Opinion
Prioritizing Information during Working Memory: Beyond Sustained Internal Attention
Nicholas E. Myers,1,2,* Mark G. Stokes,1,2 and Anna C. Nobre1,2,*

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.

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 precious limited cognitive resource.

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.

1Oxford Centre for Human Brain Activity, Oxford University, Oxford, UK
2Department 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 of 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 retrocuing 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
between the effects of retrocuing 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 retrocuing 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 retrocuing [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 individualization 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 [15,72,109,134,135], 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, [138]). 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 unceded 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 [51]). Alternatively, activation might suppress activity of competing ensembles via lateral inhibition [59]. Either way, the increased activity confers a benefit only by virtue of its suppression of competitors.
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 retrocuing 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].
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 reformattting are complete, a sustained selective bias is no longer strictly necessary.

As we argue, retrocuing benefits arise in large part due to the prospective reformattting of the cued representation for use at the time of the probe. Reformattting 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 retro cue [44,45] or an internally generated change in expectation [94]. In each case, the retro cue 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 retrocuing benefit (see Outstanding Questions).

The selection step might also account for effects of retrocuing 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 reformatatted 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
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 reactiva 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 retrocuing 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


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


97. Bays, P.M. et al. (2009) The precision of visual working memory is set by allocation of a shared resource. J. Vis. 9, 7


