* (2013) A meta-analysis of executive components of working memory. *Cereb. Cortex* 23, 264–282
53. Nee, D.E. and Jonides, J. (2008) Neural correlates of access to short-term memory. *Proc. Natl. Acad. Sci. U. S. A.* 105, 14228–14233
54. Nee, D.E. and Jonides, J. (2009) Common and distinct neural correlates of perceptual and memorial selection. *NeuroImage* 45, 963–975
55. Lepsien, J. and Nobre, A.C. (2006) Cognitive control of attention in the human brain: insights from orienting attention to mental representations. *Brain Res.* 1105, 20–31
56. Lepsien, J. *et al.* (2005) Directing spatial attention in mental representations: interactions between attentional orienting and working-memory load. *NeuroImage* 26, 733–743
57. Lepsien, J. *et al.* (2011) Modulation of working-memory maintenance by directed attention. *Neuropsychologia* 49, 1569–1577
58. Yeh, Y.-Y. *et al.* (2007) The neural correlates of attention orienting in visuospatial working memory for detecting feature and conjunction changes. *Brain Res.* 1130, 146–157
59. Wallis, G. *et al.* (2015) Frontoparietal and cingulo-opercular networks play dissociable roles in control of working memory. *J. Cogn. Neurosci.* 15, 1–16
60. Worden, M.S. *et al.* (2000) Anticipatory biasing of visuospatial attention indexed by retinotopically specific-band electroencephalography increases over occipital cortex. *J. Neurosci.* 20, 1–6
61. Poch, C. *et al.* (2014) Modulation of alpha and gamma oscillations related to retrospectively orienting attention within working memory. *Eur. J. Neurosci.* 40, 2399–2405
62. Myers, N.E. *et al.* (2015) Temporal dynamics of attention during encoding versus maintenance of working memory: complementary views from event-related potentials and alpha-band oscillations. *J. Cogn. Neurosci.* 27, 492–508
63. Mok, R.M. *et al.* (2016) Behavioral and neural markers of flexible attention over working memory in aging. *Cereb. Cortex* 26, 1831–1842
64. van Ede, F. *et al.* (2016) Temporal expectations guide dynamic prioritization in visual working memory through attenuated alpha oscillations. *J. Neurosci.* 37, 437–445
65. Kuo, B.-C. *et al.* (2012) Attention modulates maintenance of representations in visual short-term memory. *J. Cogn. Neurosci.* 24, 51–60
66. Nobre, A.C. (2008) Spatial attention can bias search in visual short-term memory. *Front. Hum. Neurosci.* 1, 4
67. Williams, M. and Woodman, G.F. (2012) Directed forgetting and directed remembering in visual working memory. *J. Exp. Psychol. Learn. Mem. Cogn.* 38, 1206–1220
68. Lepsien, J. and Nobre, A.C. (2007) Attentional modulation of object representations in working memory. *Cereb. Cortex* 17, 2072–2083
69. Kuo, B.-C. *et al.* (2014) Attention biases visual activity in visual STM. *J. Cogn. Neurosci.* 26, 1377–1389
70. Lewis-Peacock, J.A. and Postle, B.R. (2012) Decoding the internal focus of attention. *Neuropsychologia* 50, 470–478
71. Lewis-Peacock, J.A. (2014) Neural evidence for the flexible control of mental representations. *Cereb. Cortex* 25, 3303–3313
72. LaRocque, J.J. *et al.* (2013) Decoding attended information in short-term memory: an EEG study. *J. Cogn. Neurosci.* 25, 127–142
73. Sprague, T.C. *et al.* (2016) Restoring latent visual working memory representations in human cortex. *Neuron* 91, 694–707
74. Pasternak, T. and Greenlee, M.W. (2005) Working memory in primate sensory systems. *Nat. Rev. Neurosci.* 6, 97–107
75. Scolari, M. and Serences, J.T. (2010) Basing perceptual decisions on the most informative sensory neurons. *J. Neurophysiol.* 104, 2266–2673
76. Riggall, A.C. and Postle, B.R. (2012) The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. *J. Neurosci.* 32, 12990–12998
77. Christophel, T.B. (2017) The distributed nature of working memory. *Trends Cogn. Sci.* 21, 111–124
78. Lee, S.-H. *et al.* (2013) Goal-dependent dissociation of visual and prefrontal cortices during working memory. *Nat. Neurosci.* 16, 997–999
79. Ester, E.F. *et al.* (2015) Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory. *Neuron* 87, 893–905
80. Nelissen, N. *et al.* (2013) Frontal and parietal cortical interactions with distributed visual representations during selective attention and action selection. *J. Neurosci.* 33, 16443–16458
81. Higo, T. *et al.* (2011) Distributed and causal influence of frontal operculum in task control. *Proc. Natl. Acad. Sci. U. S. A.* 108, 4230–4235
82. Brady, T.F. *et al.* (2011) A review of visual memory capacity: beyond individual items and toward structured representations. *J. Vis.* 11, 4–4
83. Williams, M. *et al.* (2013) The benefit of forgetting. *Psychon. Bull. Rev.* 20, 348–355
84. Pertzov, Y. *et al.* (2016) Rapid forgetting results from competition over time between items in visual working memory. *J. Exp. Psychol. Learn. Mem. Cogn.* Published online September 26, 2016. http://dx.doi.org/10.1037/xlm0000328
85. Meiran, N. (1996) Reconfiguration of processing mode prior to task performance. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 1423–1442
86. Rogers, R.D. and Monsell, S. (1995) Costs of a predictable switch between simple cognitive tasks. *J. Exp. Psychol. Gen.* 124, 207–231
87. Monsell, S. (2003) Task switching. *Trends Cogn. Sci.* 7, 134–140
88. Allport, A. *et al.* (1994) Shifting intentional set: exploring the dynamic control of tasks. *Atten. Performance* 15, 412–452
89. Muhle-Karbe, P.S. (2016) Neural coding for instruction-based task sets in human frontoparietal and visual cortex. *Cereb. Cortex* 27, 1891–1905
90. Hollingworth, A. and Maxcey-Richard, A.M. (2013) Selective maintenance in visual working memory does not require sustained visual attention. *J. Exp. Psychol. Hum. Percept. Perform.* 39, 1047–1058
91. Makovski, T. and Pertzov, Y. (2015) Attention and memory protection: interactions between retrospective attention cueing and interference. *Q. J. Exp. Psychol. (Hove)* 68, 1735–1743
92. Souza, A.S. *et al.* (2015) Refreshing memory traces: thinking of an item improves retrieval from visual working memory. *Ann. N. Y. Acad. Sci.* 1339, 20–31
93. Duncan, J. (2013) The structure of cognition: attentional episodes in mind and brain. *Neuron* 80, 35–50
94. van Ede, F. *et al.* (2016) Temporal expectations guide dynamic prioritization in visual working memory through attenuated alpha oscillations. *J. Neurosci.* 37, 437–445
95. Souza, A.S. and Oberauer, K. (2016) In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Atten. Percept. Psychophys.* 78, 1839–1860
96. Gressmann, M. and Janczyk, M. (2016) The (un)clear effects of invalid retro-cues. *Front. Psychol.* 7, 244
97. Zhang, W. and Luck, S.J. (2008) Discrete fixed-resolution representations in visual working memory. *Nature* 453, 233–235
98. Bays, P.M. *et al.* (2009) The precision of visual working memory is set by allocation of a shared resource. *J. Vis.* 9, 7
99. Matsukura, M. *et al.* (2007) Attention effects during visual short-term memory maintenance: protection or prioritization? *Percept. Psychophys.* 69, 1422–1434
100. Matsukura, M. and Vecera, S.P. (2015) Selection of multiple cued items is possible during visual short-term memory maintenance. *Atten. Percept. Psychophys.* 77, 1625–1646

101. Brady, T.F. and Alvarez, G.A. (2011) Hierarchical encoding in visual working memory: ensemble statistics bias memory for individual items. *Psychol. Sci.* 22, 384–392
102. Brady, T.F. *et al.* (2009) Compression in visual working memory: using statistical regularities to form more efficient memory representations. *J. Exp. Psychol. Gen.* 138, 487–502
103. Li, Q. and Saiki, J. (2014) The effects of sequential attention shifts within visual working memory. *Front. Psychol.* 5, 965
104. Rerko, L. and Oberauer, K. (2013) Focused, unfocused, and defocused information in working memory. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 1075–1096
105. Souza, A.S. and Oberauer, K. (2015) Time-based forgetting in visual working memory reflects temporal distinctiveness, not decay. *Psychon. Bull. Rev.* 22, 156–162
106. Rushworth, M.F.S. (2002) Components of switching intentional set. *J. Cogn. Neurosci.* 14, 1139–1150
107. Miniussi, C. *et al.* (2005) Modulation of brain activity by selective task sets observed using event-related potentials. *Neuropsychologia* 43, 1514–1528
108. Olivers, C.N.L. *et al.* (2006) Feature-based memory-driven attentional capture: visual working memory content affects visual attention. *J. Exp. Psychol. Hum. Percept. Perform.* 32, 1243–1265
109. Olivers, C.N.L. *et al.* (2011) Different states in visual working memory: when it guides attention and when it does not. *Trends Cogn. Sci.* 15, 327–334
110. Chatham, C.H. *et al.* (2014) Corticostriatal output gating during selection from working memory. *Neuron* 81, 930–942
111. Petrides, M. and Pandya, D.N. (2002) Comparative cytoarchitectonic analysis of the human and the macaque ventrolateral prefrontal cortex and corticocortical connection patterns in the monkey. *Eur. J. Neurosci.* 16, 291–310
112. Richter, F.R. and Yeung, N. (2014) Neuroimaging studies of task switching. In *Task Switching and Cognitive Control* (Grange, J. and Houghton, G., eds), pp. 1–55, Oxford University Press
113. Buschman, T.J. and Miller, E.K. (2014) Goal-direction and top-down control. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369, 20130471
114. Liebe, S. *et al.* (2012) Theta coupling between V4 and prefrontal cortex predicts visual short-term memory performance. *Nat. Neurosci.* 15, 456–462
115. Öztekin, I. and Cowan, N. (2015) Editorial: representational states in memory: where do we stand? *Front. Hum. Neurosci.* 9, 453
116. Nee, D.E. and Jonides, J. (2013) Trisecting representational states in short-term memory. *Front. Hum. Neurosci.* 7, 796
117. Nee, D.E. and Jonides, J. (2013) Neural evidence for a 3-state model of visual short-term memory. *NeuroImage* 74, 1–11
118. Nee, D.E. and Jonides, J. (2011) Dissociable contributions of prefrontal cortex and the hippocampus to short-term memory: evidence for a 3-state model of memory. *NeuroImage* 54, 1540–1548
119. Nee, D.E. and Brown, J.W. (2013) Dissociable frontal-striatal and frontal-parietal networks involved in updating hierarchical contexts in working memory. *Cereb. Cortex* 23, 2146–2158
120. Tanoue, R.T. and Berryhill, M.E. (2012) The mental wormhole: internal attention shifts without regard for distance. *Atten. Percept. Psychophys.* 74, 1199–1215
121. Souza, A.S. *et al.* (2014) Focused attention improves working memory: implications for flexible-resource and discrete-capacity models. *Atten. Percept. Psychophys.* 76, 2080–2102
122. van Moorselaar, D. (2014) The time course of protecting a visual memory representation from perceptual interference. *Front. Hum. Neurosci.* 8, 1053
123. Spitzer, B. and Blankenburg, F. (2011) Stimulus-dependent EEG activity reflects internal updating of tactile working memory in humans. *Proc. Natl. Acad. Sci. U. S. A.* 108, 8444–8449
124. Crone, E.A. (2005) Neural evidence for dissociable components of task-switching. *Cereb. Cortex* 16, 475–486
125. Buschman, T.J. *et al.* (2012) Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. *Neuron* 76, 838–846
126. Stokes, M.G. (2015) "Activity-silent" working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn. Sci.* 19, 394–405
127. Funahashi, S. *et al.* (1993) Prefrontal neuronal activity in rhesus monkeys performing a delayed anti-saccade task. *Nature* 365, 753–756
128. Mongillo, G. *et al.* (2008) Synaptic theory of working memory. *Science* 319, 1543–1546
129. Kaufman, M.T. *et al.* (2014) Cortical activity in the null space: permitting preparation without movement. *Nat. Neurosci.* 17, 440–448
130. Bhandari, A. and Duncan, J. (2014) Goal neglect and knowledge chunking in the construction of novel behaviour. *Cognition* 130, 11–30
131. Ericsson, K.A. and Kintsch, W. (1995) Long-term working memory. *Psychol. Rev.* 102, 211–245
132. Abbott, L.F. and Regehr, W.G. (2004) Synaptic computation. *Nature* 431, 796–803
133. Citri, A. and Malenka, R.C. (2008) Synaptic plasticity: multiple forms, functions, and mechanisms. *Neuropsychopharmacology* 33, 18–41
134. Rose, N.S. *et al.* (2016) Reactivation of latent working memories with transcranial magnetic stimulation. *Science* 354, 1136–1139
135. Postle, B.R. (2015) The cognitive neuroscience of visual short-term memory. *Curr. Opin. Behav. Sci.* 1, 40–46
136. Lewis-Peacock, J.A. *et al.* (2012) Neural evidence for a distinction between short-term memory and the focus of attention. *J. Cogn. Neurosci.* 24, 61–79