Cognitive Neuroscience of Visual Working Memory

Bradley R. Postle

Departments of Psychology and Psychiatry, University of Wisconsin–Madison

To appear in: Logie, R.H., Camos, V., & Cowan, N. (in press). *Working Memory: State of the Science*. Oxford, UK: Oxford University Press.

postle@wisc.edu

Dept. of Psychology

University of Wisconsin–Madison

1202 West Johnson St.

Madison, WI. 53706

U.S.A.

Preamble.

Memory refers to the influence of past experience on current thought and behavior. Although all sentient humans have an intuition about what memory is, and what it feels like to remember something, understanding how memory relates to other aspects of cognition requires careful definition of constructs, and explicit articulation of assumptions. For example, although memory is often thought of as a (or many) cognitive system(s), the word *memory* can also be used to refer to a property of a system whose primary function is not mnemonic. For example, the gill withdrawal reflex of the *Aplysia* demonstrates habituation when the tenth instance of touching it with a probe produces a slower and smaller withdrawal motion than had the otherwise-identical first instance of touching it. The molecular events and physiological processes that underlie this experience-dependent change in the functioning of the sensorimotor circuitry that innervates the gill and its musculature are understood in exquisite detail. Importantly, however, these are best construed as elements of the gill-withdrawal system that endow it with mnemonic properties, not as the mechanisms of a memory system per se. This chapter will be guided by the perspective that behavior on visual working memory tasks arises from intrinsic properties of the visual system, of the oculomotor and skeletomotor systems, and of frontoparietal control systems, including those that function as sources of attentional control. That is, working memory may not involve any discrete *systems*, whether considered from a cognitive or a neural perspective, whose primary function is working memory. Rather, it may be a functionality resulting from the control of sensorimotor and representational systems.

1 Definition of Working Memory

Working memory is the ability to hold information in an accessible state – in the absence of relevant sensory input – to transform it when necessary, and to use it to guide behavior in a flexible, context-dependent manner.

2 Describe and explain the methods you use

I am interested in the neural bases of human cognition – how does the brain give rise to the mind? One important development in cognitive neuroscience over the past two decades has been the development of methods for applying multivariate information-based analyses to neuroimaging data. This has allowed for assessment of the neural representation of information in ways that simply weren't possible previously, and has resulted in reevaluation of several earlier findings.

2.1 Multivariate analyses of neural data

Fundamentally, what multivariate methods make possible is the assessment of distributed representations, an important advance beyond the assumption implicit in most univariate analyses, which is that fluctuations in regional aggregations of signal intensity correspond to varying levels of engagement of a single mechanism or process. For example, in the first decade-and-a-half following the earliest studies of human working memory with

neuroimaging (Jonides et al. 1993, Cohen et al. 1994), memory load-related changes in delay-period signal intensity pooled across tens -- if not hundreds or more – of voxels were assumed to index varying demands on information storage (e.g., Braver et al. 1997, Todd and Marois 2004, Postle 2006). More recently, however, studies using multivariate pattern analysis (MVPA) have demonstrated that stimulus-specific information cannot always be decoded from regions whose activity shows load sensitivity and, conversely, stimulus-specific information can be decoded from regions that show neither load sensitivity nor, indeed, elevated levels of activity during the delay (Harrison and Tong 2009, Serences et al. 2009, Riggall and Postle 2012, Emrich et al. 2013, Gosseries et al. 2018). Such findings have provided important evidence for sensorimotor-recruitment models of visual working memory (e.g., D'Esposito and Postle 2015, Postle 2015, Serences 2016).

2.1.1 Multivariate inverted encoding modeling (IEM)

Although most readers of this volume will be familiar with MVPA (e.g., Norman et al. 2006, Pereira et al. 2009), a summary of a second type of multivariate analysis method, inverted encoding modeling (IEM), will be useful, both because much of the work to be described here relies on this method, and because there exist misconceptions about how results generated with IEM are typically interpreted. IEM is a forward modeling approach that implements a dimensionality reduction on neural data to track population-level representation of stimulus characteristics. In the case of line orientation, for example, one can hypothesize that any given angular value can be represented in a unique pattern of weightings across several hypothetical broadly tuned informational channels. These channels are represented in the analysis with a basis set of overlapping broadly tuned functions, typically half-wave rectified sinusoids, each centered at a different angular value such that they span the full possible range of 180° of rotation. The model is then trained by regressing against it neural data (in our case, fMRI or EEG) acquired while a subject was performing visual working memory for line orientation. For fMRI, for example, the basis set's representation of the angular value of each stimulus presented to the subject is entered as a regressor into a general linear model that estimates the orientation tuning function of each voxel (sometimes referred to as the "population receptive field (pRF)," Dumoulin and Wandell 2008) in a region of interest (ROI). Testing of the model is carried out by inverting the matrix that maps from channel space to voxel space, then determining whether data from trials that the model has never seen generate reconstructions, in channel space, of the stimulus values used for these test trials. Successful reconstructions of test stimulus values are interpreted as reconstructions of population-level representations of these stimuli. Importantly, because the shape of the basis functions was determined a priori and used in the training of the model, quantitative values from stimulus reconstruction can be interpreted as indices of, for example, the magnitude and the precision of a neural representation (indexed by the amplitude and width of the reconstruction, respectively, Brouwer and Heeger 2009, Serences and Saproo 2012).

Recently, some critiques of the IEM approach have appeared in the literature, and although these contain mistaken assumptions and, indeed, outright misconceptions about the assumptions underlying IEM and how IEM reconstructions are interpreted, addressing some of these will be helpful for clarifying some of the useful features of this approach. One point raised

by Gardner and Liu across two papers (Liu et al. 2018, Gardner and Liu 2019) is that the results from IEM cannot be used to draw inferences about the tuning properties of individual neurons. Although this assertion is true, it has no bearing on the results that will be described here, because our experiments with fMRI and EEG are simply not intended to address this level of neural functioning, a point made with more elaboration elsewhere (Sprague et al. 2018, Sprague et al. in press). A second concern raised by Gardner and Liu (2019) gets to a fundamental issue of studying the neural representation of information. They frame this concern with the observation that IEM differs from more direct reporting of neurophysiological measures in that "the ordinate of the graph [produced by testing an IEM] is no longer a direct measure of neural activity" (p. 3). Examples of direct measures that the authors raise include firing rate, membrane potentials, reflectance changes from intrinsic signals, and fluorescence changes from voltage sensitive dyes. "Even for BOLD activity averaged across a visual area," they note, "parametric sensitivity to the strength of a visual stimulus can be assessed by plotting response magnitude as a function of stimulus properties like contrast … or motion coherence …" (p. 3). Although these statements are true, what they don't acknowledge is that to limit oneself to analyses of data formatted such that they can be expressed as a direct measure of neural activity is to preclude the ability to study much of cognition, because the neural coding of most kinds of information that one would want to study is high dimensional. That is, firing rate, membrane potential, and other first-order summaries of neural activity are inherently univariate measures, and analyses that are limited to such measures are blind to information that is represented in high-dimensional patterns of activity distributed across multiple processing units. To be concrete, if one were to average BOLD signals across a 500-voxel region of interest (ROI) that included the foveal representation of V1, it is true that one would expect to observe a monotonic increase in BOLD signal intensity in conjunction with parametric increases in contrast of a sinusoidal grating of a particular orientation – let's say 0°. The limitation of this approach, however, is clear as soon as one considers that the same pattern of monotonically increasing BOLD signal intensity would also be observed from this 500-voxel ROI for a sinusoidal grating of 30°, for one of 60°, and so on. That is, univariate measures are rarely informative about the neural representation of stimulus-specific information. Indeed, we have already considered the fact that cognitive neuroscience research on working memory has established that univariate measures can lack sensitivity (e.g., they fail to detect delay-period stimulus representation in early visual cortex) and they can lack specificity (e.g., load-sensitive activity need not contain stimulus-related information). IEM, like any other dimension-reducing approach to data analysis (e.g., principal component analysis [PCA], independent component analysis [ICA]), necessarily cannot retain the units in which levels of activity measured at individual sensors are acquired.

The goal of measuring stimulus information also motivates the choice of IEM over multivariate *decoding* approaches, such as MVPA. Although MVPA can provide important information about where in the brain stimulus- or category-level information is represented, its results are typically evaluated in terms of decoder performance, not with reference to the stimulus representation, per se. Thus, for example, when varying memory load from 1 to 2 to 3 items is seen to result in a decrease in delay-period stimulus decodability from roughly 80% to 70% to 65%, as is the case for subject #6 from Figure 5 of Emrich, Riggall, et al. (2013), one can't

know what aspect of the neural representation of the stimulus is changing to produce this pattern. (Is it a decline in precision? in the strength of the representation? in both? or in some other factor?) IEM, in v contrast, provides quantification of parameters that can be interpreted in terms of the properties of a stimulus representation. An instructive example comes from Sprague and Serences (2013), who explored the effects of attention on the representation of locations in retinotopic space by flashing a flickering checkerboard stimulus at each of 36 distinct locations while subjects either attended to the fixation point or to the flickering checkerboard (while maintaining central fixation). Two "sanity checks" indicated that IEM reproduced known facts about the visual representation of space: the size of neural representations of space increased as location of the checkerboard moved further from central fixation; and the size of the representation of any given retinotopic location was larger (i.e., coarser) in higher-level areas such as V4, MT+, IPS, and superior frontal sulcus (SFS), than in early visual areas. Importantly, with regard to the effects of attention, IEM revealed that univariate and population-level measures diverged. When attention was allocated to a region away from central fixation, the size of the pRFs of voxels in higher-level brain areas that represented the attended region increased. This trend, alone, would be difficult to reconcile with the fact that allocating attention to a location in the periphery is known to improve the precision of visual perception at that location, for the simple reason that larger receptive fields are less spatially precise. In contrast to the univariate effects at the level of individual voxels, however, IEM indicated that allocating spatial attention to a region in space resulted in an increase in the amplitude of the population-level representation of that location, but not in a change in the precision of these representations. Thus, at a population level – the level that we assume to be most important for guiding behavior and for determining subjective experience -- the effect of covert attention to a region in space is to strengthen the neural representation of that location.

2.2. Neural network modeling

One point highlighted by the recent debate about the assumptions underlying the IEM approach (Section 2.1, Liu et al. 2018, Sprague et al. 2018, Gardner and Liu 2019, Sprague et al. in press) is that IEM reconstructions are models of neural representations, but not direct measurements of representations themselves. As can be the case in many other domains of science, unexpected or atypical behavior of a model can lead one to address the same question with a different method that does not make the same assumptions as that model. Of particular relevance here, this chapter will consider studies in which the manipulation of priority between two items concurrently held in working memory has produced a systematic, but heretofore rarely reported, change in the IEM reconstruction of the unprioritized item. To better understand the factors that underlie this effect, we have turned to neural network models.

Masse and colleagues (2019) have recently noted that "recurrent neural network (RNN) models have opened a new avenue to study the putative neural mechanisms underlying various cognitive functions. Crucially, RNNs have successfully reproduced the patterns of neural activity and behavioral output that are observed in vivo, generating insights into circuit function that would otherwise be unattainable through direct experimental measurement" (p. 1159). In our case, it is not circuit-level function, but the dynamics of population-level stimulus representation for which we seek insights. Important details of implementation will be

described in section x, but the general logic is to train a RNN to perform a working memory task, then examine how the hidden layer of the RNN represents stimulus information during the delay period of the task. Although such an exercise cannot, alone, provide "proof" about how the human brain accomplishes working memory performance, it can reveal candidate processes whose biological reality can then be tested with human data.

3.Unitary vs. non-unitary nature of working memory

This strikes me as an ill-posed question, because it presupposes that working memory *is* a system, and the question to be sorted is what kind of system. To elaborate on the perspective laid out in the *Preamble*, consider this example of a parent spectating at a gymnastics competition. His challenge is to track the activity of his daughter and her teammates among the churning melee of adolescent girls, all sporting similar-looking ponytail hair styles and spangly leotards. This is not a behavior that makes overt demands on memory, although it certainly does require guidance from a priority map, believed to be instantiated in recurrent activity between the intraparietal sulcus (IPS) and the frontal eye fields (FEF; located in the superior frontal cortex (SFC) in humans), as well as the superior colliculus (e.g., Bisley and Mirpour 2019). In this scenario, if the spectator's priority map can temporarily retain information about the locations of the athletes of interest while he briefly averts his eyes to read a text message on his mobile phone, there is no need to assert that an additional system, nor that a qualitatively different neural computation, needs to be engaged for him to successfully return his gaze to the targets of interest gathered around the uneven parallel bars, halfway across the crowded gymnasium. Luck and Vogel (2013) have made a similar argument, noting that "visual working memory may not be a memory system per se, but may instead be a general-purpose visual representation system that can, when necessary, maintain information over short delays" (p. 394). (At the risk of sounding churlish, though, I would insist that there does not exist a "general-purpose visual representation system" that's different from the visual system and the oculomotor system that are also engaged in tasks not considered to require working memory.)

Neural data are also consistent with this perspective. Neurons in the FEF of nonhuman primates encode information about recent saccade targets during free viewing behavior (Mirpour et al. 2019), and MVPA of fMRI activity from SFC and from IPS in humans indicates that the neural encoding of egocentric location is highly similar whether subjects are engaged in planning a delayed saccade to a visible stimulus ("intention"), covertly attending to this stimulus in order to detect a change in its luminance ("attention"), or preparing a delayed response to the same location when it must be remembered across a delay ("retention," Jerde et al. 2012). Damage to (Mackey et al. 2016) and rTMS of (Hamidi et al. 2008) PFC in humans only disrupts spatial working memory performance when the FEF are affected.

4. The role of attention and control

4.1 Spatial attention and spatial working memory

Top-down control of mental activity arises from the dynamic interplay between dorsal (endogenous) and ventral (exogenous) attentional circuits. The strong overlap between neural systems that support oculomotor control, spatial attention, and visuospatial working memory has been thoroughly documented in several previous studies and reviews (one recent authoritative review being Jonikaitis and Moore 2019), and provides a compelling basis for the longstanding idea that working memory performance may reflect "'nothing more' than the preparation to perform an action, whether it be oculomotor, manual, verbal, or otherwise" (Theeuwes et al. 2005) (pp. 198–199). This perspective receives further support from the fMRI study of Jerde et al. (2012), as reviewed above, as well as from demonstrations that eye movements executed in the dark selectively disrupt visual working memory for locations (e.g., Postle et al. 2006).

4.2 Feature- and object-based attention and visual working memory for features and objects

Real-time visual object recognition requires interactive signaling between neural circuits at multiple levels of the visual hierarchy, from primary visual cortex to high-level distributed representations that underlie categorization and semantic memory. Importantly, feedback from higher levels to primary cortex, and even to sensory thalamus, is critical for visual perception and its attentional control (e.g., Sillito et al. 1994`, Cudeiro and Sillito 2006, Sillito et al. 2006), and this has also been shown to be important for visual working memory (van Kerkoerle et al. 2017). Because one can see the effects of object-based attention at all levels of processing involved in object perception and categorization (e.g., Çukur et al. 2013, Ester et al. 2016), the sensorimotor-recruitment framework predicts that working memory for objects and features of objects would also entail the top-down modulation of these circuits. (The constructs of feature-based attention and object-based attention are closely linked and, indeed, it's unclear if the two differ other than in the grain of detail at which elements in the visual scene must be analyzed in any given task or situation (e.g., Scolari et al. 2014). Therefore, for simplicity, this chapter will use "object-based attention" to refer generally to the two constructs, and similarly it will use "object working memory" to refer in general to working memory for visually presented objects and to working memory for the features of visually presented objects.)

Although object-based attention is ostensibly "nonspatial," the fact that visual perception is inherently grounded in spatial reference frames provides a rationale for why there may be important links between the workings of frontoparietal gaze control circuitry and feature-based attention (e.g., Moore and Zirnsak 2017, Bisley and Mirpour 2019). A compelling empirical example comes from the fact that subthreshold microstimulation of the FEF (i.e., at an amplitude that does not generate a saccade) produces an attention-like enhancement of the visually driven response of V4 neurons with receptive fields overlapping with the stimulated FEF motor field, enhancements that are greater for stimuli for which the V4 neuron is optimally tuned and/or when a distractor is present elsewhere in the visual field (Moore and Armstrong 2003). Saccade planning has also been shown to influence object working memory. For example, preparing a saccade to a stimulus location improves the subsequent recognition of the shape that had been presented at that location, even if the planned saccade is never

performed (Hanning et al. 2016). Furthermore, within this same experimental context, an intervening saccade can negate the attentional benefits that are otherwise produced by a retrocue: When sample offset is followed by a retrocue indicating which of two sample stimuli will be tested, this retrocue does not benefit recognition performance on trials for which subjects know that they will first need to make a saccade to the location that had been occupied by the uncued sample (Hanning et al. 2016).

The linkage between oculomotor control and object working memory may help explain the results from a recent fMRI study of delayed recall of orientations, which showed evidence that individual differences in the precision of behavioral performance were predicted, in part, by the representation of stimulus location, even though this contextual information was not needed to perform the task (Cai et al. 2019). One-item trials on this task began with the presentation of a sample at one of four locations, followed by a delay, followed by a recall dial that appeared at the same location as had the sample. Across subjects, the strength of the representation of sample location (indexed by MVPA) at encoding and at recall was positively related to the behavioral precision of recall. Furthermore, IEM estimates of the neural representation of orientation in occipital cortex were higher in amplitude, and more closely related to recall precision, when the IEM models that generated them included information about location context.

4.2.1 The role of microsaccades in visual attention and visual working memory

An additional factor that has recently been gaining prominence in the literature is the linkage between object-based attention, object working memory, and microsaccades. Microsaccades are small saccadic deviations from the point of fixation, typically smaller in amplitude than 1° of visual angle, that occur during fixation of a stable target. Although there are many proposed functions for microsaccades relating to possible roles in optimizing foveal vision, they have received relatively little consideration in attention research. Recently, however, it has been reported that the attentional effects of cuing are strongly tied to microsaccades, with the neurophysiological enhancement of the representation of the cued stimulus only observed when preceded by a microsaccade in the direction of the cued stimulus. Furthermore, the precise timing of the onset of this attentional enhancement was more closely tied to the execution of this microsaccade than to the onset of the attentional cue (Lowet et al. 2018). Interestingly, the periodicity of the attention-related microsaccadic activity described by Lowet and colleges (2018) is in the same 3-4 Hz range that is associated with the periodic attentional sampling of the visual scene that is characteristic of the behavior and the neurophysiology of NHP and human visual cognition (e.g., Fiebelkorn and Kastner 2019). The importance of understanding the relation of microsaccades to object working memory has recently been highlighted by the demonstration that subtle but systematic differences in location of gaze can be used to decode the delay-period representation of line orientation (Mostert et al., 2018). In Section 7 we will return to the question of whether such effects in eye-tracking data -- and, indeed, whether microsaccade-related signals in neural data -- are best treated as a potential empirical confounds in studies of working memory, or, alternatively, as factors reflecting an inherently functional role that microsaccades may play in encoding the contents of nonspatial visual working memory (Dotson, Hoffman, Goodell, & Gray, 2018).

4.3. The top-down control of attention and working memory

Several explicit computational models can produce controlled attention without resorting to a homunculus. One model that accounts in considerable detail for the empirical findings that motivated the biased competition model of visual attention (Chelazzi et al. 1993, Chelazzi et al. 1998), and that's also relevant to considerations of the overlap between gaze control and attention (e.g., Moore and Zirnsak 2017, Bisley and Mirpour 2019), is Hamker's (2005) model of biased competition in visual search, which relies on recurrent activity between feature-selective neurons in IT cortex and the FEF. In this model, top-down control emerges from the complex interaction between the representation of the search template in IT, feedforward signaling by IT, and feedback signaling from the subpopulation of "movement neurons" in FEF. (Note that although our understanding of some specific details about FEF circuitry, cell types, and connectivity with other brain areas have evolved during the ensuing years (e.g., Merrikhi et al. 2017), the principles underlying the Hamker (2005) model continue to be relevant.) A second model that's highly relevant to concerns about control without a homunculus was developed by Rougier and colleagues (2005), in part to address the fact that "a major challenge for theories of the neural bases of cognitive control …[is] how it can be explained in terms of self-organizing mechanisms that develop on their own, over time, without recourse to unexplained sources of influence or intelligence" (p. 7338). The authors used a connectionist framework with many biologically inspired properties, including synaptic learning rules (O'Reilly and Munakata 2000), and an architecture that included a "posterior cortex" hidden layer, and a separate "PFC" context layer that could influence both the hidden layer and the "response" (i.e., output) layer. Units in the "PFC" had two unique and also biologically inspired properties: an "up" state of sustained, elevated activity that was robust against interfering signals, and a readily triggered bistability between this sustained "up" state and a "down" state that enabled rapid updating of patterns of activity in the PFC. The final critical element was a dopamine (DA)-mediated reward prediction error (RPE) signal that could trigger transitions in the bistable state of PFC units, a biologically inspired element (in that DA has this influence on PFC pyramidal neurons) that incorporated principles of reinforcement learning into the simulation. With this architecture, the authors used trial-and error learning to first train the model to perform naming and same-different comparisons of stimuli that could vary according to shape, size, egocentric, location, and color. This stage of training taught the "posterior cortex" to reliably represent the different stimulus domains and the relations between ordinal levels within each (e.g., small < medium small < medium large < large), and the "PFC" to represent rules (e.g., how to decide if a medium small-sized, yellow, vertically striped square located in the upper-right quadrant is the same as or different from a small, green, vertically striped square located in the upper-right if the matching dimension is size). Finally, they introduced a Wisconsin Card-Sorting Task (WCST), and the network learned to perform it at a high level of proficiency, staying on one sorting rule as long as that rule produced positive feedback, and switching rules upon receiving negative feedback (the incorrect choice generating a RPE signal, which transiently shunted the PFC from its up to its down state). Interestingly, if one considers the simulated role of the RPE signal in the Rougier et al. (2005) model, together with empirical evidence for the role of dopamine in controlling the influence of the FEF on visual attention and working memory (e.g., Noudoost and Moore 2011, Merrikhi et

al. 2017), one can see how the Hamker (2005) model might be modified to support the flexible and context-sensitive selection of rules to guide oculomotor behavior.

Whereas the Rougier et al. (2005) model illustrated the power of incorporating principles of reinforcement learning into neural models, more recently it has been suggested that PFC has a property of "meta reinforcement learning" in which dynamic adaptations of behavior that follow the principles of reinforcement learning can be implemented by on a trial-by-trial basis by patterns of activity in the PFC (Wang et al. 2018). The gist is that unique physiological properties and anatomical connectivity of the PFC allow for dopamine-based reinforcement learning to train this region, over time, to be able to operate as a "learning system" that implements principles of reinforcement learning in patterns of activity. Although "conventional" reinforcement learning is slow, based, as it is, on incremental changes in synaptic weights, "meta reinforcement learning" can change behavior on a moment-by-moment basis, because such constructs as reward, choice history, object value, and prediction error can be represented dynamically in distributed patterns of activity (rather than in patterns of weights that bias connection strengths between different neurons). Importantly, because reinforcement learning is unsupervised, this scheme endows the PFC with the ability to control behavior without needing the "supervision" of a homunculus.

4.3.1. The neural bases of the source of object-based attention (and of object working memory?)

Mapping theoretical models like those summarized in the previous subsection onto neural systems is an important goal for cognitive neuroscience moving forward. One region that is emerging as an important node in the control of object-based attention is in posterior ventrolateral PFC, a region known as the inferior frontal junction (IFJ, at the intersection of the inferior frontal and precentral sulci) in the human, and the ventral prearcuate area (VPA) in the NHP. In humans, Baldauf and Desimone (2014) observed with magnetoencephalography (MEG) that alternating attention between superimposed streams of translucent images of faces and of houses produced the expected alternations of attention-related boosts of signal intensity in stimulus-related activity in posterior face- and house-sensitive regions, and these were tightly linked to alternations in the strength of coherence in the upper gamma band (roughly 60-100 Hz) between IFJ and these posterior regions. In the NHP, Bichot and colleagues (2015) have demonstrated that, in a visual search task, neurons in VPA showed selectivity for the search target and showed feature-based attentional modulation earlier than did neurons in FEF. Furthermore, local inactivation of VPA neurons produced marked deficits in search performance, and abolished the feature-based attention modulation of FEF that was observed prior to the inactivation (Bichot et al. 2015). It remains to be determined how closely the feature-based attention-related functions of VPA may correspond the involvement of VPA in the control of visuoobject working memory, as has been described in separate research (Mendoza-Halliday et al. 2014).

5. Storage, maintenance and loss of information in WM

5.1. Evidence for *where* influences models of *how*

In cognitive neuroscience, questions of *how* information is retained for working-memory task performance are often closely tied to *where* in the brain one can find evidence for the delay-period representation of trial-relevant information. Historically, evidence for stimulus-selective delay-period activity has often been interpreted as evidence for a storage function in that region, and such evidence, in turn, has been influential in the development of models of visual working memory and visual cognition. For example, reports of stimulus-selective delay-period activity in the dorsolateral PFC, at the level of single-unit recordings (e.g. , Funahashi et al. 1989, Funahashi et al. 1990, Wilson et al. 1993) or MVPA decoding from population recordings (Mendoza-Halliday et al. 2014), has influenced theories on a wide range of questions, including the role of working memory in high-level cognition (e.g., Goldman-Rakic 1992, Davachi et al. 2004), the functional organization of cortical contributions to high-level cognition (Katsuki and Constantinides 2012, Leavitt et al. 2018), and principles of neural coding (Murray et al. 2017, Constantinidis et al. 2018). Another important example is that the observation of patterns of load-sensitive delay-period activity measured with fMRI in IPS (Todd and Marois 2004, Todd and Marois 2005, Xu and Chun 2006), and with EEG at occipitoparietal scalp electrodes (Vogel and Machizawa 2004, Vogel et al. 2005), has led to the codification of an EEG component, the contralateral delay activity (CDA), that has been an influential tool for developing models of individual differences in working memory abilities (a.k.a. capacity limitations, e.g., Luck and Vogel 2013, Luria et al. 2016), of visual search (Woodman 2013), of automaticity (Servant et al. 2018), and for inferring roles for working memory in a wide variety of cognitive tasks that do not make overt demands on the short-term retention of information (Balaban and Luria 2019).

5.1.1. Caveats about inferring function from activity

As discussed in *Section 2.1*, the application of multivariate analyses to neuroimaging data sets has led to important advances in our understanding of the neural bases of many cognitive functions. The impressive sensitivity of these methods has also highlighted the challenge of how to go about assessing, when multiple regions can be shown to represent stimulus-specific information, whether these regions are all supporting the same function, or perhaps different functions that nonetheless all entail the active representation of the same information.

Not long after the power of MVPA for working memory research was demonstrated by delay-period decoding of stimulus information from V1 (despite the absence of elevated activity, Harrison and Tong 2009, Serences et al. 2009), Christophel and colleagues (2012, 2014) and Bettencourt and Xu (2016) published evidence for delay-period stimulus representation in parietal cortex, and Ester and colleagues published IEM evidence that "Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory" for the orientation of square-wave gratings (Ester et al. 2015), see also (Cai et al. 2019). Similarly, the remembered direction of motion is decodable with MVPA from delay period signal across multiple subregions of the IPS (Gosseries et al. 2018). Although some have interpreted these and similar findings as evidence for working memory buffers operating in IPS and frontal cortex (e.g., Riley and Constantinidis 2016, Christophel et al. , Xu 2017, Leavitt et al. 2018), other interpretations are possible. In data from this author's group, for example, the measures of delay-period stimulus information that we have observed in IPS and PFC differ from those

simultaneously measured in VOT regions in several ways. IPS and PFC representations tend to be weaker and less robust, as indexed by lower decoding scores and the failure to decode or reconstruct at memory loads higher than 1 (e.g., Gosseries et al. 2018). Furthermore, in Gosseries, Yu, et al. (2018), we operationalized memory load with trials presenting one motion patch (*1M*) vs. 1 motion patch and 2 color patches (*1M2C*), and context-binding load with *1M* vs. *3M*. (On three-item trials (i.e., *1M2C* and *3M*) samples were presented serially, and a digit indicated the item to recall, thereby requiring the binding of each sample to its ordinal context.) In this study, MVPA-decoded delay-period representations in IPS coexisted with patterns of BOLD signal whose sensitivity to context-binding load covaried with *1M*-to-*3M* declines in behavioral precision, and with *1M*-to-*3M* declines in decoding from VOT; no such relations were observed in comparison to analogous patterns of variation across *1M*-to-*1M2C*. Therefore, it's possible that delay-period representations in IPS had a more important role in the control of context binding than in stimulus representation per se. In a study by Cai et al. (2019), IPS and PFC representations did not show evidence for an influence of location context, and variation in their strength did not relate to behavior, suggesting that these representations may have been more abstract than those carrying information specific to the current trial.

5.2. Dynamic representations supporting working memory function

One argument that is made for the need for working memory buffers in non-sensory regions the brain is that this scheme would avoid the potential problem of working memory storage interfering with ongoing perception. For example, it has been argued that "… the content of [visual working memory] is fairly resistant to distraction. This is at odds with an intuitive understanding of the sensory account, which would predict a large interference between VWM storage and sensory processing of the distractor as a result of shared neural resources" (Xu 2017) p. 799. Where intuition fails, however, information theory dynamical systems theory, and computational modeling offer promising ways forward.

At a general level, one solution to intuited problem of interference between memory storage and ongoing perception is to re-represent the to-be-remembered information in a format that does not interfere with perceptual codes. One scheme that could accomplish this would be to encoding to-be-remembered information into an "activity-silent" state (Stokes 2015), possibly supported by patterns of oscillatory synchrony and/or transient changes in synaptic weights (e.g., Mongillo et al. 2008, Erickson et al. 2010, Toda et al. 2012, Barak and Tsodyks 2014). Empirical evidence consistent with this idea has been generated with perturb-and-measure studies that reveal the otherwise-subthreshold representation of information in the multivariate readout of response evoked by the delay-period perturbation (Wolff et al. 2015, Rose et al. 2016, Wolff et al. 2017), as well as with computer simulations (Manohar et al. 2019, Masse et al. 2019). A second possibility (that is not incompatible with an activity-silent mechanism, c.f. Masse et al. 2019) would be to recode to-be-remembered information it into a format that is different, and perhaps more robust, than the perceptual code. In this scheme, the recoded information could later be decoded back into its original format (e.g., Koyluoglu et al. 2017) or used in its transformed state to influence thought and/or guide behavior (e.g., Myers et al. 2017, van Ede et al. 2019). Importantly, a recoding scheme does not require engaging circuits that were not involved in the initial encoding of the information in question. In

the next section I'll consider empirical and computational evidence for one candidate recoding mechanism: priority-based remapping.

5.2.1. Priority-based remapping as a candidate mechanism for storage in working memory

Evidence for priority-based remapping was first observed in a two-item dual serial retrocuing study of delayed recall (Yu and Postle unpublished) but here we will focus on a 2-back working memory task with which this phenomenon has been studied in the greatest detail to date. In the 2-back study of Wan et al. (in-principle accepted), subjects viewed the serial presentation of oriented grating stimuli, and indicated for each, with a button press, whether is matched or non-matched the item that had appeared two positions previously in the series. In this way, each item *n* transitions through multiple states of priority: first as a recognition probe for the item *n-2*; then as an "unprioritized memory item" (UMI) while the subject compares item *n+1* against *n-1*; then as a "prioritized memory item" (PMI) in anticipation of its comparison with *n+2*. The EEG data from Wan et al. (in-principle accepted) were analyzed with an IEM that was trained on data from the delay period of an independent 1-item delayed-recognition task. The principal empirical finding of interest, which replicated Yu and Postle (unpublished), was that the IEM estimate of the neural representation of the UMI took on a value that was the opposite of its true value, then returned to its true value when it transitioned to PMI. To make this concrete, let's take the example of a hypothetical grating stimulus of orientation 30° (from a stimulus set where 0° corresponded to a horizontal orientation and 90° to a vertical one). While this 30° stimulus grating was a UMI, its neural representation was reconstructed by the IEM as most similar to the neural representation of 120°, then, when a PMI, its neural representation was once again reconstructed by the IEM as 30°.

There are two important points to emphasize about this finding. The first is to acknowledge that this pattern of IEM reconstruction of the UMI suggests an active memory trace, a finding at variance with previous studies that have failed to find evidence for an active trace of the UMI (Lewis-Peacock et al. 2012, LaRocque et al. 2013, Rose et al. 2016, LaRocque et al. 2017). The second is that this pattern would be more consistent with a remapping of how stimuli are represented within neural code than an actual recoding operation. This because an IEM trained to learn one code should fail when tested on data corresponding to a different code. A neural operation that preserved the learned neural states but changed the neural-pattern-to-stimulus-value mapping (e.g., flipping the mappings, or rotating them), however, could produce the pattern similar to what we have observed with the IEM reconstruction of the UMI (Wan et al. in-principle accepted, Yu and Postle unpublished).

5.2.1.1 Priority-based remapping: evidence from neural network modeling

To explore the mechanism that may underlie the transformation of the neural representation of the UMI, Wan and colleagues (unpublished) trained a fully recurrent neural network (RNN) with a hidden layer of 16 so-called long short-term memory (LSTM) units to perform the 2-back task (Figure 1, A. and B.). (For the present purposes, LSTM can be understood as a kind of architecture that allows artificial neural networks to process information during time steps when no new information is being fed to the network.) Once the network was trained, the representational dynamics of the network could be observed by

tracking the shifting patterns of activity over the course of an item's evolution from probe to UMI to PMI. More specifically, the dimensionality of the 16 units hidden layer was reduced using principal component analysis (PCA) and the trajectory of the first two principal components tracked. As illustrated in Figure 1.C., the network's representation of each stimulus item underwent a dynamic trajectory while held in working memory: when initially presented to the RNN, the RNN's representation of item *n* was aligned along an axis that separated "match" from "nonmatch" responses; during the ensuing delay period, when its status in the task transitioned to UMI, its representation by the RNN rotated until, during the presentation of *n+1*, it was orthogonal to the decision axis; then, when its status in the task transitioned to PMI, it continued its rotation along the same high-dimensional manifold (i.e., along the same "plane"), re-aligning with the decision axis during the presentation of *n+2*. As calculated across multiple simulations, the average distance between the axis of alignment of the UMI versus the PMI was 134°.

5.2.1.2 Priority-based remapping: evidence from human EEG

The previous subsection reviewed evidence, from a neural network simulation, rotational remapping may be a process whereby information can be held in working memory in a format that doesn't interfere with behavior. Is there, however, any evidence that a similar mechanism might be deployed by the human brain? To address this, Wan and colleagues (unpublished) applied the analysis that had been applied to the RNN to the EEG data in which IEM reconstructions of stimulus orientation difference as a function of the item's priority status (Wan et al. in-principle accepted). Conceptually, they treated the data from the 60 channels of the EEG in the same way as they had the activation values of the hidden layer of the RNN. (Some of the processing steps were necessarily different; before performing the PCA, temporal covariance matrices were first computed from the EEG data, then these covariance matricies eigendecomposed, following Cohen (2014).) The results produced a qualitative replication of the results with the RNN, with the UMI rotated relative to the PMI by 177°.

The putative priority-based mechanism described here may also account for the findings of van Loon and colleagues (2018), who acquired fMRI while subjects performed a visual search task in which each trial began with the sequential presentation of two search targets, followed by the sequential presentation of two search arrays, one corresponding to each of the two targets. Because the order of appearance of the search arrays would not be known until after the offset of the second target, each trial was assumed to entail the sequential prioritization of the working-memory representation of the two targets. Analyses with MVPA revealed that although a classifier trained on epochs when a stimulus category corresponded to the impending search array (i.e., when it was the prioritized item) could also decode evidence for that same category from epochs when it did not correspond to the impending search array (i.e., when it was the unprioritized item), the pattern of activity when this category was the unprioritized differed from when it was prioritized. This manifested in two ways: first, decoding of the unprioritized category with a decoder trained on information in the prioritized state was significantly below chance; and second, the high-dimensional representation of each category projected into opposite regions of multidimensional scaling (MDS) space during epochs when it was prioritized versus when it was unprioritized.

5.2.1.3 Priority-based remapping in working memory: a specific case of a more general mechanism?

Together, the results from IEM of neural data from two experiments using two different tasks (Wan et al. in-principle accepted, Yu and Postle unpublished), and the PCA-based analyses of 2-back data from an RNN simulation and from human EEG (Wan et al. unpublished)

suggest that the phenomenon of priority-based remapping may be implemented as a rotational remapping that is emergent from the dynamics of high-dimensional distributed representations. In keeping with the idea that working memory performance does not depend on specialized systems, a qualitatively similar phenomenon has been observed in populations of neurons in primary auditory cortex in mice exposed to sequences of tones. Although the animals were not trained on any specific task, stimuli that deviated from predictable sequences produced a transformation of the representation of previous stimuli from their perceptual codes into an orthogonal dimension. The authors propose that "This rotational dynamic may be a general principle, by which the cortex protects memories of prior events from interference by incoming stimuli (Libby and Buschman unpublished). Future research will need to assess the mechanism whereby rotational remapping can be triggered by top-down signals, which must be the case when it is observed in retrocuing tasks (Yu and Postle 2018, Sahan et al. unpublished).

6. The role of LTM knowledge in WM storage and processing

Every domain of cognition, including object recognition, depends on long-term memory (LTM). Without access to pre-existing representations of meaning, an individual viewing a scene would experience associative agnosia (Farah 1990). Indeed, it may not be possible for humans to perceive, nor to hold in working memory, a novel, abstract object or shape without associating it with prior knowledge and applying a verbal label to it (Postle et al. 2005, Postle and Hamidi 2007). It follows from this that visual working memory necessarily entails the representation of information from semantic LTM, in addition to sensory and motoric representations. One recent paper that makes a thorough and compelling set of arguments consistent with this line of thinking is from Cowan (2019).

One recent and exciting demonstration of an influence of LTM on working memory has come from the application of ideas from dynamical systems theory. Panichello and colleagues (2019) tested 90 human subjects and two NHPs on short-term recall (a.k.a. delayed estimation) of color from arrays of one versus three (for humans) or two (for NHPs) colored squares. For both species, even though stimuli were drawn evenly from the full 360° of possible colors, responses were markedly biased away from some colors and toward others. Attractor dynamics accounted for the frequency, bias, and precision of behavior better than standard mixture modeling. In particular, the greater imprecision in responses on high-load trials was shown to reflect both a drift of remembered stimulus representations toward stable attractor states and a greater influence of random diffusion (i.e., noise). The authors framed this as evidence for an error-correcting mechanism, whereby increased internal noise (manipulated here by varying load) is counteracted by drift toward stable long-term representations of color space. (From a Bayesian perspective, one could construe this as drawing on prior knowledge to counteract

uncertainty about the recently presented stimuli.) Evidence that the inferred attractor landscape reflected knowledge acquired prior to the presentation of sample stimuli on any given trial of the task came from the fact that the attractor landscape could be made to shift systematically to reflect new environmental statistics, as implemented by changing from a flat distribution of sample stimuli to a strongly biased distribution (Panichello et al. 2019).

To investigate the neural bases of these attractor dynamics, we have reanalyzed data from an fMRI study in which subjects had performed delayed recall of one versus three line orientations (Cai et al. unpublished). Analysis of the behavioral data from Cai et al. (unpublished) with the discrete attractor model developed by Panichello et al. (2019) provided a much better account of the data than did the classical three-factor mixture model (Bays et al. 2009): the discrete attractor model was estimated to be $2.9*10^6$ times more likely than the mixture model by cross-validated log-likelihood estimation. Furthermore, the drift and diffusion parameters from the discrete attractor model also related closely to load-related changes in IPS. We used within-subject correlation to relate individual differences in load-related changes in behavior to individual differences load-related changes in fMRI signal, and whereas using only the concentration parameter from the classical mixture model produced an adjusted r^2 of .25, adding the parameters from the discrete attractor model increased the adjusted r^2 to .69.

The question of whether or not working memory necessarily draws on LTM is often approached with logical argumentation, with reasonable people disagreeing over the strength of various arguments (e.g., Norris 2017, Cowan 2019). The findings reviewed here, in contrast, provide a quantitative demonstration that a model that incorporates an influence of LTM on working memory performance does a far superior job of accounting for individual differences, both in behavior and in task-related neural activity, than does a model that does not.

7. Is there evidence that is not consistent with your theoretical framework, and how does your framework address that inconsistency?

Within the domain of visual working memory, the evidence that would be most prominently inconsistent with my theoretical framework would be evidence for working-memory storage-related functions in circuits that are not associated with visual perception and/or the representation of visual knowledge. Prominent claims of such evidence come from findings of stimulus-specific delay-period activity in the IPS (e.g., Bettencourt and Xu 2016, Christophel et al. 2017, Xu 2017, Christophel et al. 2018, Xu 2018) and the PFC (Mendoza-Halliday et al. 2014, Riley and Constantinidis 2016, Mendoza-Halliday and Martinez-Trujillo 2017, Riley et al. 2017, Constantinidis et al. 2018, Leavitt et al. 2018). Importantly, I do not question the veracity of the data contained in these reports, but, rather, the interpretation often given to these data. Stated most broadly, I believe that many of the instances of delay-period stimulus-specific activity in IPS and PFC may reflect the operation of control processes rather than the operation of storage buffers per se. Alternative explanations for the functions supported by these findings appeal to the same roles in controlling behavior that these neural systems play in situations that don't make overt demands on working memory. These include protection from the influence of external interference (e.g., Malmo 1942, Chao and Knight 1995); control of perseveration and/or proactive interference (e.g., Tsujimoto and Postle 2012);

and the need to manipulate remembered information prior to using it to guide behavior (e.g., D'Esposito et al. 1999, Masse et al. 2019).

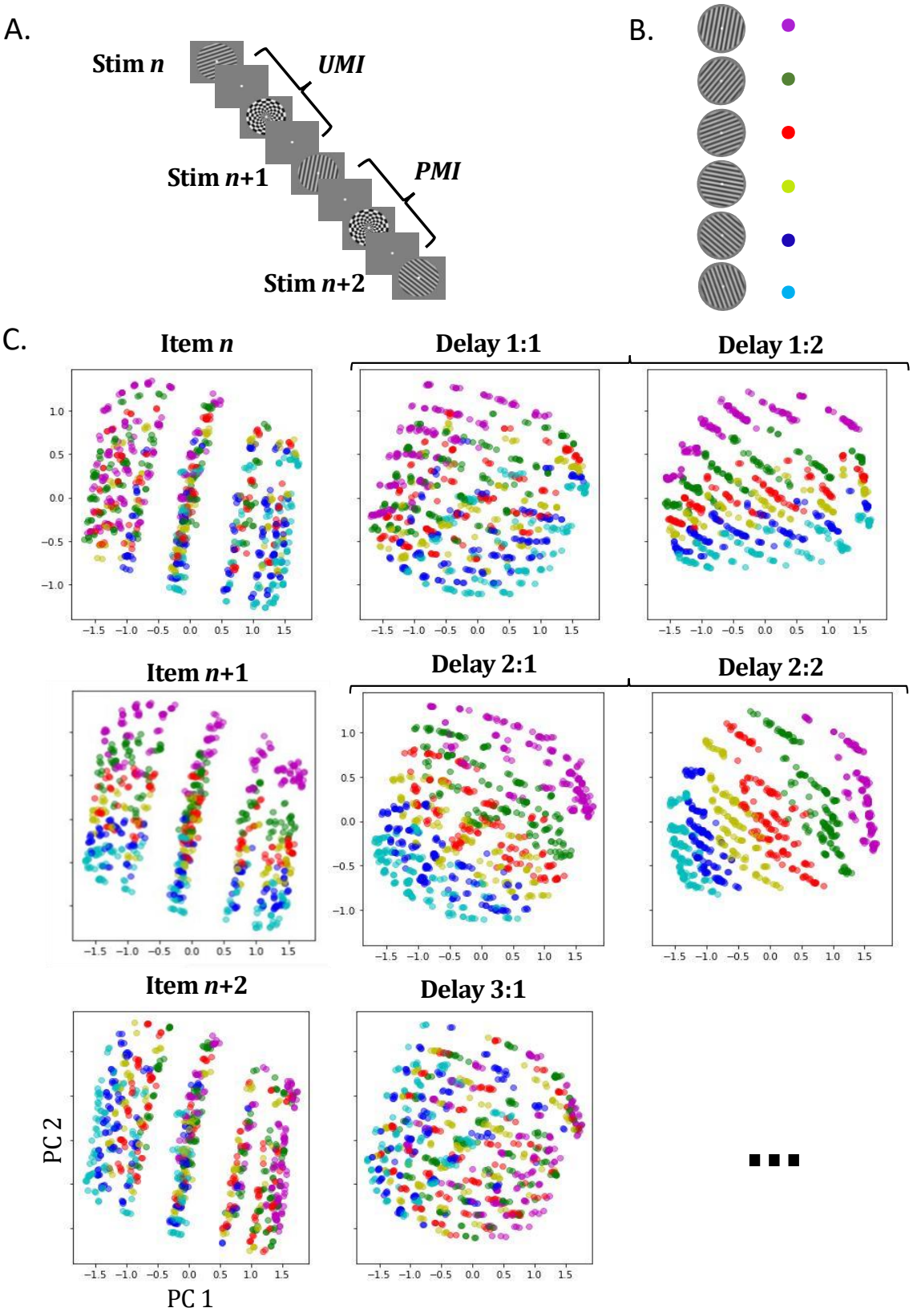
A.
Stim *n*
UMI
Stim *n*+1
PMI
Stim *n*+2
B.
C.
Item *n*
Delay 1:1
Delay 1:2
Item *n*+1
Delay 2:1
Delay 2:2
Item *n*+2
Delay 3:1
PC 2
PC 1

Figure 1. Rotational remapping as a mechanism to represent priority in working memory.

A. Cartoon of a sequence of stimuli from the 2-back task as administered by Wan et al. (in-principle accepted).
B. Correspondence between stimulus labels (left-hand column) and color codes used for RNN simulation by Wan et al. (unpublished).
C. The first two principal components extracted from activation patterns of the hidden layer of the RNN, at multiple successive time steps of simulated 2-back performance. The plot labeled "Item *n*" corresponds to the presentation of "Stim *n*" in panel A.; the plots labeled "Delay 1:1" and "Delay 1:2" correspond to the two time steps while item *n* is a UMI; and so on. Each colored dot represents a simulated trial in which that item progressed through the states of recognition probe ("Item *n*"), UMI ("Delay 1:1", "Delay 1:2", and "Item *n*+1"); and PMI ("Delay 1:1", "Delay 1:2", and "Item *n*+2"). In the "Item *n*" plot, trials for which *n* required a match response occupy the manifold (i.e., the "stripe") that aligns with the value of 0.0 of the first principal component, and trials for which *n* required a nonmatch response occupy the manifolds that appear to the left and to the right of the "match" manifold in this plot. Note that, across the time that n is processed in working memory, its representation rotates clockwise in the image plane, such that when item *n*+1 is presented, *n* is aligned with an axis that is orthogonal to the decision axis, and that when *n*+2 is presented, *n* is again aligned with the decision axis. The rotated state of the representation of *n* at the time when the network is assessing the match between *n*+1 and *n*-1 may reduce the likelihood that the identity of *n* will interfere with the decision about *n*+1.

References

Balaban, H. and R. Luria (2019). "Using the contralateral delay Activity to study online processing of items still within view." Neuromethods.

Baldauf, D. and R. Desimone (2014). "Neural mechanisms of object-based attention." Science **344**: 424-427.

Barak, O. and M. Tsodyks (2014). "Working models of working memory." Current Opinion in Neurobiology **25**: 20-24.

Bays, P. M., R. F. Catalao and M. Husain (2009). "The precision of visual working memory is set by allocation of a shared resource." Journal of Vision **9**.

Bettencourt, K. C. and Y. Xu (2016). "Decoding the content of visual short-term memory under distraction in occipital and parietal areas." Nature Neuroscience **19**: 150-157.

Bichot, N. P., M. T. Heard, E. M. DeGennaro and R. Desimone (2015). "A source for feature-based attention in prefrontal cortex." Neuron **88**: 832-844.

Bisley, J. W. and K. Mirpour (2019). "The neural instantiation of a priority map." Current Opinion in Psychology **29**: 108-112.

Braver, T., J. D. Cohen, L. E. Nystrom, J. Jonides, E. E. Smith and D. C. Noll (1997). "A parametric study of prefrontal cortex involvement in human working memory." NeuroImage **5**: 49-62.

Brouwer, G. J. and D. J. Heeger (2009). "Decoding and reconstructing color from responses in human visual cortex." The Journal of Neuroscience **29**: 13992-14003.

Cai, Y., A. D. Sheldon, Q. Yu and B. R. Postle (2019). "Overlapping and distinct contributions of stimulus location and of spatial context to nonspatial visual short-term memory." Journal of Neurophysiology **121**: 1222-1231.

Cai, Y., Q. Yu, A. D. Sheldon and B. R. Postle (unpublished). "The role of location-context binding in nonspatial visual working memory." bioRxiv.

Chao, L. L. and R. T. Knight (1995). "Human prefrontal lesions increase distractibility to irrelevant sensory inputs." NeuroReport **6**: 1605-1610.

Chelazzi, L., J. Duncan, E. K. Miller and R. Desimone (1998). "Responses of neurons in inferior temporal cortex during memory-guided visual search." Journal of Neurophysiology **80**: 2918-2940.

Chelazzi, L., E. K. Miller, J. Duncan and R. Desimone (1993). "A neural basis for visual search in inferior temporal cortex." Nature **363**: 345-347.

Christophel, T. B. and J. D. Haynes (2014). "Decoding complex flow-field patterns in visual working memory." NeuroImage **91**: 43-51.

Christophel, T. B., M. N. Hebart and J.-D. Haynes (2012). "Decoding the contents of visual short-term memory from human visual and parietal cortex." The Journal of Neuroscience **32**: 2983-12989.

Christophel, T. B., P. Iamshchinina, C. Yan, C. Allefeld and J.-D. Haynes (2018). "Cortical specialization for attended versus unattended working memory." Nature Neuroscience **21**: 494-496.

Christophel, T. B., P. C. Klink, B. Spitzer, P. R. Roelfsema and J.-D. Haynes (2017). "The distributed nature of working memory." Trends in Cognitive Sciences **21**: 111-124.

Cohen, J. D., S. D. Forman, T. S. Braver, B. J. Casey, D. Servan-Schreiber and D. C. Noll (1994). "Activation of the prefrontal cortex in a nonspatial working memory task with functional MRI." Human Brain Mapping **1**: 293-304.
Cohen, M. X. (2014). Analyzing Neural Time Series Data: THeory and Practice. Cambridge, MA, MIT Press.
Constantinidis, C., S. Funahashi, D. Lee, J. D. Murray, X.-L. Qi, M. Wang and A. F. T. Arnsten (2018). "Persistent spiking activity underlies working memory." Journal of Neuroscience **38**: 7020-7028.
Cowan, N. (2019). "Short-term memory based on activated long-term memory: A review in response to Norris (2017)." Psychological Bulletin **145**: 822-847.
Cudeiro, J. and A. M. Sillito (2006). "Looking back: corticothalamic feedback and early visual processing." Trends in Neuroscience **29**(6): 298-306.
Çukur, T., S. Nishimoto, A. G. Huth and J. L. Gallant (2013). "Attention during natural vision warps semantic representation across the human brain." Nature Neuroscience **16**: 763-770.
D'Esposito, M. and B. R. Postle (2015). "The cognitive neuroscience of working memory." Annual Review of Psychology **66**: 115-142.
D'Esposito, M., B. R. Postle, D. Ballard and J. Lease (1999). "Maintenance versus manipulation of information held in working memory: an event-related fMRI study." Brain & Cognition **41**: 66-86.
Davachi, L., L. M. Romanski, M. V. Chafee and P. S. Goldman-Rakic (2004). Domain specificity in cognitive systems. The Cognitive Neurosciences III. M. S. Gazzaniga. Cambridge, MA, The MIT Press**:** 665-678.
Dumoulin, S. O. and B. A. Wandell (2008). "Population receptive field estimates in human visual cortex." NeuroImage **39**: 647-660.
Emrich, S. M., A. C. Riggall, J. J. Larocque and B. R. Postle (2013). "Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory." Journal of Neuroscience **33**: 6516-6523.
Erickson, M. A., L. A. Maramara and J. Lisman (2010). "A single brief burst induces glur1-dependent associative short-term potentiation: A potential mechanism for short-term memory." Journal of Cognitive Neuroscience **22**: 2530-2540.
Ester, E. F., T. C. Sprague and J. T. Serences (2015). "Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory." Neuron **87**: 893-905.
Ester, E. F., D. W. Sutterer, J. T. Serences and E. Awh (2016). "Feature-selective attentional modulations in human frontoparietal cortex." Journal of Neuroscience **36**: 8188-8199.
Farah, M. (1990). Visual agnosia. Cambridge, MA, MIT Press.
Fiebelkorn, I. C. and S. Kastner (2019). "A rhythmic theory of attention." Trends in Cognitive Sciences **23**: 87-101.
Funahashi, S., C. J. Bruce and P. S. Goldman-Rakic (1989). "Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex." Journal of Neurophysiology **61**: 331-349.
Funahashi, S., C. J. Bruce and P. S. Goldman-Rakic (1990). "Visuospatial coding in primate prefrontal neurons revealed by oculomotor paradigms." Journal of Neurophysiology **63**: 814-831.
Gardner, J. L. and T. Liu (2019). "Inverted Encoding Models Reconstruct an Arbitrary Model Response, Not the Stimulus." eNeuro **6**.

Goldman-Rakic, P. S. (1992). "Working memory and the mind." Scientific American **267**: 110-117.
Gosseries, O., Q. Yu, J. J. LaRocque, M. J. Starrett, N. Rose, N. Cowan and B. R. Postle (2018). "Parieto-occipital interactions underlying control- and representation-related processes in working memory for nonspatial visual features." Journal of Neuroscience **38**: 4357-4366.
Hamidi, M., G. Tononi and B. R. Postle (2008). "Evaluating frontal and parietal contributions to spatial working memory with repetitive transcranial magnetic stimulation." Brain Research **1230**: 202-210.
Hamker, F. H. (2005). "The reentry hypothesis: The putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement." Cerebral Cortex **15**: 431-447.
Hanning, N. M., D. Jonikaitis, H. Deubel and M. Szinte (2016). "Oculomotor selection underlies feature retention in visual working memory." Journal of Neurophysiology **115**: 1071-1076.
Harrison, S. A. and F. Tong (2009). "Decoding reveals the contents of visual working memory in early visual areas." Nature **458**: 632-635.
Jerde, T., E. P. Merriam, A. C. Riggall, J. H. Hedges and C. E. Curtis (2012). "Prioritized Maps of Space in Human Frontoparietal Cortex." The Journal of Neuroscience **32**: 17382-17390.
Jonides, J., E. Smith, R. Koeppe, E. Awh, S. Minoshima and M. Mintum (1993). "Spatial working memory in humans as revealed by PET." Nature **363**: 623-625.
Jonikaitis, D. and T. Moore (2019). "The interdependence of attention, working memory and gaze control: behavior and neural circuitry." Current Opinion in Psychology **29**: 126-134.
Katsuki, F. and C. Constantinides (2012). "Unique and shared roles of the posterior parietal and dorsolateral prefrontal cortex in cognitive functions." Frontiers in Integrative Neuroscience **6**.
Koyluoglu, O. O., Y. Pertzov, S. M. Manohar, M. Husain and I. R. Fiete (2017). "Fundamental bound on the persistence and capacity of short-term memory stored as graded persistent activity." eLife **6**: e22225.
LaRocque, J. J., J. A. Lewis-Peacock, A. Drysdale, K. Oberauer and B. R. Postle (2013). "Decoding attended information in short-term memory: An EEG study." Journal of Cognitive Neuroscience **25**: 127-142.
LaRocque, J. J., A. C. Riggall, S. M. Emrich and B. R. Postle (2017). "Within-category decoding of information in different states in short-term memory." Cerebral Cortex **17**: 4881-4890.
Leavitt, M. L., D. Mendoza-Halliday and J. C. Martinez-Trujillo (2018). "Sustained activity encoding working memories: not fully distributed." Trends in Neurosciences **40**: 328-346.
Lewis-Peacock, J. A., A. T. Drysdale, K. Oberauer and B. R. Postle (2012). "Neural evidence for a distinction between short-term memory and the focus of attention." Journal of Cognitive Neuroscience **24**: 61-79.
Libby, A. and T. J. Buschman (unpublished). "Rotational dynamics reduce interference between sensory and memory representations." bioRxiv.
Liu, T., D. Cable and J. L. Gardner (2018). "Inverted Encoding Models of Human Population Respons Conflate Noise and Neural Tuning Width." Journal of Neuroscience **38**: 398–408.
Lowet, E., B. Gomes, K. Srinivasan, H. Zhou, R. J. Schafer and R. Desimone (2018). "Enhanced neural processing by covert attention only during microsaccades directed toward the attended stimulus." Neuron **99**: 207–214.

Luck, S. J. and E. K. Vogel (2013). "Visual working memory capacity: from psychophysics and neurobiology to individual differences." Trends in Cognitive Sciences **17**: 391-400.
Luria, R., H. Balaban, E. Awh and E. K. Vogel (2016). "The contralateral delay activity as a neural measure of visual working memory." Neuroscience and Biobehavioral Reviews **62**: 100-108.
Mackey, W., O. Devinsky, W. Doyle, M. Meager and C. E. Curtis (2016). "Human dorsolateral prefrontal cortex is not necessary for spatial working memory." Journal of Neuroscience **36**: 2847-2856.
Malmo, R. B. (1942). "Interference factors in delayed response in monkey after removal of the frontal lobes." Journal of Neurophysiology **5**: 295-308.
Manohar, S. G., N. Zokaei, S. J. Fallon, T. P. Vogels and M. Husain (2019). "Neural mechanisms of attending to items in working memory." Neuroscience and Biobehavioral Reviews **101**: 1-12.
Masse, N. Y., G. R. Yang, H. F. Song, X.-J. Wang and D. J. Freedman (2019). "Circuit mechanisms for the maintenance and manipulation of information in working memory." Nature Neuroscience **22**: 1159-1167.
Mendoza-Halliday, D. and J. C. Martinez-Trujillo (2017). "Neuronal population coding of perceived and memorized visual features in the lateral prefrontal cortex." Nature Communications **8**.
Mendoza-Halliday, D., S. Torres and J. C. Martinez-Trujillo (2014). "Sharp emergence of feature-selective sustained activity along the dorsal visual pathway." Nature Neuroscience **17**: 1255-1262.
Merrikhi, Y., K. Clark, E. Albarran, M. Parsa, M. Zirnsak, T. Moore and B. Noudoost (2017). "Spatial working memory alters the efficacy of input to visual cortex." Nature Communications **8**: 15041.
Mirpour, K., S. Bolandnazar and J. W. Bisley (2019). "Neurons in FEF keep track of items that have been previously fixated in free viewing visual search." Journal of Neuroscience **39**: 2114-2124.
Mongillo, G., O. Barak and M. Tsodyks (2008). "Synaptic theory of working memory." Science **319**: 1543-1546.
Moore, T. and K. M. Armstrong (2003). "Selective gating of visual signals by microstimulation of frontal cortex." Nature **421**: 370-373.
Moore, T. and M. Zirnsak (2017). "Neural mechanisms of selective visual attention." Annual Review of Psychology **68**: 47-72.
Murray, J. D., A. Bernacchia, N. A. Roy, C. Constantinidis, R. X. Romo and X.-J. Wang (2017). "Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex." Proceedings of the National Academy of Sciences (USA) **114**: 394-399.
Myers, N. E., M. G. Stokes and A. C. Nobre (2017). "Prioritizing information during working memory: Beyond sustained internal attention." Trends in Cognitive Sciences **21**: 449-461.
Norman, K. A., S. M. Polyn, G. J. Detre and J. V. Haxby (2006). "Beyond mind-reading: multi-voxel pattern analysis of fMRI data." Trends in Cognitive Sciences **10**: 424-430.
Norris, D. (2017). "Short-term memory and long-term memory are still different." Psychological Bulletin **143**: 992-1009.
Noudoost, B. and T. Moore (2011). "Control of visual cortical signals by prefrontal dopamine." Nature **474**: 372-375.

O'Reilly, R. C. and Y. Munakata (2000). Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating Brains. Cambridge, MA, MIT Press.
Panichello, M. F., B. DePasquale, J. W. Pillow and T. J. Buschman (2019). "Error-correcting dynamics in visual working memory." Nature Communications **10**: 3366.
Pereira, F., T. Mitchell and M. M. Botvinick (2009). "Machine learning classifiers and fMRI: a tutorial overview." NeuroImage **45**: S199-S209.
Postle, B. R. (2006). "Working memory as an emergent property of the mind and brain." Neuroscience **139**: 23-38.
Postle, B. R. (2015). "The cognitive neuroscience of visual short-term memory." Current Opinion in Behavioral Sciences **1**: 40-46; doi:10.1016/j.cobeha.2014.1008.1004.
Postle, B. R., M. D'Esposito and S. Corkin (2005). "Effects of verbal and nonverbal interference on spatial and object visual working memory." Memory & Cognition.
Postle, B. R. and M. Hamidi (2007). "Nonvisual codes and nonvisual brain areas support visual working memory." Cerebral Cortex **17**: 2134-2142.
Postle, B. R., C. Idzikowski, S. Della Salla, R. H. Logie and A. D. Baddeley (2006). "The selective disruption of spatial working memory by eye movements." Quarterly Journal of Experimental Psychology **59**: 100-120.
Riggall, A. C. and B. R. Postle (2012). "The relationship between working memory storage and elevated activity as measured with funtional magnetic resonance imaging." The Journal of Neuroscience **32**: 12990-12998.
Riley, M. R. and C. Constantinidis (2016). "The role of prefrontal persistent activity in working memory." Frontiers in Systems Neuroscience.
Riley, M. R., X.-L. Qi and C. Constantinidis (2017). "Functional specialization of areas along the anterior-posterior axis of the primate prefrontal cortex." Cerebral Cortex **27**: 3683-3697.
Rose, N., J. J. Larocque, A. C. Riggall, O. Gosseries, M. J. Starrett, E. Meyering and B. R. Postle (2016). "Reactivation of latent working memories with transcranial magnetic stimulation." Science **354**: 1136-1139.
Rougier, N. P., D. C. Noelle, T. S. Braver, J. D. Cohen and R. C. O'Reilly (2005). "Prefrontal cortex and flexible cognitive control: Rules without symbols." Proceedings of the National Academy of Sciences (USA) **102**: 7338-7343.
Sahan, M. I., A. D. Sheldon and B. R. Postle (unpublished). "The neural consequences of attentional prioritization of internal representations in visual working memory." bioRxiv.
Scolari, M., E. F. Ester and J. T. Serences (2014). Feature- and object-based attentional modulation in the human visual system. The Oxford Handbook of Attention. A. C. Nobre and S. Kastner.
Serences, J. T. (2016). "Neural mechanisms of information storage in visual short-term memory." Vision Research **128**: 53-67.
Serences, J. T., E. F. Ester, E. K. Vogel and E. Awh (2009). "Stimulus-specific delay activity in human primary visual cortex." Psychological Science **20**: 207-214.
Serences, J. T. and S. Saproo (2012). "Computational advances towards linking BOLD and behavior." Neuropsychologia **50**(4): 435-446.
Servant, M., P. Cassey, G. D. Logan and G. F. Woodman (2018). "The neural bases of automaticity." Journal of Experimental Psychology: Learning, Memory, and Cognition **44**: 440-464.

Sillito, A. M., J. Cudeiro and H. E. Jones (2006). "Always returning: feedback and sensory processing in visual cortex and thalamus." Trends in Neuroscience **29**(6): 307-316.
Sillito, A. M., H. E. Jones, G. L. Gerstein and D. C. West (1994). "Feature-linked synchronization of thalamic relay cell firing induced by feedback from the visual cortex." Nature **369**(6480): 479-482.
Sprague, T. C., K. C. S. Adam, J. J. Foster, M. Rahmati, D. W. Sutterer and V. A. Vo (2018). "Inverted Encoding Models Assay Population-Level Stimulus Representations, Not Single-Unit Neural Tuning." eNeuro **5**.
Sprague, T. C., G. M. Boynton and J. T. Serences (in press). "Inverted encoding models estimate sensible channel responses for sensible models." eNeuro.
Sprague, T. C. and J. T. Serences (2013). "Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices." Nature Neuroscience **16**: 1879-1887.
Stokes, M. G. (2015). "'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework." Trends in Cognitive Sciences **19**: 394-405.
Theeuwes, J., C. N. L. Olivers and C. L. Chizk (2005). "Remembering a location makes the eyes curve away." Psychological Science **16**: 196-199.
Toda, K., Y. Sugase-Miyamoto, T. Mizuhiki, K. Inaba, B. J. Richmond and M. Shidara (2012). "Differential encoding of factors influencing predicted reward value in monkey rostral anterior cingulate cortex." PLoS One **7**(1): e30190.
Todd, J. J. and R. Marois (2004). "Capacity limit of visual short-term memory in human posterior parietal cortex." Nature **428**: 751-754.
Todd, J. J. and R. Marois (2005). "Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity." Cognitive, Affective, & Behavioral Neuroscience **5**: 144-155.
Tsujimoto, S. and B. R. Postle (2012). "The prefrontal cortex and delay tasks: a reconsideration of the "mnemonic scotoma"." Journal of Cognitive Neuroscience **24**: 627-635.
van Ede, F., S. R. Chekroud, M. G. Stokes and A. C. Nobre (2019). "Concurrent visual and motor selection during visual working memory guided action." Nature Neuroscience **22**: 477-483.
van Kerkoerle, T., M. W. Self and P. R. Roelfsema (2017). "Layer-specificity in the effects of attention and working memory on activity in primary visual cortex." Nature Communications **8**: 13804.
van Loon, A. M., K. Olmos-Solis, J. J. Fahrenfort and C. N. L. Olivers (2018). "Current and future goals are represented in opposite patterns in object-selective cortex." eLife **7**: e38677.
Vogel, E. K. and M. G. Machizawa (2004). "Neural activity predicts individual differences in visual working memory capacity." Nature **428**: 748-751.
Vogel, E. K., A. W. McCollough and M. G. Machizawa (2005). "Neural measures reveal individual differences in controlling access to working memory." Nature **438**: 368-387.
Wan, Q., Y. Cai, T. T. Rogers and B. R. Postle (unpublished). "Rotational remapping as a candidate mechanism for priority-based recoding in visual working memory: empirical and computational evidence."
Wan, Q., Y. Cai, J. Samaha and B. R. Postle (in-principle accepted). "Tracking stimulus representation across a 2-back visual working memory task." Royal Society Open Science **in-principle accepted registered report**.

Wang, J. X., Z. Kurth-Nelson, D. Kumaran, D. Tirumala, H. Soyer, J. Z. Leibo, D. Hassabis and M. Botvinick (2018). "Prefrontal cortex as a meta-reinforcement learning system." Nature Neuroscience **21**: 860-868.
Wilson, F. A. W., S. P. O'Scalaidhe and P. S. Goldman-Rakic (1993). "Dissociation of object and spatial processing domains in primate prefrontal cortex." Science **260**: 1955-1958.
Wolff, M. J., J. Ding, N. E. Myers and M. G. Stokes (2015). "Revealing hidden states in visual working memory using electroencephalography." Frontiers in Systems Neuroscience **9**.
Wolff, M. J., J. Jochim, E. G. Akyürek and M. G. Stokes (2017). "Dynamic hidden states underlying working-memory-guided behavior." Nature Neuroscience.
Woodman, G. F. (2013). "Viewing the dynamics and control of visual attention through the lens of electrophysiology." Vision Research **80**: 7-18.
Xu, Y. (2017). "Reevaluating the sensory account of visual working memory storage." Trends in Cognitive Sciences **27**: 794-815.
Xu, Y. (2018). "Sensory cortex is nonessential in working memory storage." Trends in Cognitive Sciences **22**: 192-193.
Xu, Y. and M. M. Chun (2006). "Dissociable neural mechanisms supporting visual short-term memory for objects." Nature **440**: 91-95.
Yu, Q. and B. R. Postle (2018). "Different states of priority recruit different neural codes in visual working memory." bioRxiv.
Yu, Q. and B. R. Postle (unpublished). "Different states of priority recruit different neural codes in visual working memory." bioRxiv.