Trends in Cognitive Sciences

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.