ANNUAL REVIEWS

Annual Review of Vision Science

Balancing Flexibility and Interference in Working Memory

Timothy J. Buschman

Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, New Jersey 08544, USA; email: tbuschma@princeton.edu

Annu. Rev. Vis. Sci. 2021. 7:367–88

First published as a Review in Advance on June 3, 2021

The *Annual Review of Vision Science* is online at vision.annualreviews.org

https://doi.org/10.1146/annurev-vision-100419-104831

Keywords

working memory, cognitive flexibility, capacity limitations, interference, neural dynamics

Abstract

Working memory is central to cognition, flexibly holding the variety of thoughts needed for complex behavior. Yet, despite its importance, working memory has a severely limited capacity, holding only three to four items at once. In this article, I review experimental and computational evidence that the flexibility and limited capacity of working memory reflect the same underlying neural mechanism. I argue that working memory relies on interactions between high-dimensional, integrative representations in the prefrontal cortex and structured representations in the sensory cortex. Together, these interactions allow working memory to flexibly maintain arbitrary representations. However, the distributed nature of working memory comes at the cost of causing interference between items in memory, resulting in a limited capacity. Finally, I discuss several mechanisms used by the brain to reduce interference and maximize the effective capacity of working memory.

INTRODUCTION

Working memory is the ability to internally maintain task-relevant information. It acts as a workspace on which you can place thoughts and ideas, manipulate them, and then use them to guide your behavior. By decoupling behavior from the immediate sensory world, working memory plays a fundamental role in almost all complex cognitive behaviors, from decision making to reward learning to cognitive control.

Everyday life is filled with behaviors that rely on working memory. For example, shopping in a grocery store requires holding a list of needed items in working memory. As you walk through the store, you iteratively select each item from working memory and use its name and visual properties to search the shelves. This example highlights two key characteristics of working memory. First, working memory is remarkably flexible—you can hold anything in mind, even without prior experience. In our example, this is illustrated by the variety of food items that one can hold in working memory, including novel ingredients for a new recipe. You can also recombine items in unique ways in working memory. Again, this is seen in our example—the order of items in working memory is not fixed to their original order on the list but can be reordered to optimize your path through the store. Of course, there are near-infinite numbers of paths to take through the grocery store, and the flexibility of working memory allows you to represent any of these.

The second characteristic of working memory is that it has a limited capacity. The average adult human can hold between three and four items in working memory at a time (although the exact limit varies across individuals and situations, as is discussed further below). In our example, this is illustrated by the need to regularly check your written list as you do your shopping, allowing you to update the contents of working memory with the next batch of items on the list.

In this review, I discuss the neural mechanisms supporting working memory and how they give rise to both the flexibility and limited capacity of working memory. In the first part of the review, I argue that flexibility arises from the interaction of structured representations in the sensory cortex with integrative representations in associative regions. This is codified in a flexible model of working memory that can maintain arbitrary inputs. However, as detailed in the second part of the review, the integration of representations in associative regions leads to interference between items in memory. This interference, in turn, limits the capacity of working memory. Finally, in the last part of the review, I discuss several mechanisms used by the brain to mitigate interference and maximize working memory capacity.

FLEXIBILITY OF WORKING MEMORY

Working memory is remarkably flexible. You can hold anything in working memory, even from the first time that you experience it. This flexibility is fundamental to our general intelligence, as it allows us to engage in a wide variety of ever-changing behaviors.

The flexibility of working memory extends to the neural level, as neurons can encode the memory of a wide variety of stimuli, including novel stimuli. To study this, one can train monkeys to perform a working memory task, such as the delayed match to sample task. In this task, monkeys are briefly shown an image (the sample) that they must remember over a memory delay (during which nothing is shown). After the delay, the animal is presented with a second test image and reports whether the test stimulus matches the sample stimulus (similar tasks are used in humans). Electrophysiological recordings during such a task have found that neurons in the prefrontal cortex represent the identity of the sample stimulus by responding with a different number of action potentials for different stimuli. Importantly, this response is also found during the memory delay—in this way, neurons in the prefrontal cortex can be said to encode (or represent) information about the items held in working memory (note that the strength of encoding is defined as how well the

identity of the item can be decoded). Reflecting their flexibility, neurons in the prefrontal cortex have been found to encode the memory of a wide variety of stimuli, including novel stimuli that had been experienced for the first time (Miller et al. 1991).

Memory information does not seem to depend on training: Neurons in the prefrontal cortex encoded information about the memory of recently presented stimuli, even before the animal had been trained on a working memory task (Meyers et al. 2012). Interestingly, training animals to perform a working memory task did not change how much information was encoded in the prefrontal cortex about the memory of a stimulus. This suggests that learning is not necessary for prefrontal neurons to represent memory information; rather, they can flexibly maintain this information from the beginning.

Further studies have shown that the ability to represent novel stimuli extends beyond the prefrontal cortex—imaging and lesion studies in humans suggest that the medial temporal lobe plays a critical role in maintaining memories for novel stimuli (perhaps through rapid associative learning) (Hasselmo & Stern 2006, Ranganath & D'Esposito 2001, Stern et al. 2001). Altogether, these results suggest that the flexibility of working memory is reflected in the flexibility of individual neurons, particularly in associative regions like the prefrontal cortex and the medial temporal lobe.

However, working memory representations extend beyond associative regions (Christophel et al. 2017). Previous work in humans and monkeys has found working memory representations in the sensory (Harrison & Tong 2009, Lee et al. 2005, Pasternak & Greenlee 2005), temporal (Miller et al. 1991, Ranganath 2006), and parietal cortexes (Arcizet et al. 2011, Sarma et al. 2016). Subcortical regions are also involved, with mnemonic activity sustained in the thalamus, basal ganglia, and superior colliculus (Isseroff et al. 1982, Levy et al. 1997, Postle & D'Esposito 2003, Rahmati et al. 2020). In this review, I argue that the distributed nature of working memory is critical to its flexibility. In this framework, associative regions integrate information across domains, creating high-dimensional representations that provide a flexible basis for representing arbitrary inputs. This is complemented by the structured representations in the sensory cortex, which represent the content of working memory. In the remainder of this section, I detail the characteristics of the associative and sensory systems alone, before discussing how combining the two systems can support flexible working memory.

Integration in Prefrontal Cortex Increases the Dimensionality of Working Memory

As sensory information is processed, it ascends the cortical hierarchy, converging onto associative brain regions, such as the prefrontal cortex and the medial temporal lobe (Markov et al. 2013). For example, the prefrontal cortex integrates inputs from the sensory cortex, the motor cortex, the basal ganglia, and the medial temporal lobe (Miller & Cohen 2001). This confluence of inputs allows neurons in the prefrontal cortex to encode a wide variety of sensory and cognitive variables, including sustaining memory representations about the color, shape, or identity of a visual stimulus (Christophel et al. 2017).

The convergence of information is seen in the selectivity of individual prefrontal neurons. Recordings in monkeys have shown that prefrontal neurons respond to a seemingly random conjunction of different sensory inputs, context, and time (Rigotti et al. 2013). For example, a prefrontal cortex neuron may selectively respond to the color of an object, but only in conjunction with a specific task or a specific moment in time (Barone & Joseph 1989, Buschman et al. 2012, Siegel et al. 2015). The exact conjunction of features that a neuron responds to is highly variable, with each neuron responding to a different combination of inputs (without any particular relationship to the selectivity of neighboring neurons). This creates a high-dimensional space for

representing cognitive variables across the entire neural population, with the response of each neuron acting as an independent dimension in this space (Rigotti et al. 2013). In other words, information is distributed, i.e., represented in the high-dimensional vector of neural activity across the population. Consistent with this, approximately 35% of prefrontal neurons are selective for any given task variable, such as the identity of a stimulus, the current task in effect, or the animal's behavioral response (Barak et al. 2013).

High-dimensional spaces that integrate different types of information are particularly useful for flexible working memory. By integrating from so many different brain regions, they can capture the diversity of information that one might want to hold in working memory. In addition, high-dimensional spaces allow for the maintenance of unique combinations of inputs. For example, while you have likely never seen a pink elephant, your familiarity with the color pink and elephants allows you to combine them in a novel manner. In this way, convergence into a high-dimensional space allows for a combinatorial explosion of representations, playing an important role in the flexibility of working memory.

One mechanism for generating high-dimensional representations is to have neurons respond to a random set of inputs. This could create the variety of conjunctive selectivity seen in the prefrontal cortex. Random connections are also an effective way to maintain information in working memory—computational models have shown that random connections in a recurrent neural network can create reservoir networks that are able to maintain a short-term memory of an input as it echoes through the network (Enel et al. 2016). Despite their random structure, these networks do a surprisingly good job of maintaining the memory of an input, approximating the performance of an optimized network (White et al. 2004) without needing to tune the responses of individual neurons to optimally encode a specific variable. Interestingly, memory performance of these networks is improved when representations are suitably sparse, as I return to below (Ganguli & Sompolinsky 2010). Of course, the major advantage of random connections is their arbitrary nature, which means that the network can represent arbitrary inputs (including novel stimuli).

In sum, associative regions, such as the prefrontal cortex, use high-dimensional, conjunctive representations to arbitrarily combine a wide variety of information (possibly from random connections). This may provide the neural basis for flexible working memory. However, on their own, the unstructured nature of random networks means that they lose information about how different items relate to one another, which makes it hard to generalize knowledge. For example, because the ordered nature of the color spectrum would not be captured by a purely random network, decoding a particular color (such as red) would require a unique decoder that is unrelated to the decoding of other similar colors (such as orange or purple). This is in contrast to the sensory cortex, where neural representations are highly structured.

Sensory Cortex Provides Structure to Working Memory

In the visual cortex, each neuron responds to a specific region of the visual world (its spatial receptive field) and a specific region of feature space (its tuning curve). The regions of spatial and feature space that a sensory neuron responds to are contiguous, with nearby features eliciting similar responses. For example, in the visual system, V1 neurons respond in a graded fashion to the tilt of a line (Hubel & Wiesel 1959), V4 neurons respond in a continuous way to the angle of an object's corner (Pasupathy & Connor 1999), and IT neurons respond in a graded fashion to specific dimensions of facial features (Higgins et al. 2020). These representations likely emerge through experience with the world, as a result of unsupervised learning maximizing the amount of information that the network encodes about sensory inputs (Simoncelli & Olshausen 2001). In this way, tuning curves are optimized to discriminate between sensory inputs. They also capture

statistical regularities in the world, embedding knowledge about the way in which stimulus features relate to one another in the world. For example, a pink ball is unlikely to suddenly transform into a green cube but could change to a red apple as the lighting shifts. These likelihoods are reflected in the structure of neural responses; neurons in the sensory cortex respond strongly to inputs with their preferred features, and to visual stimuli with similar features, but they are inhibited by inputs with substantially different features (surround suppression) (Angelucci & Bressloff 2006, Liu et al. 2018). What this means is that, across the population of neurons, the pattern of activity in response to a pink ball is similar to that of a red apple but very different from the response to a green cube.

As noted above, the sensory cortex is also involved in representing items held in working memory (Christophel et al. 2017). Similar to perception, the structure of the sensory cortex helps to improve the accuracy of working memory. As tuning curves are optimized for perception, they are also optimized to accurately encode stimulus information in working memory. Similarly, the knowledge of the world embedded in the structure of representations in the sensory cortex constrains memories to meaningful representations (i.e., representations that make sense in the world). In particular, surround suppression limits memories to a region of feature space around an input by inhibiting dissimilar features. This acts to stabilize memories—while a memory may wander from its initial representation (e.g., pink becomes red), the surround suppression makes it unlikely that a memory will transform into a very different representation (e.g., red becomes green).

Creating Flexible Memories by Balancing Structure and Randomness

Altogether, these results suggest that associative and sensory regions play complementary roles in working memory. The conjunctive, seemingly random nature of responses in associative regions is ideal for representing arbitrary information. However, while flexible, purely random representations are unstructured and therefore missing the knowledge of the world that is captured in the sensory cortex. This necessitates the distributed nature of working memory—distribution provides a framework for incorporating the relative advantages of both associative and sensory systems.

Recently, this framework was instantiated in the flexible working memory model (FWMM) (Bouchacourt & Buschman 2019). As schematized in **Figure 1**, the FWMM flexibly maintains arbitrary inputs in working memory in the interaction between two layers, one structured and one unstructured (random). The first layer acts as a sensory cortex, containing several subnetworks of neurons, each representing a different domain of sensory processing (e.g., color, shape, motion, orientation) or a different location in space. Each sensory neuron receives specific sensory inputs that, coupled with local recurrent connections, tune the neuron to a particular part of input space (similar to tuning curves in the brain) (**Figure 1**). The sensory layer is then randomly and reciprocally connected with a layer of control neurons. Both the reciprocity and randomness of these connections are important for flexible working memory.

The reciprocal nature of the connections between the sensory and control networks is what sustains representations in working memory. In brief, activity in the sensory network drives activity in a random subset of neurons in the control network that receive excitatory connections (approximately 35%). These control neurons reciprocally feed back excitation onto the same sensory neurons. In this way, the two-layer network acts as an attractor network. As with all attractor networks (Wang 2001), representations in the FWMM are stable (i.e., they do not change over time), allowing an item to be held in working memory.

The distributed nature of the random connections means that control neurons also excite neurons in other sensory subnetworks (or in other parts of the same subnetwork). However, because these projections are random, the top-down input into other neurons is unstructured and therefore elicits roughly equal amounts of direct excitation and surround inhibition. In other words, the

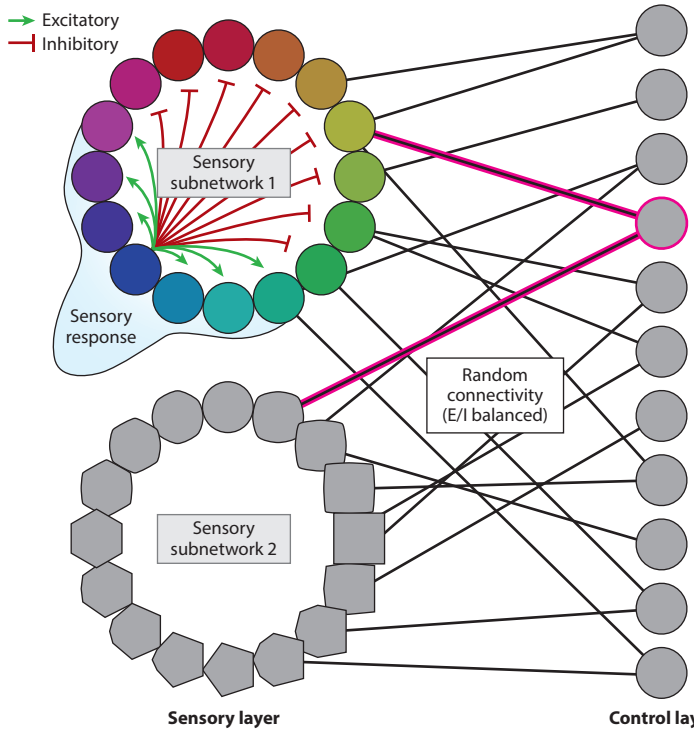

Figure 1

Schematic of flexible working memory model (FWMM). Memories are stored in the interaction between a sensory layer (*left*) and control layer (*right*). The sensory layer can contain multiple subnetworks, schematized as color and shape subnetworks. Connections within the sensory layer are excitatory (*green*) and inhibitory (*red*) for near and distal neurons, representing similar and dissimilar stimulus preferences, respectively. A schematic of a sensory neuron's tuning curve is shown along the outside of the ring. Connections between the sensory and control layers are random and reciprocal and have balanced excitation/inhibition (E/I). Control neurons receive inputs from all sensory subnetworks, resulting in conjunctive neurons. For example, the neuron highlighted in pink responds to yellow-green rounded rectangles. Figure adapted with permission from Bouchacourt & Buschman (2019).

top-down signals cancel out in the other sensory subnetworks (i.e., the sum effect is zero). From a geometric perspective, this means that each sensory subnetwork exists in a different subspace within the control network, and because these subspaces are defined randomly, they are uncorrelated with one another. The result is that the representation of one item stored in one subnetwork does not overlap with the representation of another item stored in another subnetwork (i.e., the subspaces are orthogonal).

In addition to preventing interference in the subnetworks, the randomness of the connections converging from the sensory layer to the control layer also endows the network with its flexibility. The random connections are (by definition) untuned with respect to the content of what is being held in working memory. This allows any representation in the sensory layer to be maintained in the network, regardless of the exact input or its distribution across the sensory layer (which it does without learning). Exemplifying the flexibility of the network, the FWMM structure works with different types of sensory inputs, including one-dimensional stimuli represented in a line of neurons (e.g., brightness), a circular space in a ring of neurons (e.g., color), a higher-dimensional variable in a sheet of neurons (e.g., faces), and long-term memories stored in a pattern-completing

recurrent network (e.g., Hopfield-like auto-associative networks) (for details on these other networks, see Bouchacourt & Buschman 2019).

Importantly, the flexible model of working memory only works because of the combination of structured representations in the sensory layer with the unstructured, random representations in the control layer. Projecting a structured representation, such as in the sensory network, into a random space maintains the topological structure of the inputs. This means that nearby representations in input space are nearby in the space defined by the random projections (and, likewise, distant representations are distant). For example, the representation of red would stay between orange and purple in a random projection while remaining far from green. In this way, the control network inherits the structure embedded in the sensory networks, providing meaning to the representations in the control network and constraining them to the space of reasonable representations (i.e., a red apple may become a pink ball but not a green cube).

Despite its simplicity, the FWMM captures many of the behavioral and physiological characteristics of working memory. First, as highlighted above, the network is able to maintain arbitrary inputs, capturing the flexibility of working memory seen in behavior (and in neural activity). The model also allows for unique combinations of representations to be maintained in working memory—inputs converge onto the control layer from multiple sensory subnetworks, allowing for the maintenance of novel conjunctive stimuli (e.g., a pink elephant).

Second, as seen in the brain, representations in the FWMM are distributed across the sensory and control networks, similar to working memory representations distributed across sensory and prefrontal/parietal cortices (Christophel et al. 2017). Consistent with both regions being critical to working memory, lesions of the prefrontal cortex impair working memory behavior (Pribram et al. 1952), and optogenetically suppressing the sensory cortex in mice during the memory delay impaired working memory (Zhang et al. 2019). The FWMM also predicts that holding an item in working memory should increase the interaction between prefrontal or parietal control regions and the structured sensory cortex. Consistent with this, previous work has shown that synchrony between the prefrontal cortex and visual cortex (V4) in monkeys predicts the monkeys' performance on a visual working memory task (Liebe et al. 2012). Similarly, functional connectivity between the prefrontal cortex and the fusiform face area (FFA) is increased when maintaining a face in working memory (Gazzaley et al. 2004), and causally disrupting the functional connectivity between the frontal and sensory cortexes with repetitive transcranial stimulation impairs performance on a working memory task in humans (Zanto et al. 2011).

Third, neural responses in the FWMM are similar to experimental observations: Neurons in the sensory layer show tuned responses, similar to neurons recorded in the visual cortex, while neurons in the control layer have complex, conjunctive responses, similar to recordings in the prefrontal cortex (Rigotti et al. 2013). Fourth, as in the brain, sensory networks are engaged in their own sensory domain (and not others), while control networks (such as the prefrontal cortex) are engaged in all sensory domains (Christophel et al. 2017, Postle et al. 2003). Finally, memories in the network drift slowly over time, accruing error. This matches experimentally observed drift in behavioral responses (discussed further in the next section) (Bays et al. 2009, Wimmer et al. 2014).

It is important to note that there are biological mechanisms for generating random connections in the brain. For example, protocadherins are cell-adhesion molecules that help determine synaptic connectivity (Zipursky & Sanes 2010). They are recombined in individual neurons, giving each neuron a random, unique barcode that is matched to other neurons to determine connections (and avoid self-connections). In this way, neurons in sensory and control networks with matching barcodes could form random, reciprocal connections.

Altogether, these results provide support for the FWMM. However, more work is needed to test predictions specific to this model (some of which have been detailed in previous manuscripts;

see, e.g., Bouchacourt & Buschman 2019). One of the strongest predictions of the model is that the limited capacity of working memory is due to interference in the control network.

INTERFERENCE LIMITS THE CAPACITY OF WORKING MEMORY

As noted in the Introduction, working memory has a surprisingly small capacity—the average adult human can hold only three to four items in working memory at once (Cowan 2001, Luck & Vogel 1997). Given the importance of working memory to cognition, why is it that working memory is so limited? In this section, I argue that the capacity limitations are a result of interference between representations when they converge onto a shared representation in the prefrontal and parietal cortexes. Thus, while convergence facilitates flexibility, it also limits the capacity of working memory, suggesting that there is a necessary trade-off between these two characteristics of working memory.

Interference in working memory falls into two broad categories (**Figure 2*a***). First, overlap in the representation of two items held in working memory can cause representational interference. Second, items compete for representation in the same neural population, leading to a competitive interference (even if representations are not overlapping). In this section, I discuss both forms of interference in working memory and how they arise from the integration of memories in a single network.

Interference from Overlapping Representations

Interference can occur when two (or more) items have overlapping representations in working memory. Such representational interference impairs the ability to read out information about an item. This is schematized in **Figure 2*b***. In this case, as in integrative regions like the prefrontal cortex, the identities of two different items are represented in the activity of the whole population of neurons. Specifically, the identity of each item is represented in an encoding subspace in the N-dimensional space of activity in the neural population; this is schematized in **Figure 2*b*** as an axis encoding a 1D variable (e.g., size, luminance, etc.). Therefore, to read out the identity of an item, one simply has to project the activity of the neural population onto its encoding axis

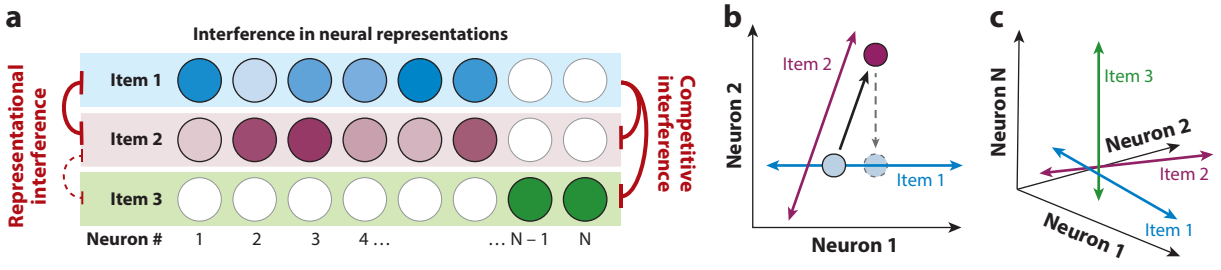

Figure 2

Interference between neural representations. (*a*) Schematic of representational and competitive interference. Representational interference occurs when two representations overlap in neural space, as between items 1 and 2 but not items 2 and 3. Competitive interference occurs between any representations in the same region, regardless of their overlap, as between item 1 and items 2 and 3. (*b*) Representational interference can induce changes in the content of memories. When the identity of item 1 is remembered alone (*blue dot*), it can be decoded by projecting onto the item 1 encoding subspace (*blue axis*). Adding a second item to memory that overlaps with item 1 causes representational interference—movement of the population vector in item 2's encoding subspace (moving along the *purple axis* to the *purple dot*) induces a change in the decoded value of item 1 (*dashed blue dot*). This is due to alignment of the encoding subspaces. (*c*) Representations in independent neural populations (item 3) exist in a subspace that is orthogonal to the other items (*green arrow*), avoiding representational interference (although still causing competitive interference, as shown in panel *a*).

(**Figure 2*b***). If the encoding axes of two items are orthogonal, then the neural population can represent both items at the same time, without interference, as the identity of one item does not affect the identity of the other item. However, overlap in the neural representation of two items, such as when the representations of two items are correlated, will cause the encoding axes of the items to align (i.e., the angle between the axes will be acute). In this case, the two items interfere: Changing the activity of the neural population to update the identity of one of the items will also change the representation of the other item (**Figure 2*b***). In this way, overlapping representations cause interference.

The effect of representational interference is that memories change over time. For example, in models of attractor networks, two memories that are similar to one another tend to merge to a single, intermediate value (e.g., red and yellow merge to orange; this occurs when representations fall within the extent of lateral excitatory connections) (Compte et al. 2000). In contrast, more dissimilar memories tend to repel one another (as they are pushed away by lateral inhibition). A similar effect is seen in the FWMM when the sensory layer is a two-dimensional surface (e.g., representing location in the visual field) (Bouchacourt & Buschman 2019). This is consistent with experiments—when subjects are asked to remember multiple spatial locations simultaneously, nearby locations are attracted toward one another, while distant locations repulse one another (Almeida et al. 2015). Similar effects are also seen across trials—the residual memory of the saccade location on the previous trial influences the memory of the saccade location on the current trial (Papadimitriou et al. 2015), reflected in the response of neurons in the prefrontal cortex (Papadimitriou et al. 2017).

In addition to impairing decoding of items from working memory, representational interference can also impair learning. Correlations in the representation of two items will cause learning signals to affect both items, which can lead to spurious learning or, in the worst-case scenario, catastrophic forgetting (Duncker et al. 2021). For example, learning an association between one item and a response could cause a spurious association to be formed with an interfering item. If that item was previously associated with a different response, then the spurious learning could overwrite the old association.

Competitive Interference Causes Regularization of Neural Responses

A second form of competitive interference occurs whenever multiple items are represented within a single neural population. Classically, this has been observed in the domain of visual processing, where increasing the number of objects in a visual scene causes an increase in interference between the representations of the objects. This interference is reflected in behavior, such as when subjects are blind to large changes in a visually crowded scene (Whitney & Levi 2011), and in neural activity, such as when neural responses are reduced when a stimulus is presented with other stimuli (Reynolds et al. 1999).

Competitive interference is also seen in working memory (Bays 2015, Kamiński et al. 2017). For example, when monkeys were asked to remember a single item alone, neurons in the prefrontal cortex strongly represented the identity of the item. However, as animals were asked to remember a second or third item, the neural response to the first item decreased (**Figure 3*a***, *top*). This reduction in signal led to a reduction in information about the first item as more items were remembered (Buschman et al. 2011) (**Figure 3*a***, *bottom*). Similar effects have been seen in humans, where increasing the number of items held in working memory decreased the decodability of each individual item (Sprague et al. 2014). Together, these results suggest that competitive interference acts on the representation of items in working memory, limiting the total information held in working memory.

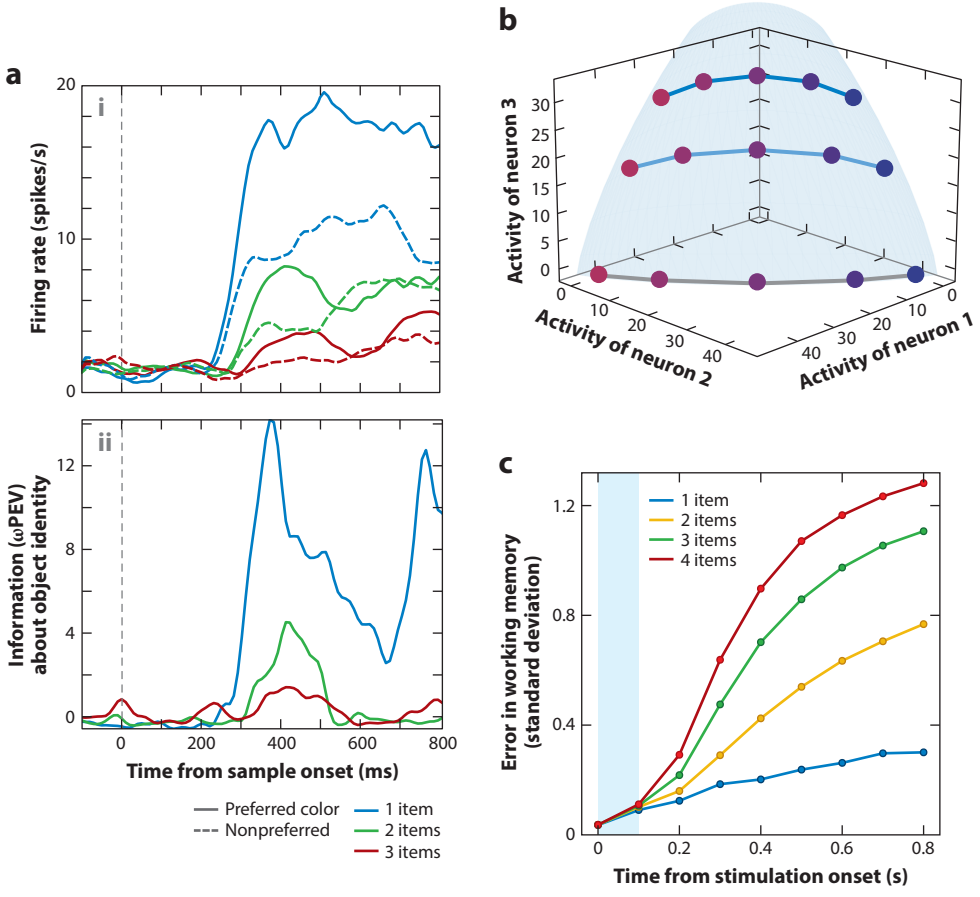

Figure 3

Competitive interference limits working memory capacity. (*a*) Increasing the number of items held in working memory causes interference, reducing the information about each item. (*i*) Firing rate response of a neuron in the prefrontal cortex to remembering an item in working memory by itself (*blue*) or with one (*green*) or two (*red*) other items. Increasing the number of items in working memory reduced the overall activity of the neuron, as well as the separability in response to the preferred (*solid*) and nonpreferred (*dashed*) stimulus. (*ii*) The suppression of activity reduced the amount of information about the item in memory. Panel adapted with permission from Buschman et al. (2011). (*b*) Geometric intuition of divisive normalization. Shown is neural space, defined as the activity of three schematic neurons. Because of divisive normalization, the population activity is constrained to a region below a hypersurface (*light blue*). The hypersurface captures the response of the neural population to a varying stimulus input into neurons 1 and 2 (*colored dots*) and a second stimulus input delivered to neuron 3 (increasing in contrast from no stimulus to a full-contrast stimulus along the *z* axis). The response of neurons was modeled based on the normalization function estimated by Busse et al. (2009). The separability of information about the first stimulus encoded in the first two neurons (schematized as the separation of dots) is high when the second stimulus is absent (neuron 3 is zero; *gray line*). Encoded information is reduced when the second stimulus is stronger (i.e., when the activity of the third neuron is high, the separability of the dots is reduced; *blue lines*). (*c*) Error in memory increases as a function of the number of items in working memory in the flexible working memory model (FWMM). Error accumulates over time, accumulating more quickly with more items in working memory. Panel adapted with permission from Bouchacourt & Buschman (2019).

The competition between neural representations is well-explained by a divisive normalization model. The normalization model hypothesizes that there is an upper bound on the total neural activity in the network at any moment in time (Busse et al. 2009, Carandini & Heeger 2012). To implement normalization, the activity of each individual neuron is divided by the integrated activity of all neurons (thus, all responses are normalized). In neural circuits, normalization is thought to be mediated by inhibitory interneurons, which integrate local excitatory drive and provide lateral inhibition of activity (Wilson et al. 2012). From a normative perspective, divisive normalization may be important for restricting neural activity to a domain that facilitates behavior, decoding, and learning (Carandini & Heeger 2012).

The neural mechanisms underlying normalization in working memory are likely the same as those for sensory processing: Convergent inputs into a single brain region lead to competitive interference. However, while classic models of divisive normalization rely on algebraic normalization (i.e., dividing neural activity by a function of the summed activity), this is not the only mechanism that can normalize neural activity. As seen in the FWMM, balanced excitation and inhibition can achieve similar regularization of neural responses. Recent work has found balance in excitation/inhibition (E/I) is important to avoid runaway excitation (or collapsing inhibition) (Fritschy 2008) and maximize the coding efficiency of neural responses (Denève & Machens 2016). In the FWMM, the result of E/I balance is that, as the number of to-be-remembered items is increased, there is a balanced increase in both excitatory and inhibitory drive from the sensory networks, leading to divisive-normalization-like regularization of neural activity in the control network.

Normalization of neural representations is thought to impact working memory in several ways. First, it reduces the separability of memory representations, making it more difficult to discriminate memories in the control network. This is schematized in **Figure 3*b***. In this case, the activity of two neurons (neurons 1 and 2) encodes the identity of an item in memory, with the spectrum of possible stimulus identities shown in the array of colored dots (**Figure 3*b***). When a single item is held in working memory, the possible identities of that item are well-separated, facilitating accurate discrimination. However, when another item is also maintained in working memory, the activity of other neurons increases (shown as an increase in the activity of neuron 3). Due to normalization, the activity of the neural population is constrained to a region of neural space (**Figure 3*b***). This constraint means that the increased activity of other neurons causes the range of the neural response to the first item to decrease, leading to decreased discriminability (**Figure 3*b***). This matches experimental observations that information about an item decreases as the number of items in working memory increases (Buschman et al. 2011, Sprague et al. 2014) (**Figure 3*a***). Note that the reduction in separability is not due to direct overlap in the neural representations, and the effect does not depend on the decoder (i.e., there is no projection of the representation that would eliminate the effect of normalization).

The second effect of normalization is that it reduces the strength of neural activity in integrative regions, such as the prefrontal and parietal cortices. This, in turn, reduces the strength of top-down feedback onto sensory regions. As these feedback connections help sustain the memory representation in the sensory cortex, a decrease in feedback causes the magnitude of the memory representation to be reduced, making it more susceptible to noise and allowing the memory to drift and/or be lost (Edin et al. 2009). Again, these effects are encapsulated in the FWMM. Increasing the number of items in working memory leads to normalization in the control layer (due to E/I balance), which reduces feedback onto the sensory layer. This increases the effect of noise in the memories as neural responses become weaker and more sporadic in the sensory subnetworks. Increased noise leads to a concomitant increase in the drift of the memory within sensory space (**Figure 3*c***). This increase in drift matches the experimentally

observed increase in error with working memory load—when more items are held in working memory, the error distribution of the reported memory gets wider (Bays et al. 2009), and the accumulation of error is faster (Pertzov et al. 2017).

Furthermore, in the model, if the drop in feedback is severe enough, then the memory representation can be degraded to such an extent that it is no longer sustained (i.e., the memory is lost). This is consistent with experiments showing that the likelihood of forgetting an object increases as memory load increases (Adam et al. 2017, Luck & Vogel 1997). Altogether, these results suggest that the limited capacity of working memory reflects representational and competitive interference between representations, likely due to convergence of sensory inputs into associative regions.

EXPANDING THE MIND: OVERCOMING INTERFERENCE TO MAXIMIZE THE CAPACITY OF WORKING MEMORY

The limited capacity of working memory limits cognition. This is reflected in the fact that an individual's working memory capacity is highly correlated with their general fluid intelligence (Conway et al. 2003, Engle et al. 1999). In many ways, this makes sense—the more items you can hold in mind, the more you can manipulate and combine them to find novel solutions to a problem. Given that interference is the main constraint on the capacity of working memory, one might expect the brain to develop mechanisms to mitigate the effect of interference and, by doing so, expand the capacity of the mind. In this section, I discuss several of these mechanisms.

Reducing Interference by Increasing Sparsity

One way to minimize interference is to increase the sparsity of neural representations. The sparsity of a neural representation is a measure of how many neurons are involved in the representation (e.g., high sparsity means that very few neurons are involved). Increasing sparsity can reduce both competitive and representational interference (Ganguli & Sompolinsky 2012). First, as sparsity decreases the total amount of activity needed to represent a stimulus, this decrease reduces normalization. Second, as the representation becomes sparser, there is less opportunity for overlap with other representations, and therefore, the degree of representational interference decreases. This is especially true if the response of specialized neurons is sparse across the domain of all possible sensory stimuli (i.e., these neurons only respond to a few stimuli; in the limit, this would yield so-called grandmother cells).

The neural architecture of a region influences the sparsity of representations, suggesting that some brain regions may be better suited for avoiding interference than others (Norman & O'Reilly 2003). Sparsity may also be increased with experience. In visual perception, the first time that an object is seen, its representation is distributed across the population of neurons in the visual cortex, with many of the neurons responding to the stimulus to a moderate degree. However, over multiple exposures, experience increases the sparsity of neural representations: Most neurons reduce their response, while a few specialized neurons become highly responsive (Freedman et al. 2006, Lim et al. 2015, Woloszyn & Sheinberg 2012).

It is likely that the change in sparsity of visual representations extends to memory representations. If this is true, then the familiarity of an object should make it easier to remember the object. Indeed, subjects have better working memory for the faces of famous people compared to unfamiliar faces (Jackson & Raymond 2008). Similarly, visual working memory encoding is faster, and capacity limitations are higher, for letters drawn from a familiar alphabet compared to those from a nonfamiliar alphabet (Ngiam et al. 2019). Recently, it was found even limited experience with a stimulus can improve working memory performance (Brady et al. 2009). In a classic working

memory paradigm, subjects were asked to report the color of one of several items. Importantly, each item consisted of a pair of colors that formed a single bullseye item (i.e., one color in the center and another color in the surrounding annulus). Critically, some of the bullseye items had consistent colors, such that there was a statistical regularity between the inner and outer color. One way to conceptualize this is that the representation could be compressed by forming a single object representation that captured the inner and outer color together (e.g., object A has a red center and green annulus) (Quiroga et al. 2008). As expected, working memory performance for these objects was better than for other objects. However, even when subjects were tested on the color of other items in the display, the presence of a regularly occurring object improved memory performance. This suggests that the grouped colors, acting as a single object, reduced the overall load on working memory, improving performance for all of the items in memory.

Altogether, these results are consistent with the idea that experience with visual objects creates a sparse representation. The sparsity of this representation means that it is less likely to cause interference, improving working memory performance.

Reducing Interference by Rotational Dynamics

One of the most prevalent forms of representational interference is between items in working memory and new sensory inputs. As noted above, recent experimental work has found that working memory engages the sensory cortex, which is likely important for providing structure to memory representations (Christophel et al. 2017, Harrison & Tong 2009, Zhang et al. 2019). However, this poses a problem—new sensory inputs will engage the same sensory regions that hold the memory of previous sensory inputs. This could lead to catastrophic interference, as the new stimulus representation will directly overlap with the memory representation, causing the memory to change or be lost. Behaviorally, working memory is largely robust to such interference (although distracting stimuli presented in the memory delay can impact working memory performance; Soto et al. 2011). How then does the brain avoid such catastrophic interference? One possibility is that memories are represented in a form that avoids overlap with sensory inputs.

Recent work has provided support for such an idea, showing that sensory representations dynamically transform into an independent, orthogonal memory representation. Using a cross-temporal classification analysis, a study in monkeys found that the representation of an item in the prefrontal cortex changes over time (Stokes et al. 2013). As seen in **Figure 4*a***, the representation of a stimulus when the animal was encoding it was not related to the representation of the memory of the stimulus. This suggests that sensory representations dynamically transform into an independent, orthogonal memory representation. More recently, a similar effect was seen in the auditory cortex of mice (Libby & Buschman 2021). When mice were presented with a sequence of sounds, auditory cortex neurons represented both the immediate sensory input and the memory of recent sounds. Sensory representations were prone to interference—new inputs overwrote the representation of the previous sensory input. However, the memory representation of previous sounds avoided interference from new sensory inputs because it was stored in a separate memory subspace in the high-dimensional space of neural activity, which was orthogonal to the sensory subspace. These results are also consistent with previous work that has shown two orthogonal subspaces for preparing and executing motor movements (Kaufman et al. 2014). In this study, the prospective working memory representation of an upcoming movement (i.e., the motor plan) was represented in one subspace, and then, when the animal began to engage in the motor movement itself (i.e., the memory became active), the representation rotated into the second subspace for engagement.

There are several different circuit mechanisms that might generate orthogonal sensory and memory representations. First, the representations could be orthogonal because they are

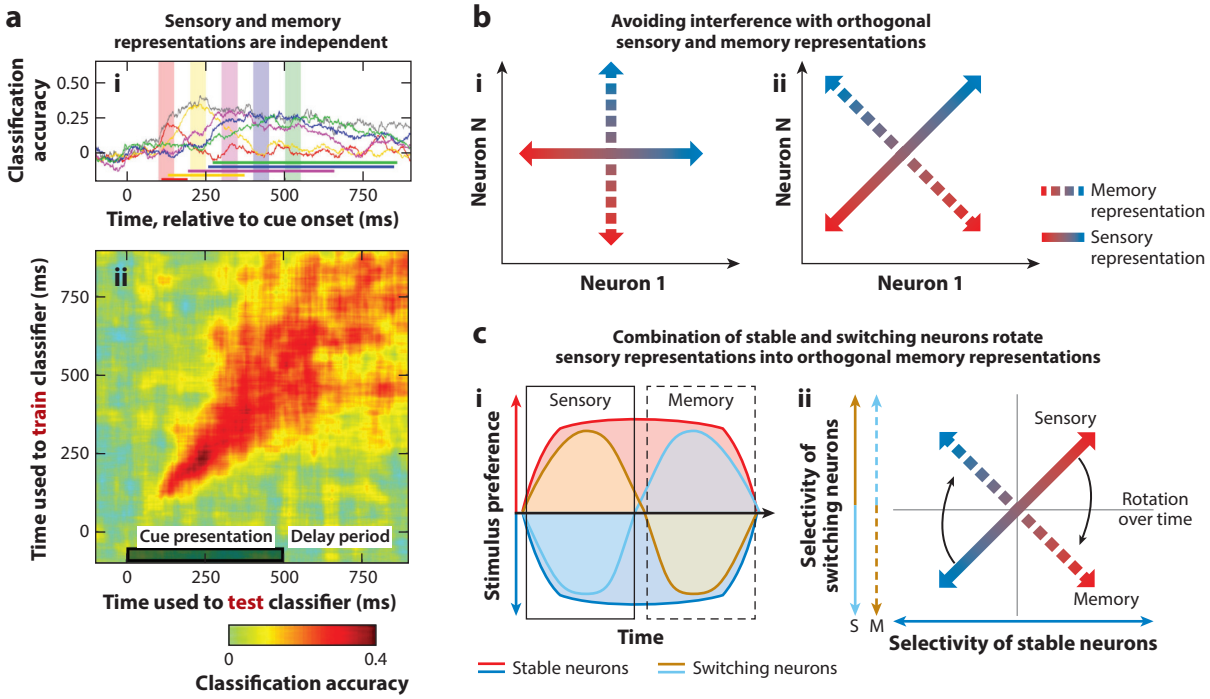

Figure 4

Sensory and memory representations are orthogonal. (*a*) Stokes et al. (2013) trained classifiers to decode which stimulus was presented, based on the activity of prefrontal cortex neurons. Different classifiers were trained at different time periods (*i*: *shaded regions*; *ii*: *vertical axis*) and then tested on all time periods (*horizontal axes*). Cross-temporal classification performance (*i*: *vertical axis*; *ii*: *color axis*) showed that classifiers defined during sensory processing did not decode memories (and vice versa). This suggests that sensory and memory representations are orthogonal. Panel adapted from Stokes et al. (2013). (*b*) Mechanisms for generating orthogonal representations. Sensory and memory representations could be stored in independent populations of neurons (*i*) or in orthogonal dimensions within the same population of neurons (*ii*). (*c*, *i*) Recordings from the auditory cortex in mice found two populations of neurons: stable neurons that maintained their stimulus preference throughout the sensory and memory time periods (*red* and *blue lines*) and switching neurons that inverted their selectivity from the sensory to the memory time periods (*gold* and *light blue lines*). (*ii*) The combination of stable and switching neurons facilitated a rotation of the sensory representation (*solid arrow*) into a memory representation (*dashed arrow*). This was due to the inversion of selectivity of the switching neuron (*y* axis), which caused the population representation to rotate.

represented in independent populations of neurons (**Figure 4*b***, *left*). This would reduce the amount of interference between representations, as they would lie in different dimensions in neural space (as seen in **Figure 3*c***; note that this is similar to increasing sparsity). Second, the sensory and memory representations could exist within the same population of neurons, but sensory representations could be dynamically transformed into an orthogonal memory representation (**Figure 4*b***, *right*). One possible mechanism for this transformation is to have random representations for both sense and memory. As noted above, random vectors are largely uncorrelated with one another in high-dimensional spaces and, therefore, would allow for orthogonal representations of sensory and memory. Indeed, such a mechanism underlies the FWMM detailed above—the randomness of connections between different sensory subnetworks means that they are generally orthogonal and, therefore, have little interference between subnetworks.

Electrophysiological recordings from the auditory cortex in mice suggested another mechanism—memories transform through structured rotation (Libby & Buschman 2021). In this case, the rotation of a sensory representation into an orthogonal memory representation

was facilitated by two types of neurons within the sensory cortex: stable neurons, which maintained their stimulus preference across sensory and memory, and switching neurons, which inverted their stimulus preference from sensory to memory. As schematized in **Figure 4*c***, the combination of these two subpopulations of neurons led to the rotation of the representation, allowing it to get out of the way of new sensory inputs.

Reducing Noise and Interference by Dynamically Changing Memories

In addition to helping avoid interference from sensory inputs, dynamics in the neural representation can also mitigate the effect of noise. Noise in neural representations is thought to cause memories to diffuse randomly away from their initial value over time, which leads to error in the memory report (Bays 2015, Wimmer et al. 2014). However, memories also show systematic drift. For example, when subjects are asked to remember a color, such as pink, this memory will drift over time toward a more canonical color, such as red (Bae et al. 2014, Taylor & Bays 2018). Recent work found that this drift could be explained by attractor dynamics within mnemonic space—i.e., there is an attractor at red that pulls pink toward it over time (Panichello et al. 2019). While these attractor dynamics induce a systematic error into memory reports by pulling memories to an attractor, they can also mitigate the effect of diffusion: Once a memory representation is at an attractor, the effect of random diffusion away from the attractor will be reduced (as the memory will consistently drift back toward the attractor). Importantly, behavioral studies showed that the locations of attractors in mnemonic space were not fixed; instead, they adapted to the statistics of the environment such that they moved to the locations of regularly occurring stimuli. This meant that attractors pulled memories toward statistically likely values (as if dynamics were integrating prior beliefs into the posterior over time). In this way, attractor dynamics in mnemonic space can reduce the total amount of error in working memory (by optimizing systematic drift and minimizing random diffusion). More recently, functional imaging found that drift and diffusion were represented in different brain regions, suggesting that they may have different underlying mechanisms (Yu et al. 2020).

Dynamics in representations can also mitigate interference between two items held in working memory. As detailed above, when multiple items are held in working memory, normalization causes the items to compete for representation. One way to reduce this competitive interference is to reduce the total number of neurons involved in their (combined) representation. While this can occur through sparsification (see above), memories may also evolve over time such that their representations overlap, reducing the total number of neurons involved in the representation (Bouchacourt & Buschman 2019). Note that, while these dynamics reduce competitive interference, they also increase representational interference.

These results suggest that several different forces act on memory representations, causing them to change over time: A diffusive force reflects noise in neural activity; a drift force integrates expectations into memories; and competitive interference causes memories to drift toward sparser, overlapping representations. In the end, memories likely evolve in a way that balances all of these forces.

Multiple Working Memory Systems Optimize the Trade-Off in Flexibility and Capacity

As discussed above, convergence is important for the flexibility of working memory, yet it also leads to the interference that limits the capacity of working memory. This framework then suggests that there is a trade-off between flexibility and capacity. Working memory systems designed

to maximize flexibility rely on integrative connections, which come at the cost of interference and a limited capacity. In contrast, a capacity-maximizing system should limit interference. Above, I focus on mechanisms that minimize interference in an integrative network. However, another way to limit interference would be to build specialized circuits for maintaining specific information. When independent circuits are constructed, these items would have no competitive or representational interference with other memories (regardless of whether these other memories were maintained in other independent circuits or in a distributed, flexible memory system). However, as detailed below, such dedicated circuits are expensive, so this solution may be reserved for maintaining information that is critical to behavior.

Several lines of experimental evidence point to dedicated memory networks for task-critical information. For example, one type of critical information for visual perception is maintaining a representation of the location of attention (Buschman & Kastner 2015). In many experiments, the location of attention is constrained to a series of locations around a circle. Thus, it has been well-modeled by a dedicated ring attractor network (Compte et al. 2000). Ring attractor networks consist of a single layer, with a population of neurons arranged in a ring according to their preference for a one-dimensional, circular variable (e.g., the angular location of attention). Inputs into a neuron in the ring are sustained via local, recurrent, excitatory connections with other neurons at nearby locations on the ring. Inhibitory connections to distal neurons in the ring help to stabilize memories through surround suppression (as described above for the FWMM). In this way, the ring attractor can represent circular variables, such as the angular location of attention, in the position of a bump of activity around the ring and sustain this representation over a memory delay.

The ring attractor network captures many of the experimental observations of how neurons represent the location of sustained attention. As in the ring attractor, neurons in the prefrontal and parietal cortexes are tuned to the location of attention and sustain this representation during a memory delay (Bisley & Goldberg 2003, Funahashi et al. 1989; although see Stokes 2015 for recent work that has questioned whether memory representations are actually sustained). In addition, trial-by-trial variability in the location of attention is well-modeled by memories diffusing along the ring attractor (Wimmer et al. 2014).

Evidence for specialized memory networks, such as ring attractor networks, is not limited to primates. Electrophysiological recordings in the subiculum and entorhinal cortex of mice find neurons that respond to the angular direction of the animal's head (Taube & Bassett 2003). Similar representations are also found in the ellipsoid body of the *Drosophila* fruit fly (Kim et al. 2017). As heading direction is also a circular variable, it can be represented by a ring attractor. Strikingly, as predicted by the model, neurons in the ellipsoid body of *Drosophila* are arranged in a physical ring, with local recurrent connectivity sustaining representations (Kim et al. 2017).

These results suggest that specialized networks for behaviorally critical information may be conserved across evolution, driving the development of specialized circuitry to sustain representations of the location of attention or heading direction. The dedication of specialized circuits to sustain this information may also help to avoid interference with other memories and, thus, avoid limits in the capacity of working memory. Indeed, you would not want to forget what direction you were heading in, or where you were attending, because you were also trying to remember an email address.

The success of specialized networks in maintaining memories begs the question of why all memories are not stored in independent attractor networks. The largest drawback to such a system is that it would require dedicated circuits for each type of information that one wants to remember. This is exemplified in the ring attractor—the structure of the network is designed to maintain information about circular variables and nothing else. While it may be reasonable to construct specialized networks for a few key cognitive variables that are conserved across evolution, this

approach begins to break down as you increase the variety of items that you want to hold in mind (e.g., numbers, shopping lists, etc.). As representations become more complicated (e.g., visual objects, like a pink elephant), it also becomes more difficult to construct a continuous, balanced network to maintain these representations.

In addition, specialized systems, such as ring attractor networks, still have fixed capacity limitations, as they do not avoid interference within the same network. For this reason, these networks are typically designed to hold a single item at a time. In most cases, this is not a problem—for example, your head can only be directed to a single location or direction at a time. However, if you wanted to store more than one item (e.g., the direction of both your attention and your head), then this would require duplicating the network (or suffering the same interference seen in the flexible memory system). Therefore, while multiple specialized memory systems would allow you to store each item without interference, they would also require an inordinate amount of circuitry dedicated to remembering each instance of all the possible types of information. In addition, these circuits would have to be learned, as many cognitive variables (such as letters) are not evolutionarily defined.

Given the difficulty in building an efficient working memory system from specialized networks alone, the brain likely developed a secondary, flexible system for working memory. This system relied on the structure of sensory systems combined with an integrative prefrontal or parietal cortex to maintain arbitrary inputs, yet with a limited capacity.

CONCLUSION

Working memory is at the core of cognition, acting as a flexible workspace on which thoughts can be held, manipulated, and used to guide behavior. In this review, I present experimental and theoretical evidence that the flexibility of working memory emerges from the interaction of high-dimensional representations in associative regions with structured representations in the sensory cortex. However, this flexibility comes at a cost. The convergence of inputs onto associative regions causes interference between memories, limiting the number of items that one can simultaneously maintain in working memory.

Given the importance of working memory to cognition, the brain should minimize the effect of interference. In this review, I discuss four ways in which interference can be reduced. First, neural representations become sparser with experience, reducing interference. Second, sensory representations rotate over time, moving into an orthogonal memory subspace that avoids interference with new sensory inputs. Third, dynamics of memories within mnemonic space can mitigate the effect of noise and reduce interference between memories. Fourth, I review evidence that evolutionarily critical information may be maintained in dedicated networks. Together, these mechanisms optimize the use of working memory while still allowing for the flexibility of cognition.

There are several open questions remaining (see the Future Issues). In particular, while this review focuses on visual working memory, future work is needed to test whether the principles and mechanisms described in this review apply to working memory in general, including working memory for other sensory domains, for motor actions, or for episodic memories. As noted above, the FWMM may provide a framework for these other domains. In addition, while this review focuses on the flexibility and limited capacity of working memory, there are other characteristics of working memory that are not discussed. For example, working memory is tightly controlled to maximize the use of this limited resource (Gazzaley & Nobre 2012, Hochreiter & Schmidhuber 1997, Vogel et al. 2005). Future work is needed to understand how working memory is controlled and how this control supports its flexibility and reduces interference.

SUMMARY POINTS

1. Working memory is flexible, able to hold anything. This flexibility arises from the interaction between high-dimensional, integrative representations in the prefrontal cortex and structured representations in the sensory cortex.
2. While the convergence of memory representations in the prefrontal cortex supports the flexibility of working memory, it also causes interference between items, limiting the number of items that can be remembered simultaneously.
3. The brain maximizes the efficiency of working memory by reducing interference between memory representations in several ways, including increasing the sparsity of memory representations to reduce overlap and dynamically transforming memory representations to orthogonal subspaces.
4. There is a trade-off between the flexibility and capacity of working memory; constructing specialized networks for working memory can reduce interference and increase capacity, but this necessarily reduces the flexibility of these networks.

FUTURE ISSUES

1. In this review, I focus on visual working memory. Future work is needed to understand how the principles discussed above extend to other sensory domains (auditory, olfactory, etc.), to motor movements, and to working memory for episodic memories.
2. To efficiently use working memory, what gets into and out of working memory is tightly controlled (Gazzaley & Nobre 2012). Future work is needed to understand the mechanisms controlling working memory, particularly how a controller may be flexible enough to act on arbitrary representations.
3. Oscillations are thought to flexibly couple neural populations representing the contents of working memory (Buschman et al. 2012, Salazar et al. 2012) and reduce interference by temporally separating working memory representations (Siegel et al. 2009). Future work is needed to relate oscillations to the attractor dynamics models of working memory discussed in this review.
4. Although there is evidence for biological mechanisms that can create random connections (e.g., protocadherins), more research is needed to understand the mechanisms and rules that define connections between neurons and how these might constrain models of working memory.
5. The flexible working memory model, like all attractor networks, relies on balanced, reciprocal connections. Future work is needed to understand how learning, whether unsupervised experiential learning or supervised associative or reward learning, can create or shape these connections without disrupting the function of the network.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Thanks to Flora Bouchacourt, Caroline Jahn, Alex Libby, Matthew Panichello, Eleni Papadoyannis, Motoaki Uchimura, and Sarah Henrickson for helpful conversations. This work is supported by National Institutes of Mental Health grant R01MH115042.

LITERATURE CITED

Adam KCS, Vogel EK, Awh E. 2017. Clear evidence for item limits in visual working memory. *Cogn. Psychol.* 97:79–97

Almeida R, Barbosa J, Compte A. 2015. Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. *J. Neurophysiol.* 114(3):1806–18

Angelucci A, Bressloff PC. 2006. Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate V1 neurons. *Prog. Brain Res.* 154:93–120

Arcizet F, Mirpour K, Bisley JW. 2011. A pure salience response in posterior parietal cortex. *Cereb. Cortex* 21(11):2498–506

Bae G-Y, Olkkonen M, Allred SR, Wilson C, Flombaum JI. 2014. Stimulus-specific variability in color working memory with delayed estimation. *J. Vis.* 14(4):7

Barak O, Rigotti M, Fusi S. 2013. The sparseness of mixed selectivity neurons controls the generalization-discrimination trade-off. *J. Neurosci.* 33(9):3844–56

Barone P, Joseph J-P. 1989. Prefrontal cortex and spatial sequencing in macaque monkey. *Exp. Brain Res.* 78(3):447–64

Bays PM. 2015. Spikes not slots: noise in neural populations limits working memory. *Trends Cogn. Sci.* 19(8):431–38

Bays PM, Catalao RFG, Husain M. 2009. The precision of visual working memory is set by allocation of a shared resource. *J. Vis.* 9(10):7

Bisley JW, Goldberg ME. 2003. Neuronal activity in the lateral intraparietal area and spatial attention. *Science* 299(5603):81–86

Bouchacourt F, Buschman TJ. 2019. A flexible model of working memory. *Neuron* 103(1):147–160.e8

Brady TF, Konkle T, Alvarez GA. 2009. Compression in visual working memory: using statistical regularities to form more efficient memory representations. *J. Exp. Psychol. Gen.* 138(4):487–502

Buschman TJ, Denovellis EL, Diogo C, Bullock D, Miller EK. 2012. Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. *Neuron* 76(4):838–46

Buschman TJ, Kastner S. 2015. From behavior to neural dynamics: an integrated theory of attention. *Neuron* 88(1):127–44

Buschman TJ, Siegel M, Roy JE, Miller EK. 2011. Neural substrates of cognitive capacity limitations. *PNAS* 108(27):11252–55

Busse L, Wade AR, Carandini M. 2009. Representation of concurrent stimuli by population activity in visual cortex. *Neuron* 64(6):931–42

Carandini M, Heeger DJ. 2012. Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* 13(1):51–62

Christophel TB, Klink PC, Spitzer B, Roelfsema PR, Haynes J-D. 2017. The distributed nature of working memory. *Trends Cogn. Sci.* 21(2):111–24

Compte A, Brunel N, Goldman-Rakic PS, Wang X-J. 2000. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb. Cortex* 10(9):910–23

Conway ARA, Kane MJ, Engle RW. 2003. Working memory capacity and its relation to general intelligence. *Trends Cogn. Sci.* 7(12):547–52

Cowan N. 2001. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.* 24(01):87–114

Denève S, Machens CK. 2016. Efficient codes and balanced networks. *Nat. Neurosci.* 19(3):375–82

Duncker L, Driscoll L, Shenoy KV, Sahani M, Sussillo D. 2021. Organizing recurrent network dynamics by task-computation to enable continual learning. *Adv. Neural Inf. Process. Syst.* 33. In press

Edin F, Klingberg T, Johansson P, McNab F, Tegnér J, Compte A. 2009. Mechanism for top-down control of working memory capacity. *PNAS* 106(16):6802–7
Enel P, Procyk E, Quilodran R, Dominey PF. 2016. Reservoir computing properties of neural dynamics in prefrontal cortex. *PLOS Comput. Biol.* 12(6):e1004967
Engle RW, Tuholski SW, Laughlin JE, Conway AR. 1999. Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *J. Exp. Psychol. Gen.* 128(3):309–31
Freedman DJ, Riesenhuber M, Poggio T, Miller EK. 2006. Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex. *Cereb. Cortex* 16(11):1631–44
Fritschy J-M. 2008. Epilepsy, E/I balance and GABAA receptor plasticity. *Front. Mol. Neurosci.* 1:5
Funahashi S, Bruce CJ, Goldman-Rakic PS. 1989. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* 61(2):331–49
Ganguli S, Sompolinsky H. 2010. Short-term memory in neuronal networks through dynamical compressed sensing. *Adv. Neural Inf. Process. Syst.* 23:667–75
Ganguli S, Sompolinsky H. 2012. Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annu. Rev. Neurosci.* 35:485–508
Gazzaley A, Nobre AC. 2012. Top-down modulation: bridging selective attention and working memory. *Trends Cogn. Sci.* 16(2):129–35
Gazzaley A, Rissman J, D'Esposito M. 2004. Functional connectivity during working memory maintenance. *Cogn. Affect. Behav. Neurosci.* 4(4):580–99
Harrison SA, Tong F. 2009. Decoding reveals the contents of visual working memory in early visual areas. *Nature* 458(7238):632–35
Hasselmo ME, Stern CE. 2006. Mechanisms underlying working memory for novel information. *Trends Cogn. Sci.* 10(11):487–93
Higgins I, Chang L, Langston V, Hassabis D, Summerfield C, et al. 2020. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal neurons. arXiv:2006.14304 [q-bio]
Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–80
Hubel DH, Wiesel TN. 1959. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* 148(3):574–91
Isseroff A, Rosvold HE, Galkin TW, Goldman-Rakic PS. 1982. Spatial memory impairments following damage to the mediodorsal nucleus of the thalamus in rhesus monkeys. *Brain Res.* 232(1):97–113
Jackson MC, Raymond JE. 2008. Familiarity enhances visual working memory for faces. *J. Exp. Psychol. Hum. Percept. Perform.* 34(3):556–68
Kamiński J, Sullivan S, Chung JM, Ross IB, Mamelak AN, Rutishauser U. 2017. Persistently active neurons in human medial frontal and medial temporal lobe support working memory. *Nat. Neurosci.* 20(4):590–601
Kaufman MT, Churchland MM, Ryu SI, Shenoy KV. 2014. Cortical activity in the null space: permitting preparation without movement. *Nat. Neurosci.* 17(3):440–48
Kim SS, Rouault H, Druckmann S, Jayaraman V. 2017. Ring attractor dynamics in the *Drosophila* central brain. *Science* 356(6340):849–53
Lee H, Simpson GV, Logothetis NK, Rainer G. 2005. Phase locking of single neuron activity to theta oscillations during working memory in monkey extrastriate visual cortex. *Neuron* 45(1):147–56
Levy R, Friedman HR, Davachi L, Goldman-Rakic PS. 1997. Differential activation of the caudate nucleus in primates performing spatial and nonspatial working memory tasks. *J. Neurosci.* 17(10):3870–82
Libby A, Buschman TJ. 2021. Rotational dynamics reduce interference between sensory and memory representations. *Nat. Neurosci.* 24:715–26
Liebe S, Hoerzer GM, Logothetis NK, Rainer G. 2012. Theta coupling between V4 and prefrontal cortex predicts visual short-term memory performance. *Nat. Neurosci.* 15(3):456–62
Lim S, McKee JL, Woloszyn L, Amit Y, Freedman DJ, et al. 2015. Inferring learning rules from distributions of firing rates in cortical neurons. *Nat. Neurosci.* 18(12):1804–10
Liu LD, Miller KD, Pack CC. 2018. A unifying motif for spatial and directional surround suppression. *J. Neurosci.* 38(4):989–99
Luck SJ, Vogel EK. 1997. The capacity of visual working memory for features and conjunctions. *Nature* 390(6657):279–81

Markov NT, Ercsey-Ravasz M, Essen DCV, Knoblauch K, Toroczkai Z, Kennedy H. 2013. Cortical high-density counterstream architectures. *Science* 342(6158):1238406
Meyers EM, Qi X-L, Constantinidis C. 2012. Incorporation of new information into prefrontal cortical activity after learning working memory tasks. *PNAS* 109(12):4651–56
Miller EK, Cohen JD. 2001. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24:167–202
Miller EK, Li L, Desimone R. 1991. A neural mechanism for working and recognition memory in inferior temporal cortex. *Science* 254(5036):1377–79
Ngiam WXQ, Khaw KLC, Holcombe AO, Goodbourn PT. 2019. Visual working memory for letters varies with familiarity but not complexity. *J. Exp. Psychol. Learn. Mem. Cogn.* 45(10):1761–75
Norman KA, O'Reilly RC. 2003. Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychol. Rev.* 110(4):611–46
Panichello MF, DePasquale B, Pillow JW, Buschman TJ. 2019. Error-correcting dynamics in visual working memory. *Nat. Commun.* 10(1):3366
Papadimitriou C, Ferdoash A, Snyder LH. 2015. Ghosts in the machine: memory interference from the previous trial. *J. Neurophysiol.* 113(2):567–77
Papadimitriou C, White RL, Snyder LH. 2017. Ghosts in the machine II: neural correlates of memory interference from the previous trial. *Cereb. Cortex* 27(4):2513–27
Pasternak T, Greenlee MW. 2005. Working memory in primate sensory systems. *Nat. Rev. Neurosci.* 6(2):97–107
Pasupathy A, Connor CE. 1999. Responses to contour features in macaque area V4. *J. Neurophysiol.* 82(5):2490–502
Pertzov Y, Manohar S, Husain M. 2017. Rapid forgetting results from competition over time between items in visual working memory. *J. Exp. Psychol. Learn. Mem. Cogn.* 43(4):528–36
Postle B, Druzgal T, Desposito M. 2003. Seeking the neural substrates of visual working memory storage. *Cortex* 39(4–5):927–46
Postle BR, D'Esposito M. 2003. Spatial working memory activity of the caudate nucleus is sensitive to frame of reference. *Cogn. Affect. Behav. Neurosci.* 3(2):133–44
Pribram KH, Mishkin M, Enger H, Kaplan SJ. 1952. Effects on delayed-response performance of lesions of dorsolateral and ventromedial frontal cortex of baboons. *J. Comp. Physiol. Psychol.* 45(6):565–75
Quiroga RQ, Kreiman G, Koch C, Fried I. 2008. Sparse but not "Grandmother-cell" coding in the medial temporal lobe. *Trends Cogn. Sci.* 12(3):87–91
Rahmati M, DeSimone K, Curtis CE, Sreenivasan KK. 2020. Spatially specific working memory activity in the human superior colliculus. *J. Neurosci.* 40(49):9487–95
Ranganath C. 2006. Working memory for visual objects: complementary roles of inferior temporal, medial temporal, and prefrontal cortex. *Neuroscience* 139(1):277–89
Ranganath C, D'Esposito M. 2001. Medial temporal lobe activity associated with active maintenance of novel information. *Neuron* 31(5):865–73
Reynolds JH, Chelazzi L, Desimone R. 1999. Competitive mechanisms subserve attention in macaque areas V2 and V4. *J. Neurosci.* 19(5):1736–53
Rigotti M, Barak O, Warden MR, Wang X-J, Daw ND, et al. 2013. The importance of mixed selectivity in complex cognitive tasks. *Nature* 497(7451):585–90
Salazar RF, Dotson NM, Bressler SL, Gray CM. 2012. Content-specific fronto-parietal synchronization during visual working memory. *Science* 338(6110):1097–100
Sarma A, Masse NY, Wang X-J, Freedman DJ. 2016. Task-specific versus generalized mnemonic representations in parietal and prefrontal cortices. *Nat. Neurosci.* 19(1):143–49
Siegel M, Buschman TJ, Miller EK. 2015. Cortical information flow during flexible sensorimotor decisions. *Science* 348(6241):1352–55
Siegel M, Warden MR, Miller EK. 2009. Phase-dependent neuronal coding of objects in short-term memory. *PNAS* 106(50):21341–46
Simoncelli EP, Olshausen BA. 2001. Natural image statistics and neural representation. *Annu. Rev. Neurosci.* 24:1193–216

Soto D, Greene CM, Chaudhary A, Rotshtein P. 2011. Competition in working memory reduces frontal guidance of visual selection. *Cereb. Cortex* 22(5):1159–69
Sprague TC, Ester EF, Serences JT. 2014. Reconstructions of information in visual spatial working memory degrade with memory load. *Curr. Biol.* 24(18):2174–80
Stern CE, Sherman SJ, Kirchhoff BA, Hasselmo ME. 2001. Medial temporal and prefrontal contributions to working memory tasks with novel and familiar stimuli. *Hippocampus* 11(4):337–46
Stokes MG. 2015. "Activity-silent" working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn. Sci.* 19(7):394–405
Stokes MG, Kusunoki M, Sigala N, Nili H, Gaffan D, Duncan J. 2013. Dynamic coding for cognitive control in prefrontal cortex. *Neuron* 78(2):364–75
Taube JS, Bassett JP. 2003. Persistent neural activity in head direction cells. *Cereb. Cortex* 13(11):1162–72
Taylor R, Bays PM. 2018. Efficient coding in visual working memory accounts for stimulus-specific variations in recall. *J. Neurosci.* 38(32):7132–42
Vogel EK, McCollough AW, Machizawa MG. 2005. Neural measures reveal individual differences in controlling access to working memory. *Nature* 438(7067):500–3
Wang X-J. 2001. Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci.* 24(8):455–63
White OL, Lee DD, Sompolinsky H. 2004. Short-term memory in orthogonal neural networks. *Phys. Rev. Lett.* 92(14):148102
Whitney D, Levi DM. 2011. Visual crowding: a fundamental limit on conscious perception and object recognition. *Trends Cogn. Sci.* 15(4):160–68
Wilson NR, Runyan CA, Wang FL, Sur M. 2012. Division and subtraction by distinct cortical inhibitory networks in vivo. *Nature* 488(7411):343–48
Wimmer K, Nykamp DQ, Constantinidis C, Compte A. 2014. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat. Neurosci.* 17(3):431–39
Woloszyn L, Sheinberg DL. 2012. Effects of long-term visual experience on responses of distinct classes of single units in inferior temporal cortex. *Neuron* 74(1):193–205
Yu Q, Panichello MF, Cai Y, Postle BR, Buschman TJ. 2020. Delay-period activity in frontal, parietal, and occipital cortex tracks noise and biases in visual working memory. *PLOS Biol.* 18(9):e3000854
Zanto TP, Rubens MT, Thangavel A, Gazzaley A. 2011. Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. *Nat. Neurosci.* 14(5):656–61
Zhang X, Yan W, Wang W, Fan H, Hou R, et al. 2019. Active information maintenance in working memory by a sensory cortex. *eLife* 8:e43191
Zipursky SL, Sanes JR. 2010. Chemoaffinity revisited: Dscams, protocadherins, and neural circuit assembly. *Cell* 143(3):343–53