Neural Mechanisms of Prioritization in Working Memory

By

Quan Wan

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Psychology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2023

Date of final oral examination: 12/14/2023

The dissertation is approved by the following members of the Final Oral Committee:

Bradley R. Postle, Professor, Psychology & Psychiatry
Yuri B. Saalmann, Associate Professor, Psychology
Timothy T. Rogers, Professor, Psychology
Joseph L. Austerweil, Associate Professor, Psychology
Matthew I. Banks, Professor, Anesthesiology

Dedicated to my partner Gavin and my parents Fengrong and Xiaofeng.

Acknowledgements

My years in Madison have been a thrilling, rewarding, and humbling journey, along which I have harvested much knowledge and growth. I am truly grateful to my professors, colleagues, administrators, and friends in and outside of the Postle lab. This thesis would not have been possible without all their help and support.

First and foremost, I feel endlessly fortunate to have had Brad as my PhD advisor. His passion for science and commitment to integrity and equality have shined like a beacon over my academic and professional path. It has been a privilege to be a recipient of his masterful mentoring and he has taught me to be a more rigorous thinker and a courageous explorer into uncharted territories. I am deeply indebted to his unwavering love and support throughout the years, especially during the difficult times. My heartfelt appreciation to my other committee members: Yuri Saalmann, Tim Rogers, Joe Austerweil and Matt Banks for their precious input and training.

I would like to thank every member of the Postle lab, for all their kind help and friendship: Jason Samaha, Qing Yu, Ying Cai, Jackie Fulvio, Chunyue Teng, Mattia Pietrelli, Jiangang Shan and Yun Ding. I have learned so much from them all and greatly enjoyed their company.

I have also been blessed with amazing collaborators: Jorge Menendez and Adel Ardalan have imparted to me much wisdom and knowledge on training RNNs and dimensionality reduction, upon which this thesis is heavily built. It has been fun and intellectually stimulating to work with them both. Thank you also to Helena Olraun for embarking on a challenging project together with me.

Finally, and most of all, my deepest gratitude to my loving partner, Gavin, for always being there for me and believing in me, to my friends who always lend a listening ear, and to my parents in China, for their unconditional love and indispensable support.

Table of Contents

Abstract

The ability to prioritize among contents in working memory (WM) is critical for successful control of thought and behavior. Recent work has demonstrated that prioritization in WM can be implemented by representing different states of priority in different representational formats. However, its neural mechanisms remain unclear on an algorithmic level. In this thesis, I take a novel approach to studying WM prioritization by turning to artificial neural networks (ANNs) for mechanistic insights. Chapter 1 sets up the cognitive and neural architectures of WM and introduces the problem of WM prioritization and extant neuroimaging work on the neural representations of differently prioritized memory items. Chapter 2 presents an empirical study where I trained recurrent neural networks (RNN) with a long short-term memory (LSTM) architecture to perform the 2-back WM task. Visualization of LSTM hidden layer activity using principal component analysis (PCA) and demixed PCA had confirmed that stimulus representations in RNNs also undergo representational transformations – a reversal of stimulus coding axes – when transitioning between priority states. In Chapter 3, I further explored the mechanisms underlying WM prioritization by simulating the double serial retrocuing (DSR) task with RNNs. PCA visualization of stimulus representational dynamics revealed that the network represented trial context (order of presentation) and priority via different mechanisms. Ordinal context was accomplished by *segregating* representations into orthogonal subspaces, and priority by *separating* representations into different manifolds within each subspace. We assessed the generality of these mechanisms by applying dimensionality reduction and multiclass decoding to fMRI and EEG datasets and found that priority and context are represented differently along the dorsal visual stream, and that behavioral performance is sensitive to trial-by-trial efficacy of priority coding, but not context coding. Chapter 4 discusses the significance of context coding in WM and future directions in studying the control of WM prioritization. The work presented in this thesis significantly adds to our mechanistic understanding of the neural bases of WM prioritization.

Chapter 1

Introduction

Working memory (WM) refers to our ability to temporarily retain a limited amount of information in the absence of sensory input, to flexibly update and manipulate it, and to use it to guide behavior. Often thought of as a stable individual trait, WM ability is predictive of many laboratory and real-world metrics, including general fluid intelligence (Cowan et al., 2006; Fukuda et al., 2010; Shipstead et al., 2012; Shipstead & Engle, 2013) and academic achievement (Gathercole & Pickering, 2000). WM is also an essential component of many cognitive functions, including cognitive control, problem solving and planning. Deficits in WM are characteristic of many neurological and psychiatric disorders, including Alzheimer's Disease (AD), Attention Deficit/Hyperactivity Disorder (ADHD), Major Depressive Disorder (MDD) and schizophrenia (e.g., Devinsky et al., 2003; Gold et al., 2019). Due to its substantial real-world relevance, WM is widely studied in psychology, neuroscience, and medicine.

The cognitive architecture of WM

The most influential model that characterizes the cognitive architecture of WM is the multiple component model, which was initially proposed by Baddeley and Hitch (1974) and later revised and updated a few times by Baddeley and collaborators (e.g., Baddeley, 2000, 2007; Baddeley & Hitch, 2007, 2019). According to this model, WM is implemented in several domain-specific sub-systems that operate as buffers for both

the storage and manipulation of information. They include the phonological loop for verbalizable information, the visuospatial sketchpad for visuospatial information and the later added episodic buffer. In addition, a Central Executive serves to coordinate among and manipulate the contents of these sub-systems. More specifically, the phonological loop comprises a phonological store for the storage function and an articulatory loop for rehearsal of verbal information. The visuospatial sketchpad echoes the "what" and "where" of visual processing and can be further divided into a visual cache for object features and an inner scribe to hold spatiotemporal information (Logie, 2003). The episodic buffer serves to integrate information across different domains to form new structural representations (which requires the work of the Central Executive) and buffer information retrieved from episodic and semantic long-term memory (LTM; Baddeley, 2000).

In contrast to the memory-systems perspective that the Baddeley model provides stand the state-dependent theoretical frameworks, from which this dissertation approaches WM. An influential model under this framework is Cowan's embedded-processes model (Cowan, 1988, 1995, 1999), which emphasizes links between attention and LTM. His model posits that working memory emerges from the temporary activation of the preexisting representations in the cognitive system, which makes the critical distinction between the "activated LTM" (aLTM) and the Focus of Attention

(FoA). The aLTM refers to representations that are only activated whereas FoA represents a privileged state where information has a limited capacity, and can be manipulated and access awareness. aLTM and FoA differ in a few respects: (1) FoA's capacity is limited to about 4 discrete chunks whereas aLTM is not subject to such limits; (2) information in aLTM can be forgotten via interference and decay, against which FoA contents are protected; (3) aLTM only includes simply activated long-term knowledge, while FoA is capable of binding information together and forming novel structural representations.

Another state-dependent theoretical framework for WM was proposed by Oberauer (2002, 2005, 2013; Figure 1.1), which distinguishes three states of WM information: activated LTM (similar to Cowan's model), the region of direct access, and the focus of attention. The region of access operates by generating and keeping temporary bindings to create new representations. The interference between bindings determines its limited capacity. The focus of attention, on the other hand, is a mechanism that can select specific components of the representation currently maintained in the region of access.

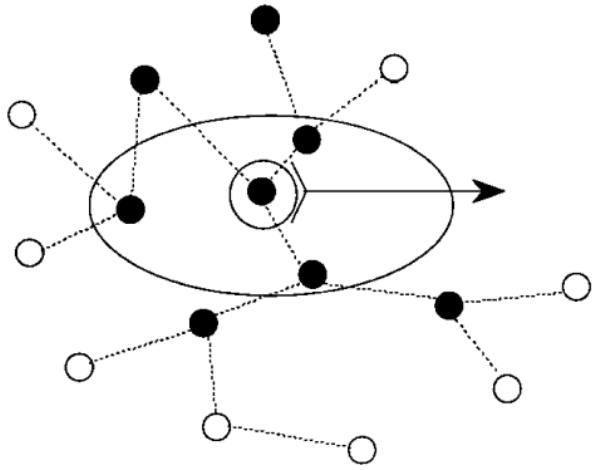


Figure 1.1. A concentric model of working memory. Nodes and lines represent a network of long-term memory representations, with black nodes indicating activated representations. Some of the black-node items are within the region of direct access (large oval). Within the region of direct access, one item is selected for processing by the focus of attention (small oval). Activated items outside the region of direct access form the activated part of long-term memory, accessible only indirectly through associative links (dotted lines) to representations in the more central regions. Figure and caption adapted from Oberauer, 2002.

The neural architecture of WM

A current dominant account for the neural architecture of WM are the sensorimotor recruitment models. The fundamental premise of these models is, congruent with the state-dependent perspectives, that "the systems and representations engaged to perceive information can also contribute to the short-term retention of that information (D'Esposito & Postle, 2015)." This account has been supported by a wealth of cognitive neuroscience studies (see Postle, 2006 for a review of evidence on the systems level). For example, using multivariate pattern analysis (MVPA), two studies showed that on delayed recognition tasks, representations of the color or orientation of target stimuli can be found in the primary visual cortex (V1) across the delay period

(Harrison & Tong, 2009; Serences et al., 2009). These results were replicated with other types of stimuli: for example, the short-term maintenance of complex visuospatial patterns can be decoded from the occipital and parietal cortices (Christophel et al., 2012) and the short-term retention of familiar objects, houses, faces, scenes, and body stimuli are decodable from the ventral occipitotemporal cortex (Han et al., 2013; Lee et al., 2013; Nelissen et al., 2013; Sreenivasan et al., 2014).

The important role of prefrontal cortex (PFC) in WM originates from two early studies in monkey electrophysiology. In one study, Fuster and Alexander (1971) showed that PFC neurons sustained firing during the active maintenance period of a delayed-response task where information was no longer present but still relevant for completing the task. A similar finding from a delayed alternation task was reported by Kubota and Niki (1971). It has long been thought that the persistent activity in PFC serves the storage function of working memory, but this role has been called into question by studies that support the sensorimotor recruitment models of WM. As mentioned before, Harrison and Tong (2009) and Serences et al. (2009) were able to decode visual stimulus information from V1 during the delay-period despite the lack of persistent elevated signals in this region. Later studies (Emrich et al., 2013; Riggall & Postle, 2012) not only replicated these findings but also failed to find stimulus information from the elevated delay-period activity in frontoparietal areas. These

findings imply that persistent PFC activity might serve functions other than storage per se. For example, it might represent various task variables that are not directly related to the memoranda such as task rules (Warden & Miller, 2010), contingent motor responses (Romo et al., 1999), and stimulus-response mappings (Wallis et al., 2001). Koechlin and colleagues (2003) have hypothesized that the frontal cortex might be organized rostro-caudally in a hierarchical manner all the way leading to action (also see Fuster, 2004). Furthermore, PFC has long been understood to provide a source of top-down signals to influence processing in other cortical and subcortical brain regions (Braver et al., 2008; Duncan, 2001; Shallice, 1982) and support the cognitive control functions of working memory, such as selective attention, updating, inhibition, manipulation etc., with contributions from parietal cortex (Collette et al., 2005; Koenigs et al., 2009) and basal ganglia (Chatham et al., 2014; Chatham & Badre, 2015).

Neural bases of WM prioritization

One hallmark of WM is the ability to flexibly prioritize among its contents to appropriately guide behavior. To accomplish this, one needs to keep information in a readily accessible state while preventing it from interfering with ongoing behavior. This is a cognitive operation that we engage in frequently throughout our daily life. For example, say that you've just completed a talk at a conference, and you see two people

simultaneously approaching each of two microphones to ask a question. You turn to the moderator and wait for them to indicate who will ask the first question, and based on this your shift of gaze is guided by your memory of the location of the cued microphone. In the meantime, you keep the location of the uncued microphone in working memory which you will turn to later while you answer the first question. Then after you've answered the question, the moderator cues the other microphone, and you use your memory of its location to shift your gaze again. In this example, the memory of the locations of microphones is dynamically prioritized/deprioritized according to the moderator's cue. This thesis centers on how this process is accomplished in the brain. What neural mechanisms enable the flexible prioritization of contents in working memory?

As the example above illustrates, working memory has an intertwined relationship with attention (Oberauer, 2019). As mentioned above, working memory has been thought to be supported by sustained, elevated neuronal activity that persists throughout the retention interval (Curtis & D'Esposito, 2003; Fuster & Alexander, 1971; Vogel et al., 2005). But this theory is running the risk of confounding working memory with attention, because in most studies that demonstrate delay-period activity, the memoranda are often behaviorally relevant, thus in the "focus of attention". This calls

into question whether the delay-period activity is reflecting selective attention, rather than working memory per se.

Indeed, a series of neuroimaging studies exploring the effect of prioritization of working memory have demonstrated that unprioritized information in WM can be maintained without an "active neural trace," corresponding to the idea of "activated LTM". These studies often employ a retrocuing task. A dual serial retrocuing (DSR) task begins with the presentation of two sample items, followed by a retrocue that signals which of the two items will be tested first, rendering it the "prioritized memory item" (PMI). Because there is a .5 probability that the initially unprioritized memory item (UMI) may be tested later in the trial, it is assumed that the UMI is retained in working memory. If the same item (PMI) is tested later, it is termed a 'repeat' trial, whereas if the other item (UMI) is tested later, it is called a 'switch' trial.

In a functional magnetic resonant imaging (fMRI) study, Lewis-Peacock and colleagues (2012) used English words, pronounceable pseudowords and line segments as stimuli and had subjects perform a short-term recognition before the retrocuing task. Using multivariate pattern analysis (MVPA), they trained linear classifiers on data from the recognition task and tested them throughout the DSR task time course. They found that upon the presentation of the sample items, the classifier evidence for both items was elevated above chance level (Figure 1.2). Then, after the first cue, evidence for the

initially cued item (PMI) stays elevated while that for the uncued item (UMI) drops to baseline levels (and cannot be distinguished from evidence for the item from the irrelevant category). After the second cue, in 'repeat' trials, initially prioritized item can still be decoded with above-chance classifier evidence; in 'switch' trials, intriguingly, the initially uncued item can now be significantly decoded while the evidence for the other item drops to baseline.

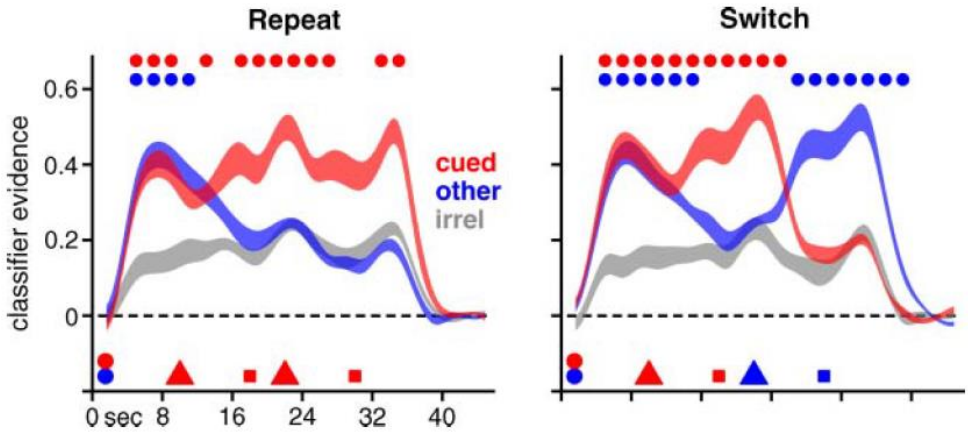


Figure 1.2. Classifier decoding performance for Lewis-Peacock et al. (2012) Experiment 2. Results from both repeat (left) and switch (right) trials are shown. Classifier evidence values for phonological, semantic and visual stimuli were relabeled and collapsed across all trials into three new categories: *cued* (red, the category of the memory item selected by the first cue), *other* (blue, the category of the other presented item), and *irrel* (gray, the category irrelevant for the trial). The colored shapes along the horizontal axis indicate the onset of targets (red and blue circles), the first cue (red triangle) and the first recognition probe (red square), the second cue (red or blue triangle) and the last recognition probe (red or blue square). Data for each category are shown as ribbons and their thickness show ±1 *SEM* across subjects, interpolated across 23 discrete data points in the trial-averaged data. Statistical comparisons of evidence values focused on within-subject differences. The color-coded circles on the top of each figure show that the classifier's evidence for the *cued* or *other* categories, respectively, was reliably stronger than the evidence for the trial-irrelevant category (*irrel*). Figure and caption adapted from Lewis-Peacock et al. (2012).

This finding of the lack of an active representation of the UMI is at odds with the idea that all items held in working memory are represented with an active trace, and was replicated a number of times with both fMRI (LaRocque et al., 2017; Rose et al., 2016) and with EEG (Larocque et al., 2014; Rose et al., 2016). The fact that this initially uncued item remains in working memory is demonstrated in several ways. First, as mentioned before, an active trace of this item returns when it is cued by the second cue of the DSR task, and recognition performance on such "switch" trials is almost as high as is performance on the second probe of "repeat" trials, when the same item is cued twice (e.g., Larocque et al., 2014, 2017; Lewis-Peacock et al., 2011). Second, when a single pulse of TMS is applied during the first delay, MVPA evidence of the UMI is transiently elevated, and on these trials the false-alarm rate UMI foil probes is higher than it is on no-TMS trials (Rose et al., 2016).

It has been proposed that the UMI could be represented via "activity-silent" mechanisms (e.g., Stokes, 2015), perhaps, for example, through short-lived synaptic modifications (Mongillo et al., 2008; Barak & Tsodyks, 2014), similar to long-term memory. Specifically, in this computational model, encoding activity temporarily alters synaptic activity of the neural network, creating a short-lived synaptic memory trace through activity-dependent short-term synaptic plasticity (STSP). According to this model, activity-silent WM mechanisms are more energetically efficient than active

coding schemes such as persistent neuronal activity. Supposedly, an "activity-silent" trace (Myers et al., 2017) might be less likely to interfere with the currently active "prioritized memory item" (PMI). Although some previous neuroimaging studies have reported that the prioritization of one item held in WM leads to a decrease-to-baseline of the activity level of the UMI (LaRocque et al., 2013; Lewis-Peacock et al., 2012; Rose et al., 2016), whether an "activity-silent" mechanism may contribute functionally to WM remains a topic of vigorous debate (Christophel et al., 2018; Schneegans & Bays, 2017b; Sprague et al., 2016; Stokes et al., 2020).

More recently, however, studies presenting evidence for an active trace of the UMI have begun to emerge. In one fMRI study, Christophel and colleagues (2018) reported that, whereas an active trace of only the PMI can be decoded from early visual areas V1-V4, the UMI (together with the PMI) can be decoded from the intraparietal sulcus (IPS) and frontal eye fields (FEF). From this, they suggested that "sensory cortex maintains a high-resolution representation of the currently attended memory item, whereas parietal cortex has low-resolution representations of both attended and unattended items" (p. 496). A different possibility, one that we will refer to as "priority-based remapping," is suggested by two other recent fMRI studies.

In one study, van Loon and colleagues (2018) acquired functional magnetic resonance imaging (fMRI) data while first presenting subjects two target images

sequentially (e.g., first a flower then a cow), then indicating with a cue whether memory for the first or second presented image would be tested first (Figure 1.3A). Had the cue been a "1", subjects would next see a test array of six flowers and indicate whether the target flower appeared in the test array, and finally a test array of six cows. On this trial, the target cow spent time as UMI, because the cue indicated that memory for the flower would be tested first. When van Loon et al. (2018) applied multivariate pattern analysis (MVPA) to fMRI data from posterior ventral temporal lobe, they found that a decoder trained on trials when an item was a PMI performed statistically below chance when that item was a UMI. Furthermore, a representational dissimilarity analysis indicated that, within their set of 12 stimuli (four cows, four skates, four dressers), each item's high-dimensional representation in one state (e.g., as a PMI) was maximally different from its representation in the other state (i.e., as a UMI). Using a similar retrocuing procedure, Yu, Teng and Postle (2020) found, with multivariate inverted encoding modeling (IEM) of fMRI data from early visual cortex, that the reconstructed orientation of a grating "flipped" when it was a UMI relative to a PMI (e.g., a 30° orientation reconstructed as 120° while a UMI; Figures 1.3B and 1.3C). Furthermore, for data from the intraparietal sulcus (IPS), they observed that the IEM reconstruction of the location where an item had been presented also flipped when an item's priority status transitioned to UMI.

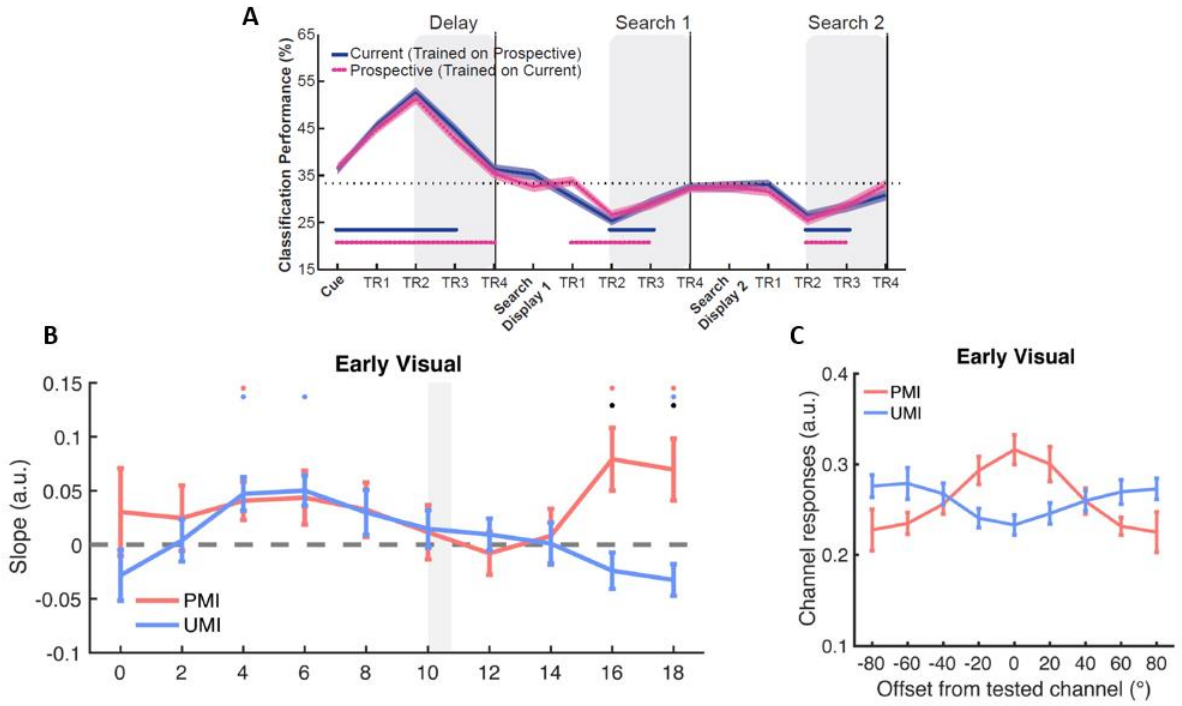


Figure 1.3. Neural evidence for active UMI representations. (A) Time course of cross-relevance category decoding from van Loon et al. (2018) Experiment 1. The classifier trained on current relevance (PMI), tested on prospective relevance (UMI), or vice versa. They observed above-chance decoding during the delay before search (suggesting similar UMI and PMI representations) but below-chance decoding during Search 1 and Search 2 (indicating opposite representational formats). Blue and pink shaded areas indicate within-subject SEM. Blue and pink horizontal lines at the bottom of the line graphs show timepoints significantly differing from chance ($p < .05$, uncorrected). (B) Time course of PMI-trained IEM reconstructions of stimulus orientation in early visual cortex from Yu, Teng and Postle (2020). Gray shaded area indicates the presentation of *Cue1*, during the *Delay 1.2* after which PMI can be positively reconstructed while negative UMI reconstructions were obtained. Red, blue and black dots show $p < .05$ for PMI and UMI reconstructions and the difference between the two, respectively. (C) PMI-trained IEM reconstructions of stimulus orientation in *Delay 1.2* (corresponding to 18 s in (B)) in early visual ROI from Yu, Teng and Postle (2020). All error-bars in (B) and (C) indicate SEM. Figure (A) and caption adapted from van Loon et al. (2018) and Figure (B)(C) and captions adapted from Yu, Teng and Postle (2020).

Shifts of priority are also characteristic of continuous-performance tasks, for which shifts of priority are dictated by task rules rather than by explicit cues. One example,

which features prominently in the work presented here, is the 2-back WM task from Wan and colleagues (2020). Electroencephalography (EEG) signals were recorded while subjects viewed the serial presentation of oriented gratings and judged for each one whether it was a match or a non-match to the item that had appeared two positions previously in the series. This task entails a predictable transition through priority states for each item: When an item *n* is initially presented, it serves as probe to compare against the memory of item *n – 2*; after the *n*-to-*n – 2* decision is made, item *n* becomes a UMI while item *n – 1* is prioritized for the upcoming comparison with *n + 1.* Next, once the *n + 1-to-n – 1* comparison is completed, item *n* becomes a PMI for its impending comparison with item *n + 2.* To analyze the EEG data, an IEM was trained on the raw EEG voltages from a separate 1-item delayed-recognition task, and then tested on the delay periods separating *n* and *n + 1* and separating *n + 1* and *n + 2* (i.e., when item *n* assumed the status of UMI, then PMI). The results, reminiscent of van Loon et al. (2018) and Yu, Teng and Postle (2020), indicated that the IEM reconstruction of the UMI was "flipped" relative to the training data (Figure 1.4). We referred to this transition from PMI to UMI as "priority-based remapping" (rather than "recoding" or "code morphing"; c.f. Parthasarathy et al., 2017), reasoning that the IEM reconstruction of the UMI would fail if it were represented in a neural code different from the trained model.

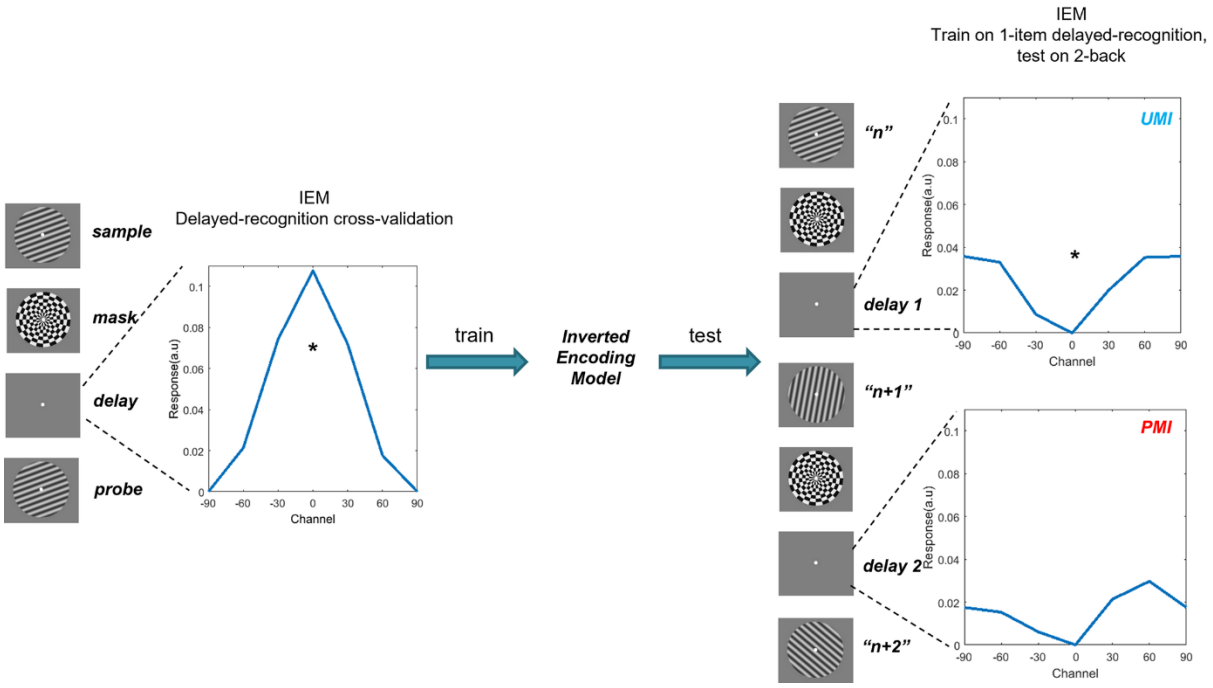


Figure 1.4. IEM reconstruction of EEG recorded during the 2-back task (N = 42). Left panel: IEM reconstruction of the stimulus during the delay in a separate one-item delayed-recognition task. This model was used to reconstruct the stimulus in the 2-back task. Right panel: Concatenation of the item *n* and item *n + 1* stimulus events to form a trial, across which *n* transitions from probe to UMI to PMI in the 2-back. On the right are IEM reconstructions corresponding to the two 2 s windows centered in two 2.8 s post-mask ISIs before and after item *n + 1*, respectively. "*" indicates $p < .01$ (two-tailed t test), FDR-corrected for multiple comparisons. As the figure shows, IEM reconstruction of stimulus *n* is "flipped" relative to the training data (IEM reconstruction from delayed recognition) when it is a UMI, demonstrating priority-based remapping. (Reconstruction of the PMI was unsuccessful.) For delayed-recognition IEM reconstruction (940 – 1040 ms from stimulus onset), $t(41) = 4.12$, $p < 0.001$. For UMI reconstruction of 2-back (-2400 – -400 ms relative to *n + 1* onset), $t(41) = -3.02$, $p = 0.009$; for PMI reconstruction of 2-back (1150 – 3150 ms from *n + 1* onset), $t(41) = -1.60$, $p = 0.117$. Figure and caption adapted from Wan et al. (2022).

Using RNNs to model WM prioritization

This thesis takes a different and novel approach to studying the neural substrates WM prioritization by resorting to artificial neural networks, which have been playing an increasingly prominent role in providing mechanistic insights into, and generating novel

hypotheses of, phenomena in cognition and neuroscience (Kell & McDermott, 2019; Mante et al., 2013; Richards et al., 2019; Sussillo et al., 2015; Yang et al., 2019). Since the 1970s, certain variants of recurrent neural networks (RNNs) have been used to model cognition (Hinton & Sejnowski, 1983; Hopfield, 1982; Wilson & Cowan, 1972). In the past decade or so, RNNs have been solidified as a useful tool to study brain and cognition in humans as well as other animals (Barak, 2017; Sussillo, 2014; Yang & Molano-Mazón, 2021). RNNs share many similarities with biological networks (Sussillo, 2014). First, RNNs have many internal units that perform nonlinear computations, which are analogous to neurons in the nervous system. Second, there are many feedback connections between the units that enable the generation of complex dynamics, and third, the units, which are simple themselves, work together in a parallel and distributed fashion, thus performing sophisticated computations. RNNs are trained to satisfy certain objective functions which provides a normative account and allow us to evaluate whether the algorithm that humans use reaches optimality. Because each RNN unit can receive connections from any other unit, it not only reflects the processing of the current input stimulus, but also the state of the entire network (Barak, 2017). In this sense, RNNs are especially suited to model computations that unfold over time, such as maintaining an item in working memory.

In the empirical work presented in Chapters 2 and 3, we trained RNNs (LSTM and vanilla RNNs) on the previously mentioned tasks (2-back and DSR) where the priority of the memoranda was manipulated. Using dimensionality reduction techniques (PCA and demixed PCA) to visualize the representational dynamics embedded in the recurrent units, we demonstrated that stimulus representations in RNNs also undergo representational transformations when transitioning between priority states (Chapters 2 and 3), reminiscent of the neuroimaging findings above. Specifically, the simulations in Chapter 3 highlighted the important role of representations of stimulus context and identified distinct transformational mechanisms of context and priority. Armed with these observations, we returned to extant fMRI and EEG datasets and found that priority and context are represented differently along the dorsal visual stream, and that behavioral performance is sensitive to trial-by-trial efficacy of priority coding, but not context coding. The work presented in this thesis significantly adds to our understanding of WM prioritization on a mechanistic/algorithmic level.

Chapter 2

Priority-based transformations of stimulus representation in visual working memory

Quan Wan, Jorge A. Menendez, Bradley R. Postle

Published in 2020: *PLOS Computational Biology 18*(6): e1009062.

Abstract

How does the brain prioritize among the contents of working memory (WM) to appropriately guide behavior? Previous work, employing inverted encoding modeling (IEM) of electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) datasets, has shown that unprioritized memory items (UMI) are actively represented in the brain, but in a "flipped", or opposite, format compared to prioritized memory items (PMI). To acquire independent evidence for such a priority-based representational transformation, and to explore underlying mechanisms, we trained recurrent neural networks (RNNs) with a long short-term memory (LSTM) architecture to perform a 2-back WM task. Visualization of LSTM hidden layer activity using Principal Component Analysis (PCA) confirmed that stimulus representations undergo a representational transformation – consistent with a flip – while transitioning from the functional status of UMI to PMI. Demixed (d)PCA of the same data identified two representational trajectories, one each within a UMI subspace and a PMI subspace, both undergoing a reversal of stimulus coding axes. dPCA of data from an EEG dataset also provided evidence for priority-based transformations of the representational code, albeit with some differences. This type of transformation could allow for retention of unprioritized information in WM while preventing it from interfering with concurrent

behavior. The results from this initial exploration suggest that the algorithmic details of how this transformation is carried out by RNNs, versus by the human brain, may differ.

Author Summary

How is information held in working memory (WM) but outside the current focus of attention? Motivated by previous neuroimaging studies, we trained recurrent neural networks (RNNs) to perform a 2-back WM task that entails shifts of an item's priority status. Dimensionality reduction of the resultant activity in the hidden layer of the RNNs allowed us to characterize how a stimulus item's representation follows a transformational trajectory through high-dimensional representational space as its priority status changes from memory probe to unprioritized to prioritized. This work illustrates the value of artificial neural networks for assessing and refining hypotheses about mechanisms for information processing in the brain.

Introduction

The ability to flexibly select and prioritize among information held in working memory (WM) is critical for guiding behavior and thought. To do this successfully, the cognitive system must solve a fundamental computational problem of how to maintain information in a readily accessible state while also preventing it from interfering with ongoing behavior. The primary goal of the work presented here is to investigate how this might be accomplished. If two items are currently held in WM, one possible solution could be to encode the "unprioritized memory item" (UMI) into a pattern of synaptic weights (Barak & Tsodyks, 2014; Stokes, 2015), an "activity-silent" trace (Myers et al., 2017) that might be less likely to interfere with the currently active "prioritized memory item" (PMI). Although some previous neuroimaging studies have reported that the prioritization of one item held in WM leads to a decrease-to-baseline of the activity level of the UMI (LaRocque et al., 2013; Lewis-Peacock et al., 2012; Rose et al., 2016), whether an "activity-silent" mechanism may contribute functionally to WM remains a topic of vigorous debate (Christophel et al., 2018; Schneegans & Bays, 2017b; Sprague et al., 2016; Stokes et al., 2020). In the present report, we evaluate an algorithmically different solution for prioritization: the *priority-based transformation* of the UMI into a representational format that, although active, is different from that of the PMI. Such a

transformation could minimize the likelihood that the UMI interferes with ongoing behavior.

Experimental tasks used to study prioritization in WM necessarily include multiple steps, such that information not needed for the impending response (i.e., the UMI) might nevertheless be needed to guide a subsequent response. This is often done with retrocues, and two recent studies using a retrocuing procedure have provided evidence consistent with priority-based transformation. In one, van Loon and colleagues (2018) acquired functional magnetic resonance imaging (fMRI) data while first presenting subjects two target images sequentially (e.g., first a flower then a cow), then indicating with a cue whether memory for the first or second presented image would be tested first. Had the cue been a "1", subjects would next see a test array of six flowers and indicate whether the target flower appeared in the test array, and finally a test array of six cows. On this trial, the target cow spent time as UMI, because the cue indicated that memory for the flower would be tested first. When van Loon et al. (2018) applied multivariate pattern analysis (MVPA) to fMRI data from posterior ventral temporal lobe, they found that a decoder trained on trials when an item was a PMI performed statistically below chance when that item was a UMI. Furthermore, a representational dissimilarity analysis indicated that, within their set of 12 stimuli (four cows, four skates, four dressers), each item's high-dimensional representation in one state (e.g., as

a PMI) was maximally different from its representation in the other state (i.e., as a UMI). Using a similar retrocuing procedure, Yu, Teng and Postle (2020) found, with multivariate inverted encoding modeling (IEM) of fMRI data from early visual cortex, that the reconstructed orientation of a grating "flipped" when it was a UMI relative to a PMI (e.g., a 30° orientation reconstructed as 120° while a UMI). Furthermore, for data from the intraparietal sulcus (IPS), they observed that the IEM reconstruction of the location where an item had been presented also flipped when an item's priority status transitioned to UMI.

Shifts of priority are also characteristic of continuous-performance tasks, for which shifts of priority are dictated by task rules rather than by explicit cues. One example, which features prominently in the work presented here, is the 2-back WM task from Wan and colleagues (2020; Figure 2.1). Electroencephalography (EEG) signals were recorded while subjects viewed the serial presentation of oriented gratings and judged for each one whether it was a match or a non-match to the item that had appeared two positions previously in the series. This task entails a predictable transition through priority states for each item: When an item *n* is initially presented, it serves as probe to compare against the memory of item *n – 2*; after the *n*-to-*n – 2* decision is made, item *n* becomes a UMI while item *n – 1* is prioritized for the upcoming comparison with *n + 1.* Next, once the *n + 1-to-n – 1* comparison is completed, item *n* becomes a PMI for its

impending comparison with item *n + 2.* To analyze the EEG data, an IEM was trained on the raw EEG voltages from a separate 1-item delayed-recognition task, and then tested on the delay periods separating *n* and *n + 1* and separating *n + 1* and *n + 2* (i.e., when item *n* assumed the status of UMI, then PMI). The results, reminiscent of van Loon et al. (2018) and Yu, Teng and Postle (2020), indicated that the IEM reconstruction of the UMI was "flipped" relative to the training data (Figure 2.2). The authors referred to this transition from PMI to UMI as "priority-based remapping" (rather than "recoding" or "code morphing"; c.f. Parthasarathy et al., 2017), reasoning that the IEM reconstruction of the UMI would fail if it were represented in a neural code different from the trained model.

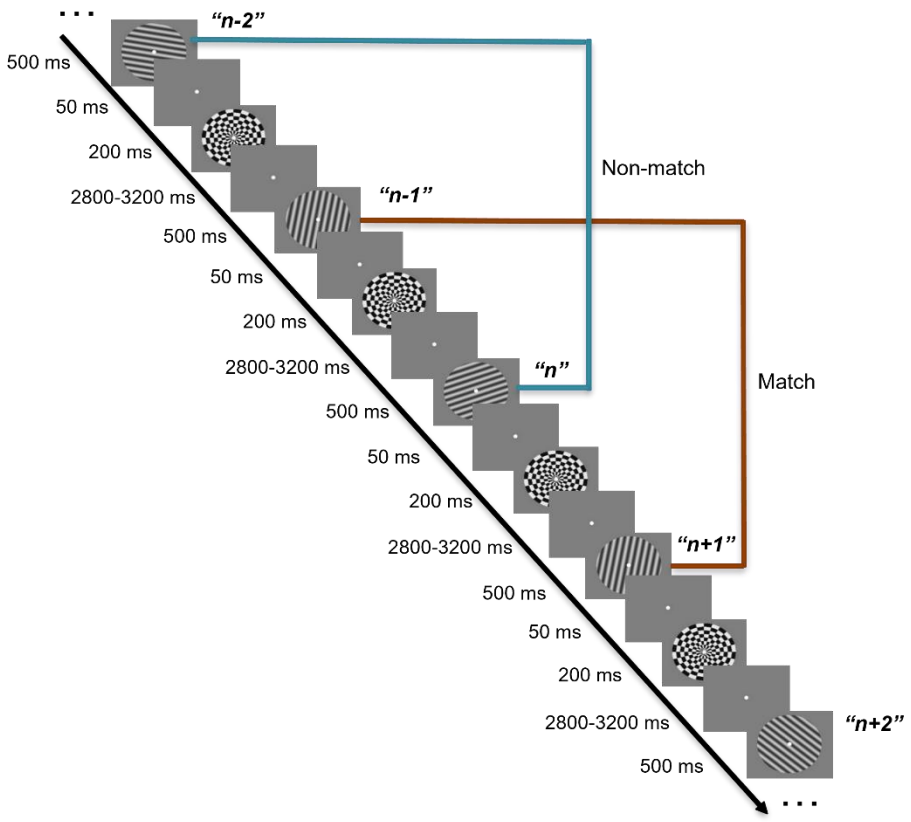


Figure 2.1. 2-back task structure in the Wan et al. (2020) EEG study. The presentation of each stimulus is followed by a 50 ms blank screen, a 200 ms radial checkerboard mask, a variable delay from 2.8 to 3.2 s, and then the next stimulus was presented, upon which the match vs. non-match response is to be made.

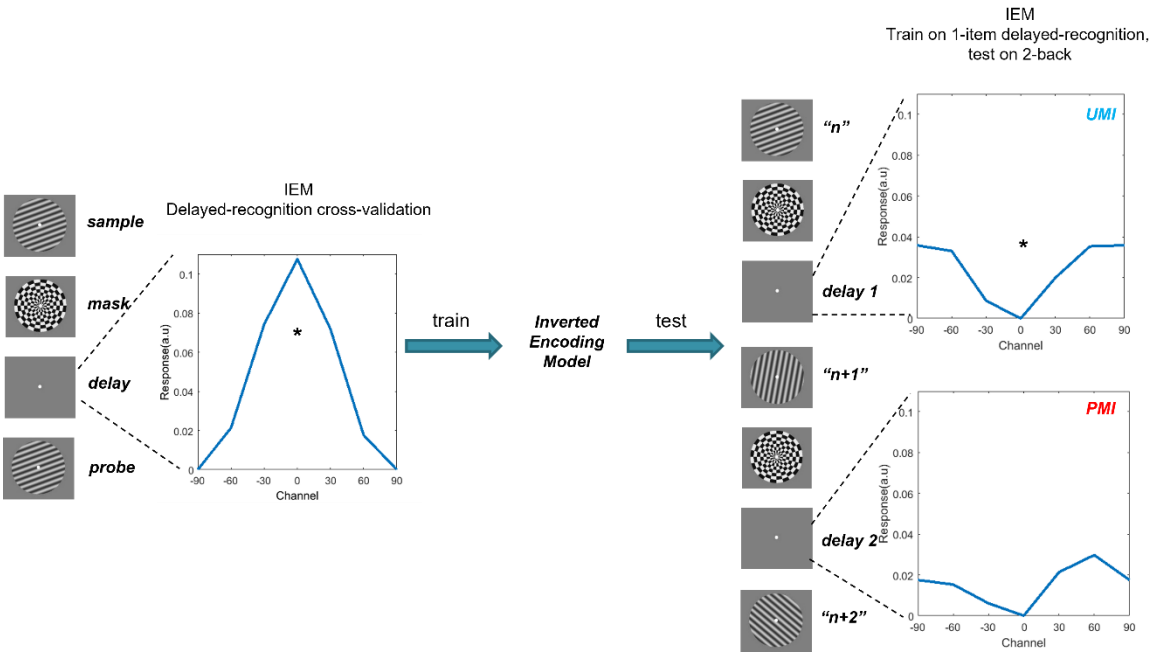


Figure 2.2. IEM reconstruction of EEG recorded while subjects performed the 2-back task (N = 42, combining data from the pilot study and the preregistered experiment from Wan et al., 2020). In IEM, voltage from each EEG electrode is construed as a weighted sum of responses from six orientation channels (modelled by a half-wave-rectified sinusoid raised to the 6th power), each tuned to a specific stimulus orientation, comprising the basis set. Left panel: IEM reconstruction of the stimulus during the delay in a separate one-item delayed-recognition task. This model was used to reconstruct the stimulus in the 2-back task. Right panel: Concatenation of the item *n* and item *n + 1* stimulus events to form a trial, across which *n* transitions from probe to UMI to PMI in the 2-back. On the right are IEM reconstructions corresponding to the two 2 s windows centered in two 2.8 s post-mask ISIs before and after item *n + 1*, respectively. "*" indicates $p < .01$ (two-tailed *t* test), FDR-corrected for multiple comparisons. As the figure shows, IEM reconstruction of stimulus *n* is "flipped" relative to the training data (IEM reconstruction from delayed recognition) when it is a UMI, demonstrating priority-based remapping. (Reconstruction of the PMI was unsuccessful.) For delayed-recognition IEM reconstruction (940 – 1040 ms from stimulus onset), $t(41) = 4.12$, $p < 0.001$. For UMI reconstruction of 2-back (-2400 – -400 ms relative to *n + 1* onset), $t(41) = -3.02$, $p = 0.009$; for PMI reconstruction of 2-back (1150 – 3150 ms from *n + 1* onset), $t(41) = -1.60$, $p = 0.117$.

Two recently published computational models offer some insight into the empirical phenomena that we have described up to this point. One model, by Lorenc and colleagues (2020), was designed to account for a similar flipped IEM reconstruction observed in an fMRI study using a retrocuing task. This approach was inspired by evidence from nonhuman primates (NHP) performing WM tasks, in which top-down signals from FEF were shown to alter several receptive field properties of neurons in extrastriate visual areas V4 and MT (Merrikhi et al., 2017). They created simulated data for training IEMs using the basis set that was employed for IEM reconstructions of empirical data, and subsequently created a test dataset where the basis function parameters for memory strength, gain, receptive field width, and receptive field centers were varied. When these parameters were fitted to experimental data, the best solution was a selective down-modulation of gain in feature-tuned sensory channels paired with a weakly excitatory top-down signal (i.e., memory strength). A second model, from Manohar and colleagues (2019), simulated WM performance in a network comprised of hard-coded feature-selective units and a pool of freely conjunctive units that can form a plastic attractor to keep one item, a PMI, in a state of elevated activity. When attention shifted away from an item (making it a UMI), it remained briefly encoded in a residual pattern of strengthened connections, and, under some conditions, inhibition from activity in other parts of the network produced an “inverted” representation of UMI.

Although this model successfully reproduced other empirical findings using simulated data, such as the temporary reactivation of the UMI by a nonspecific pulse of excitation, it was not used to account for empirical neural data.

It is instructive to consider the two models reviewed above from the perspective of the framework of Marr and Poggio (1976): They address distinct *computational* problems – prioritization within WM (Manohar et al., 2019) vs. removal from WM (Lorenc et al., 2020) – they propose different *algorithmic* solutions – inhibition via biased competition (Manohar et al., 2019) vs. excitation paired with selective gain modulation (Lorenc et al., 2020) – yet they observe similar patterns of neural *implementation* – flipping. Of particular relevance for our interests here is that although the details of their algorithmic operations differ, both models are constrained to finding only one class of solution: changing the strength of attention. Importantly, neither allows for the alternative that we will test here, which is the transformation of an item's representational geometry.

Previous WM research has implicated representational transformation as a solution to a third computational problem for WM: the retention of information in the face of distraction (e.g., Libby & Buschman, 2021; Parthasarathy et al., 2017). Our goal with the present work was to explore the possibility that the computational problem of prioritization in WM might also be solved algorithmically via representational

transformation. To accomplish this we turned to artificial neural networks (ANNs), which have been playing an increasingly prominent role in providing mechanistic insights into, and generating novel hypotheses of, phenomena in cognition and neuroscience (Kell & McDermott, 2019; Mante et al., 2013; Richards et al., 2019; Sussillo et al., 2015; Yang et al., 2019). In the current work, we use recurrent neural networks (RNNs) with an LSTM architecture (Hochreiter & Schmidhuber, 1997) to perform a 2-back WM task modeled on Wan et al. (2020). LSTMs can generate flexible behavior guided by long-range temporal dependencies, and can solve complex tasks such as speech recognition (Graves et al., 2013) and machine translation (Sutskever et al., 2014). Moreover, LSTM might be a good model for WM tasks due to its gating-based architecture, reminiscent of the cortico-striatal mechanisms believed to gate information into and out of WM (Chatham & Badre, 2015; O'Reilly & Frank, 2006). By comparing the stimulus representational schemes embedded in the EEG and RNN data, we hope to reveal whether humans and RNNs might employ similar algorithmic principles. Given that the RNNs are optimized to solve the 2-back task, we can also potentially use the RNN results to evaluate whether the algorithm that humans use reaches optimality.

Our approach was to train RNNs to perform the 2-back task, and then first use Principal Component Analysis (PCA) of the activity of the RNN's hidden layer to

visualize its representational dynamics. This revealed a smooth rotational transformation of stimulus representations over the course of the trial (Figure 2.4). This trajectory provided novel, independent evidence that the transition of the functional role of an item from memory probe to UMI to PMI is accompanied by transformations of its representational format. However, because PCA does not allow for the isolation and quantification of variation attributable to specific task dimensions (of particular interest here, priority status and the match/nonmatch decision), we carried out two additional sets of analyses. First, we applied demixed Principal Component Analysis (dPCA; Kobak et al., 2016) to the RNN data in order to identify distinct low-dimensional subspaces occupied by the neural representations of the UMI, the PMI, and the RNN's decision. We then quantified the temporal dynamics of these representations within the subspaces and the geometric relationships between the subspaces. Finally, we used this analysis of the RNN data to derive quantitative hypotheses with which to assess evidence that the EEG data from Wan et al. (2020) may also show evidence of priority-based transformation. The results of these hypothesis tests provide novel insights about priority-based transformations of stimulus information that are carried out by the human brain.

Methods

Behavioral task

In each experimental block of the 2-back WM task, both human subjects (N = 42) and RNNs (N = 20) were serially presented a sequence of stimuli drawn from a closed set of six different identities (128-stimulus blocks for humans, 20-stimulus blocks for RNNs). The task was to indicate, for each stimulus, whether or not it matched the identity of the stimulus that had been presented 2 positions earlier in the series. Each EEG subject performed 4 blocks and each RNN performed 200 blocks.

Recurrent neural network (RNN) model

RNN architecture

Twenty RNNs with an LSTM architecture were trained and simulated using the Python-based machine learning package PyTorch. Specifically, we used the default LSTM architecture in PyTorch with its default initializations. Initially, we trained 10 networks that consisted of 6 input neurons and 7 LSTM hidden units, which were linearly rectified and linearly read out to a single output neuron (Figure 2.3). We initially chose to use 7 units because this was the smallest number that solved the task with network solutions that were highly consistent across training instances (as evaluated by representational dynamics from the PCA visualization). Networks with other numbers of hidden units (up to 256) gave qualitatively similar results.

Subsequently, we repeated the procedure with RNNs with 60 LSTM hidden units to match the dimensionality of our EEG data, and with an input structure simulating the orientation stimuli in the human 2-back task of Wan et al. (2020).

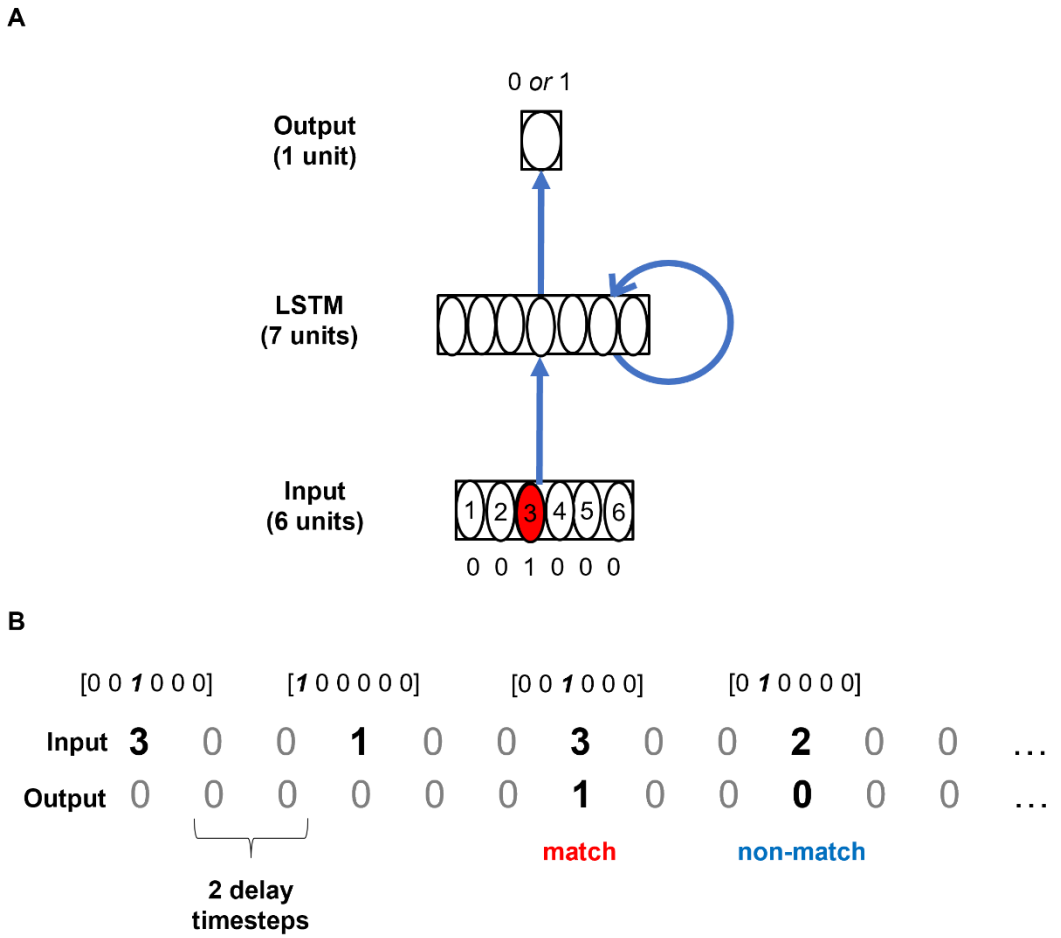


Figure 2.3. RNN model architecture. Shown is the architecture of the 7-hidden-unit RNNs. (A) One-hot vectors corresponding to each of the 6 stimulus types are fed into the input layer, which projects to an LSTM layer with 7 hidden units. This hidden layer in turn projects to an output unit with a binary target activation (0 = non-match, 1= match). (B) Example input and target output sequences. Two delay timesteps were installed after each stimulus presentation timestep to emulate the delay period in the 2-back EEG task. 60-hidden-unit RNNs have the same architecture except that they have 60 LSTM hidden units, and two input units that take a vector [cos 2θ, sin 2θ] (θ denoting the angle of grating orientations used in Wan et al., 2020) instead of a one-hot vector.

Stimuli

For 7-hidden-unit networks, the identity of each stimulus presented to the network was denoted by an integer randomly generated between 1 and 6. The stimulus input took the form of a one-hot vector, with only the unit corresponding to the stimulus identity activated (e.g., [0, 0, 1, 0, 0, 0] for stimulus #3; we also explored RNNs trained on metrically varying input vectors following the basis function used to build IEMs in Wan et al. (2020), and these yielded similar results, see Supplementary Materials S2.1). For 60-hidden-unit networks, to simulate the orientation stimuli, we instead employed 2 input units taking the vector [cos 2θ, sin 2θ], where θ denotes the orientation angle of each stimulus used in Wan et al. (2020; 10°, 40°, 70°, 100°, 130°, 160°). We multiply the angle θ by 2 to reflect the circular structure of the oriented grating stimuli, which have a period of 180° (i.e., a K° stimulus is identical to a 180° + K° stimulus). The multiplication by 2 ensures that this is also true for the corresponding inputs: cos (2 * (180° +K°)) = cos(2K°) and sin (2 * (180° + K°)) = sin (2K°) (see Supplementary Materials S2.2 for more details). To simulate the delay period in the human task, we installed 2 "delay" timesteps following the presentation of each stimulus (with an input of [0, 0, 0, 0, 0, 0]; no delay timesteps after the last stimulus in the sequence). A "stimulus event" consisted of the presentation of stimulus *n* and its following two delay timesteps. To evaluate the UMI-to-PMI representational transition of stimulus *n*, we refer to the concatenation of each two consecutive "stimulus events" as a "trial".

A scalar output was read out from the LSTM internal state by linearly rectifying the hidden state and then applying a linear layer. The network was then trained to output 1 (match) or 0 (non-match) during stimulus presentation depending on whether the presented stimulus (*n*) matched the stimulus presented two stimulus events back (*n - 2*). Each stimulus sequence comprised 18 "trials" (as defined above – note that because no delay period followed stimulus #20; the last "trial" contains stimuli #18 and #19), and only 16 trials were analyzed (because the first two stimulus events had no target outputs: not enough stimuli preceded them for there to be a match/non-match decision). We generated 200 random stimulus sequences for training the RNNs and 200 random sequences for testing the trained networks. Because the human 2-back task had a ratio of 1:2 between match and non-match trials, we generated random sequences that satisfied the criterion that each sequence had to contain at least 5 match trials. The outcome was that training sequences had an average of 5.55 match trials (*SD* = 0.78) and testing sequences an average of 5.46 match trials (*SD* = 0.70).

RNN training and testing

The internal state of the RNNs was initialized to 0, and weights and biases were initialized to random values, following the standard initialization of the PyTorch LSTM implementation. The 7-hidden-unit RNNs were trained using the Adam stochastic

gradient descent (SGD) algorithm for 5000 iterations (Kingma & Ba, 2014; learning rate = 10^{-3}). In each iteration, a batch of 20 sequences was randomly selected (with replacement) from the 200 training sequences. The loss function minimized was the mean squared error between output activity and target output across all timesteps.

We observed that after 5000 iterations of training, most RNNs had excellent performance on the training data. We therefore stopped training at this point and evaluated each RNN on an independently sampled set of 200 test stimulus sequences to assess generalization to arbitrary stimulus sequences. The network's performance accuracy was calculated as the percentage of trials (across all 200 sequences in the test set) on which the network made a correct response, where a response was deemed correct if the absolute difference between the activation of the output neuron and the target output was smaller than 0.5. We set a criterion level of performance accuracy of 99.5% for the networks. A total of 12 7-hidden-unit networks were trained, 2 of which were discarded due to below-criterion performance, leaving 10 RNNs for our analysis. All RNNs trained had the same architecture, hyperparameters and training/testing sequences. The only thing that differs across these 10 networks is the random initialization of the RNN weights prior to training. For subsequent analyses, the activity timeseries of the LSTM hidden layer units from all 3200 trials (16 trials x 200 sequences) in the training data set were used.

After analysis of the 10 successfully trained 7-hidden-unit networks, we repeated these training procedures and trained 10 RNNs with 60 units in the LSTM layer (batch size = 20, learning rate = 10^{-3}, 1500 iterations), so as to generate RNN data matching the dimensionality of our EEG data sets.

PCA visualization of the LSTM layer activity

We extracted from each network the activity of the 7 hidden units in the LSTM layer from all 200 training sequences and used Principal Component Analysis (PCA; implemented using Python's 'scikit-learn' library) to project these 7-dimensional activity patterns onto the top two dimensions accounting for the most variance across all training sequences and timesteps. We then visualized each stimulus *n*'s transition from probe to UMI to PMI within this subspace by plotting the dimensionality-reduced activity across the 9-timestep time course of a trial. These 9 timesteps comprised the presentations of stimulus *n, n + 1, n + 2* and the delay timesteps that followed each (i.e., *delay 1:1 and delay 1:2; delay 2:1 and delay 2:2; and delay 3:1 and delay 3:2*; Figure 2.4, "unlabeled" column). Note that, once a decision has been made about item *n + 2*, item *n* is no longer relevant for the task, so the *delay 3:1* and *delay 3:2* timesteps illustrate the evolution of the representational structure of *n* after it has presumably been "dropped from WM".

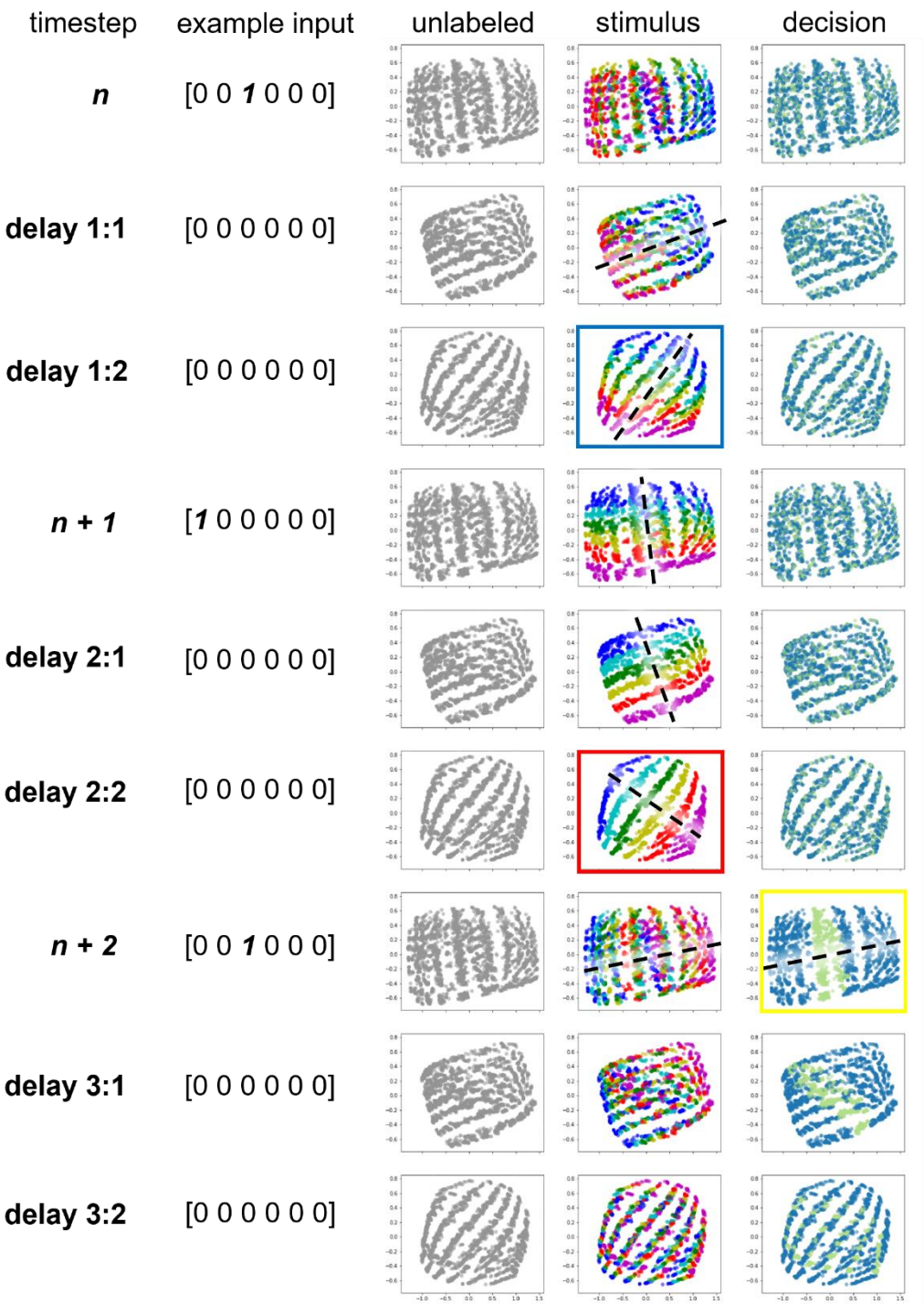


Figure 2.4. PCA visualization of LSTM hidden layer activity of an example 7-hidden-unit network (#7). Shown is a 9-timestep time course of the 2-back task, running from stimulus *n* to *delay 3:2*. Column 1 and 2: timestep labels and example input vectors. Column 3: Time course of dimensionality-reduced LSTM hidden layer activity. Each dot in the figures indicates the unit activity from a single trial. Column 4: Same as Column 3 but now each color corresponds to one of the six stimulus types, indicating stimulus *n*'s identity, and the black dashed lines illustrate the "schematic" stimulus coding axis. Blue and red squares highlight the two delay timesteps used to identify UMI and PMI dPCs, respectively. Column 5: Same as Column 4 except that the

colors now correspond to stimulus *n*'s status for the *n*-to-*n + 2* comparison that occurs at timestep *n + 2* (green: match trials, blue: nonmatch trials). Black dashed line at timestep *n + 2* (yellow square) illustrates the decision-coding axis. As can be seen in Column 5, the stimulus coding axis rotates counterclockwise (in the image plane) over time such that it becomes "perpendicular" to the decision axis at timestep *n + 1* and aligns with it at timestep *n + 2*. Percent variance explained: PC1 – 72.2%, PC2 – 15.7%.

To see how the representation of stimulus *n* evolves as it transitions from being a UMI to a PMI, we colored the activity patterns according to the identity of stimulus *n* (Figure 2.4, "stimulus" column). As explained in the Introduction, the memory of stimulus *n* is a UMI during the delay period after the presentation of stimulus *n* (i.e., during *delay 1:1 and delay 1:2*; because it is not needed for the upcoming *n – 1*-to-*n + 1* comparison), and then becomes a PMI during the delay period after the presentation of stimulus *n + 1* (i.e., during *delay 2:1 and delay 2:2*; in preparation for the imminent comparison with *n + 2*). We focused on the *delay 1:2* and *delay 2:2* timesteps (highlighted by blue and red squares) to characterize the UMI-to-PMI representational transformation.

To visualize the representation of the RNN's decision, we re-plotted the same activity patterns but colored them according to the correct response ("match" or "non-match") to the *n*-to-*n + 2* comparison when *n + 2* was presented (Figure 2.4, "decision" column; *n*-to-*n + 2* comparison timestep highlighted by yellow square). Note that, by construction, the RNN's actual response is the correct response in at least 99.5% of

trials, so this coloring can be effectively thought of as the RNN's true response in each of these trials.

WM-specific dimensionality reduction via dPCA

Demixed Principal Component Analysis (dPCA; Kobak et al., 2016) was employed to identify dimensions of RNN and EEG activity relevant to the stimulus representation in WM. Traditional PCA identifies dimensions that maximize the total variance of the recorded activity patterns across all task variables, such as time, stimulus, and decision. Demixed PCA, on the other hand, identifies dimensions of activity that contain variability specific to individual task variables. Given a task variable of interest (e.g., stimulus identity), the dPCA algorithm groups recorded activity patterns according to this variable and then extracts dimensions that maximize variance *across* groups (e.g., activity patterns evoked by different stimuli) while also minimizing variance *within* groups (e.g., activity patterns evoked by the same stimulus, but at different points in time and with different decisions). Here, we used this method to identify dimensions of activity that were strongly modulated by the identity of the UMI or PMI during the delay period.

To extract the demixed Principal Components (dPCs) of UMI-related variance, we minimize the following loss function:

$$V^{UMI}, W^{UMI} = \arg\min_{V,W} \sum_{s,t} ||(\bar{x}^s - \bar{x}) - VW^T(x_t^s - \bar{x})||^2$$

where x_t^s is the neural activity at time *t* averaged over all trials in which stimulus *s* (*s* being one of the 6 stimuli) was the UMI (trial averaging was necessary to average away noise), $\bar{x}^s = \frac{1}{T}\sum_{t=1}^{T} x_t^s$ is its mean over time, and $\bar{x}$ is the global mean over all trials and timepoints. This least squares optimization problem is called reduced-rank regression, and admits a closed-form solution (Kobak et al., 2016). This objective seeks to capture fluctuations in activity, $\bar{x}^s - \bar{x}$, arising from changes in the UMI stimulus and independent of time, as we expect the WM representation to stay stable over the late delay period. We refer to the columns of W^{UMI} as the *UMI dPCs*, and call the subspace spanned by the columns of V^{UMI} the *UMI subspace*. We similarly extracted *PMI dPCs*, W^{PMI}, and a *PMI subspace*, V^{PMI}, by exactly repeating the above operation but with the index *s* now indexing the PMI stimulus rather than the UMI stimulus.

In order to extract dimensions of activity specific to WM, we sought to restrict the above optimization to activity patterns during the late delay period. For the RNN's, this led us to utilize a single timepoint: the second timestep of the delay period (i.e., t only takes on a single index; cf. *delay 1:2* (for UMI)*, delay 2:2* (for PMI) in Figure 2.4). For the EEG data, we used timepoints from the second half of the delay: $t \in$ [-1400ms, 0ms] (for UMI) and $t \in$ [2150ms, 3550ms] (for PMI) relative to stimulus *n + 1* onset.

For the purposes of visualization we extracted only two dPCs (i.e. V and W each have two columns only), so as to obtain two-dimensional projections, z_t^s, of the neural activity. These projections were computed using the dPCs (W^{UMI} or W^{PMI}) as follows,

$$z_t^s = W^T(x_t^s - \bar{x})$$

It is these two-dimensional vectors that are plotted in Figure 2.5 using the simulated RNN data (x_t^s is the 7- or 60- dimensional internal LSTM state vector) and in Figure 2.6 using the EEG data (x_t^s is the 60-dimensional vector of signals recorded at each EEG channel).

For estimating the geometric relationships between stimulus and decision subspaces (Figure 2.7), we estimated decision dPCs, W^{dec}, and a decision subspace, V^{dec}, by capturing variability arising from changes in the subject's decision, $x_t^{s,d} - \bar{x}_t^s$, as follows,

$$V^{dec}, W^{dec} = \arg\min_{V,W} \sum_{s,t} ||\left(x_t^{s,d} - \bar{x}_t^s\right) - VW^T\left(x_t^{s,d} - \bar{x}\right)||^2$$

where $x_t^{s,d}$ is the is the neural activity at time t averaged over all trials in which stimulus s was the probe and response d ("match" or "non-match") was the decision made by the subject (or RNN), $\bar{x}_t^s = \frac{1}{2}(x_t^{s,\text{"match"}} + x_t^{s,\text{ "non-match"}})$ is its mean over the two decisions, and $\bar{x}$ is again the global mean over all trials and timepoints. We again used two dPCs, in accordance with previous analyses of WM subspaces (Panichello &

Buschman, 2021). In this case, we only considered timepoints during the decision time period: $t \in [200\text{ms}, 700\text{ms}]$ relative to stimulus onset for EEG and t = stimulus presentation timestep for RNN. See "UMI/PMI/decision subspace analysis" section below on how the relationships between the different subspaces (V^{UMI}, V^{PMI}, and V^{dec}) were then quantified.

Percent variance explained calculations were performed as follows. Percent global variance explained by the *i*th dPC, w_i (i.e. the *i*th row of the decoder matrix *W*), was calculated using the corresponding column v_i from the encoder matrix *V* by

$$1 - \frac{\sum_{s,t} \left\| (x_t^s - \bar{x}) - v_i w_i^T (x_t^s - \bar{x}) \right\|^2}{\sum_{s,t} \| (x_t^s - \bar{x}) \|^2}$$

The percent *stimulus* variance explained was defined as

$$1 - \frac{\sum_{s} \left\| (\bar{x}^s - \bar{x}) - v_i w_i^T (\bar{x}^s - \bar{x}) \right\|^2}{\sum_{s} \| (\bar{x}^s - \bar{x}) \|^2}$$

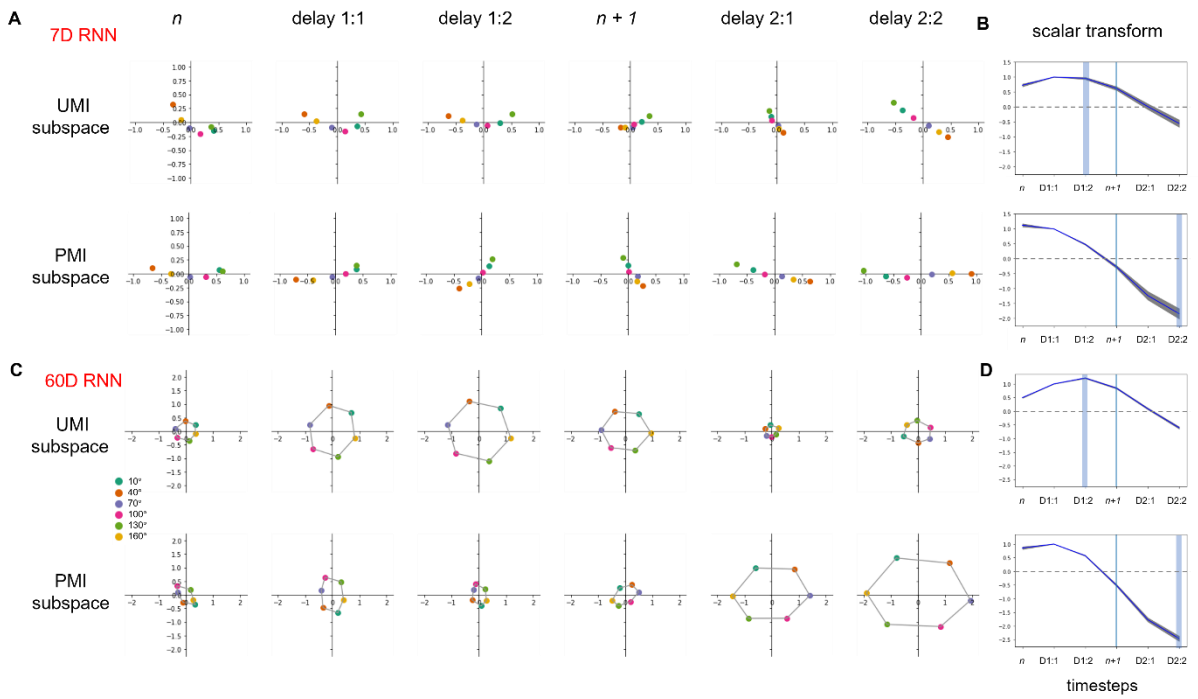


Figure 2.5. Stimulus trial-averages of RNN hidden layer activity projected into UMI and PMI subspaces over the course of a trial (stimulus *n* to *delay 2:2*). (A) Results from an example 7-hidden-unit network. Dot color indicates stimulus *n*'s identity. (B) Time course of scalar transform over the course of a trial, averaged across 10 networks. Blue vertical line indicates the timestep when stimulus *n + 1* is presented. Light blue shading shows the timesteps that were used to identify the dPCs. The gray shading around the curve indicates standard error of the mean. The gray dashed lines indicate a scalar transform of 0; the stimulus representational format is reversed after crossing this line. (C, D) Same as (A, B) but for the 60-hidden-unit RNNs. In C, data points of adjacent stimulus orientation angles are connected by a gray line.

Characterizing the dynamics of the UMI-to-PMI transformation

To characterize the continuous dynamics of the UMI-to-PMI transformation in stimulus-relevant dimensions, we quantified the evolving geometry of the stimulus representation visualized in Figure 2.5 and 2.6 by fitting a scalar transform, *k*, that minimizes the squared difference between the stimulus representation at a given timepoint and the transformed early-delay UMI representation,

$$\hat{k}_t = \arg\min_k \sum_s ||W^T(x_t^s - \bar{x}_t) - kW^T(\bar{x}^s - \bar{x})||^2$$

where $\bar{x}^s$ here is our estimate of the UMI, estimated by averaging activity patterns, x_t^s, over all timesteps *t* during the first half of the first delay (*delay 1:1* for RNN and -2800 to -1400 ms relative to stimulus *n + 1* onset for EEG). The index *s*, which refers to the stimulus presented prior to this delay, therefore corresponds to the identity of the UMI stimulus. In Figure 2.5B, 2.5D and 2.6B we plot these best-fitting scalars as a function of time over a whole trial, as the activity transitions from representing the stimulus as a UMI to representing it as a PMI. To isolate structure within the WM-relevant subspaces, we fit this transformation to low-dimensional projections through the UMI dPCs, W^{UMI}, or PMI dPCs, W^{PMI}.

Note that before computing these projections we center the activity vectors by subtracting their mean at the corresponding timepoint, $\bar{x}_t = \frac{1}{S}\sum_{s=1,..,S} x_t^s$. This is because we are specifically interested in how the representational format of the stimulus in WM changes over time, rather than changes in the absolute encoding of these stimuli. This analysis only sought to capture how the relationships between the different stimulus representations change over time.

UMI/PMI/decision subspace analysis

To quantify the relationship between UMI, PMI and decision subspaces calculated from equations above, we used a metric developed by Panichello and Buschman (2021). This metric measures the alignment between corresponding pairs of dPC encoding vectors as follows:

$$\text{UMI-PMI subspace alignment} = |v_1^{UMI} \cdot v_1^{PMI}||v_2^{UMI} \cdot v_2^{PMI}|$$

$$\text{UMI-decision subspace alignment} = |v_1^{UMI} \cdot v_1^{dec}||v_2^{UMI} \cdot v_2^{dec}|$$

$$\text{PMI-decision subspace alignment} = |v_1^{PMI} \cdot v_1^{dec}||v_2^{PMI} \cdot v_2^{dec}|$$

where the dot denotes the Euclidean dot product, and the bars denote absolute value. Here, v_1^{UMI}, v_2^{UMI} are the 1st and 2nd UMI dPC encoding vectors, i.e., the two columns of the matrix V^{UMI}. The analogous definition holds for the PMI and decision dPCs: v_1^{PMI}, v_2^{PMI} are the columns of V^{PMI}; v_1^{dec}, v_2^{dec} are the two columns of V^{dec}. Note that under the standard dPCA formulation used by Kobak et al. (2016) and used here, the encoding vectors are all norm 1. These dot products can therefore be interpreted as cosines of angles between the pairs of vectors, and the subspace alignment metric can be interpreted as a product of two cosines.

To turn this metric into an angle, we took the inverse cosine of each alignment metric in the three equations above. These are the angles plotted in Figure 2.7.

EEG dataset

The experimental protocol for the Wan et al. (2020) EEG study (the data from which was analyzed in this paper), along with the informed consent form, was approved by the University of Wisconsin–Madison Health Institutional Review Board (protocol no. 2016-0500). Prior to each experimental session, informed consent was obtained by lab personnel listed on the IRB-approved protocol.

60-channel EEG data were acquired and preprocessed as per procedures described in Wan et al. (2020). Raw EEG voltages were used for all analyses. Because data from the pilot and replication experiments from Wan et al. (2020) yielded very similar IEM reconstruction results, they were combined to yield a dataset of 42 subjects. As is the case with the RNN data, after excluding the first two stimuli from each block there were 126 stimulus events and hence 125 trials per block. Each stimulus event (stimulus presentation followed by a delay) lasted 3550 ms. A third of the trials in each block were 'match' trials and the other two thirds were 'non-match' trials. EEG data from all trials (both correct and incorrect) were included in the analyses. For each stimulus *n*, during the delay period after its onset, stimulus *n – 1* had the status of PMI and *n* had the status of UMI.

Results

Behavioral results of EEG study

Mean accuracy was 86.1% (*SD* = 5.6%), mean *d'* was 2.40 (*SD* = 0.65), and mean response time was 0.82 s (*SD* = 0.18 s).

Visualizing LSTM activity using PCA

PCA was carried out on the 7D LSTM hidden layer activity from the training data, and the resultant dimension-reduced activity from all 3200 trials projected onto the 2D-space constructed by the first 2 principal components (Figure 2.4, "unlabeled" column). This revealed that representations tended to cluster into band-like manifolds that appeared to rotate over the course of the trial (i.e., from timestep *n* to timestep *n + 2*). Next, to get a sense of the stimulus representational structure and how it evolves over time, we colored the data points for each trial according to the identity of stimulus *n* (Figure 2.4, "stimulus" column). This revealed that, across trials, stimulus representations were organized into stimulus-specific "stripes" that at some timesteps cut across the band-like manifolds (*delay 1:1* and *delay 1:2*), and at others were perfectly overlaid on them (*delay 2:1* and *delay 2:2*). These "stripes" thus defined a stimulus-coding axis. (That is, a stimulus's identity can be read out based on its location along this axis. A schematic illustration of this axis is superimposed on some of the timesteps from Figure 2.4, "stimulus" column, with a black dashed line.) It is noteworthy that, at timestep *n + 2*, the configuration of individual trials is different than at timestep

n. This reflects the fact that items serve different functions at these two timesteps – *probe* at timestep *n* and *memorandum* at timestep *n + 2.* Indeed, if one were to re-color timestep *n + 2* according to stimulus *n + 2*'s identity, this frame would be identical to the configuration of stimulus *n* at timestep *n*, which means that *n + 2* and *n* are in opposite locations in PCA space (e.g., in Figure 2.4, the azure-colored stimulus trials occupying the right side of PCA space at timestep *n* are on the left side of the space at timestep *n + 2*).

Finally, to get a sense of how the RNN's decision was represented, we colored each data point according to whether or not the correct response at the end of the trial was "match" (i.e., at timestep *n + 2*; Figure 2.4, "decision" column). This revealed that, when *n* is compared with *n + 2*, activity in trials requiring a "match" response converged onto the two central bands, whereas activity in non-match trials converged to the flanking bands. This organization thus defined a decision-coding axis, in that the correct response at a given trial can be read out based on the location of the RNN's internal state along this axis. A schematic illustration of this axis is superimposed on timestep *n + 2* from Figure 2.4, ("decision" column, with a black dashed line).

Over the course of a trial, *n*'s stimulus-specific axis appeared to rotate counterclockwise (in the PCA plane) as it transitioned from UMI (during *delay 1:1 and delay 1:2*) to PMI (during *delay 2:1 and delay 2:2*). This likely reflects, in part, transitions

between the functional roles of probe (timestep *n*), then UMI, then PMI. Thus, we can hypothesize the following functional account of the representational trajectory through a trial of, say, an azure-colored stimulus from Figure 2.4. At timestep *n*, its representational structure puts it on one of the central bands if it matches item *n – 2* (and therefore elicits an output of [1]), or on a band to the right of center if it does not match item *n – 2.* These two locations are separated along the decision-coding axis. Next, as it acquires the functional status of UMI, it transitions to a configuration that is not compatible with decision-making, as evidenced by the fact that every azure stimulus is located along a "stripe" that is parallel to the decision-coding axis at timestep *n + 1* (stated another way, the stimulus-coding axis at timestep *n + 1* is orthogonal to the decision-coding axis). During *delay 2:1* and *delay 2:2* the item's representation continues to rotate in the same counterclockwise direction on a trajectory that brings it back into alignment with the decision axis, but now on the "opposite side" of the PCA space, reflecting the fact that it is a PMI. (I.e., for azure items, probes cluster on the right side of PCA space, PMIs on the left side.) At timestep *n + 2*, the band occupied by this item will depend on its match/nonmatch status. From this we can further hypothesize that the function of this rotational trajectory might be to prevent the remembered representation of *n* from influencing the *n – 1* versus *n + 1* decision (at timestep *n + 1*).

Whatever the intuitive appeal of these hypotheses, PCA is not well suited to reveal the structure most relevant for representing a given task variable (e.g., UMI/PMI status, decision, …), because PCA is completely agnostic about which task variables the neural activity depends on. We therefore next sought to more directly visualize the structure of the UMI, PMI, and decision representations by incorporating these task-relevant variables into our dimensionality reduction method.

Visualizing LSTM representations using dPCA

Unlike PCA, which attempts to capture all variability across all time and all trials, dPCA seeks to capture variability dependent on specific task variables. By applying this dimensionality reduction method to neural activity during the delay period – during which the stimulus is held in memory – we can identify the dimensions most relevant to the representation of the stimulus in WM. By projecting the timeseries data into the subspaces spanned by these dimensions, we can visualize the temporal evolution of the geometry of the stimulus representation. This would allow us to test quantitatively the hypothesis that, for a given item *n*, its representational format while it is a UMI is transformed into a representational format that, although active, is different from that of the PMI.

RNN with 7 LSTM units

We applied dPCA to the 7D data from the RNNs to identify the top two UMI-selective dPCs (at the *delay 1:2* timestep) and the top two PMI-selective dPCs (at the *delay 2:2* timestep). The first two dPCs of the UMI subspace accounted for 97.4% of the total stimulus variance of the trial-averaged data. The first two dPCs of the PMI subspace accounted for 99.8% of the total stimulus variance (see Supplementary Materials S2.3 for additional information). Comparison of the trial-averaged population activity during the first delay period (*delay 1:2*) and second delay period (*delay 2:2*) reveals that, the way in which the stimulus is represented changes over time as it transitions from an unprioritized (UMI, in the first delay period) to a prioritized state (PMI, in the second delay period). The stimulus can be read out at both of these timepoints, but the relationship between stimulus and population activity is reversed (e.g., in Figure 2.5A, top row, the ordering along the 1st dPC at *delay 1:2* is *orange-yellow-purple-pink-teal-green*, whereas at *delay 2:2* it is *green-teal-pink-purple-yellow-orange*). This is true regardless of whether we project this representation through the UMI dPCs (Figure 2.5A, top row) or the PMI dPCs (Figure 2.5A, bottom row). Iteratively projecting trial-averaged activity from each timestep onto these two dPC subspaces suggested that the evolution of stimulus representational format across the trial is such that its projection onto the 1st dPC of the PMI – the axis that is critical for readout of the

memory item against which the impending probe is to be compared at timestep *n + 2* -- is minimal at timestep *n + 1.* (Note that this corresponds to the 0-crossing of the scalar transform, as described in the next paragraph.)

To quantify these dynamic changes in the stimulus representation across the various stages in the trial, we fit a scalar transformation from the trial-averages at timestep *delay 1:1* to the trial-averages at every other timestep (see Methods). The value of this best-fitting scalar transform for each timestep is plotted in Figure 2.5B. Relative to the UMI subspace (i.e., dPCA on timestep *delay 1:2*), an item's representational format was relatively stable (i.e., unchanging) for the first half of the trial, with the scalar transform close to 1.0, then, after timestep *n + 1*, shifted to a steady rate of transformation for the remainder of the trial, with the 0-crossing of the scalar transform (indicating the reversal of the stimulus-activity mapping) occurring at timestep *delay 2:1.* Relative to the PMI subspace, the representational format began contracting during *delay 1*, flipped just before timestep *n + 1*, and steadily expanding through *delay 2*. Together, these results confirm that an item's representational transformation across the trial proceeds at a relatively steady rate (consistent with the smooth rotation observed with the PCA (Figure 2.4)).

RNN with 60 LSTM units and a circular stimulus set

Although the results from the 7D RNN data produced quantitative predictions about the priority-based transformation of information held in WM, their direct applicability to the EEG data from Wan et al. (2020) would be complicated by two factors. First, there would be a difference in dimensionality between the two datasets (7D for the RNN, 60D [corresponding to 60 channels] for the EEG). Second, whereas the six stimuli used to train the RNNs with 7 LSTM units were unrelated to each other, the six stimuli used in Wan et al. (2020) were orientations equally spaced within the circular range of 180°. Therefore, our next step was to repeat the procedure described up to this point, but with 10 RNNs with 60 LSTM units each, trained on six stimuli drawn from a circular space. Results with the resultant 60D data would constitute the hypotheses that we would then test with the EEG data from Wan et al. (2020).

To incorporate circular stimuli into our RNN model, we used 2D inputs taking the value [cos 2θ, sin 2θ], matching the periodicity of the oriented grating stimuli used in the task (where a 0° stimulus is equivalent to a 180° stimulus). We then constructed the six stimulus inputs by simply plugging in the six stimulus angles used in the EEG experiment. With these modifications, we trained 10 LSTMs with 60 hidden units to perform the 2-back task at > 99.5% correct. We then applied dPCA to the resultant 60-dimensional data from these RNNs. In this case, we found the UMI and PMI representations to have circular structure (Figure 2.5C), spreading across both dPCs

rather than just one as we saw in the previous simulations. More concretely, the first two dPCs of the UMI subspace accounted for 98.7% of the total stimulus variance of the trial-averaged data. The first two dPCs of the PMI subspace accounted for 99.0% of the total stimulus variance (see Supplementary Materials S2.3 for additional information).

Based on our analysis of these RNN's, we derived two predictions that we next sought to test in the EEG data from Wan et al. (2020):

• *UMI-to-PMI representational reversal:* as in the 7D RNN, the stimulus representation reverses as it transitions from being unprioritized (*delay 1:2*) to prioritized (*delay 2:2*). That is, when the stimulus is a PMI (*delay 2:2*), the colored points in Figure 2.5C are flipped with respect to the *x*- and *y*-axes compared to when the stimulus is a UMI (*delay 1:2*). This can be quantified by characterizing the stimulus representation at each timestep as some scalar transformation of the representation early in the first delay period (*delay 1:1*). This scalar transform remains positive and near 1.0 during the entirety of the first delay period and then gradually decreases in value and reversing sign during the second delay period (Figure 2.5D), illustrating a reversal in the representation of the stimulus as it transitions from being unprioritized to prioritized. This holds true both within the UMI and PMI subspaces.

• *Differential alignment of UMI and PMI subspaces with the decision subspace*: the role of the UMI representation in the 2-back task is to hold information about stimulus *n* in

memory that is irrelevant for the impending decision (i.e., when the subject has to make a judgment about stimuli *n + 1* and *n - 1*). An important property of this representation, then, is that it should not interfere with that decision. Conversely, the role of the PMI representation is to provide information necessary for the impending decision – it should therefore be able to contribute to that decision. We might thus expect, then, that the activity dimensions that are used to compute the decision should overlap substantially less with the UMI subspace than the PMI subspace. To assess whether this was the case in the trained LSTMs, we used dPCA to extract a decision subspace and then used the metric of Panichello & Buschman (2021) to measure the alignment of the UMI and PMI subspaces with this decision subspace. As expected, we find that, whereas the UMI subspace is largely orthogonal to the decision subspace (81.26° ± 2.06° SD), the PMI subspace has substantial overlap with it (35.25° ± 13.25° SD; Figure 2.7). The UMI subspace was also largely orthogonal to the PMI subspace (84.13° ± 3.91° SD).

Visualizing EEG activity using dPCA

The EEG data from Wan et al. (2020) were markedly noisier than the RNN data: The first two dPCs of the UMI subspace accounted for 69.1% of the total stimulus variance of the trial-averaged data; and the first two dPCs of the PMI subspace accounted for 69.4% of the total stimulus variance of the trial-averaged data.

Regarding the experimental predictions derived from the RNNs in the previous section, we evaluated whether they held in these data:

• *UMI-to-PMI representational reversal:* inspection of the data from a single subject (Figure 2.6A) shows no signatures of a representational reversal within the first two UMI or PMI dPCs. To confirm this across all subjects, we fit a scalar transformation from the representation in the first half of the first delay period to the representation at every other timestep. The average scalar transformation for each timestep is plotted in Figure 2.6B. Relative to the UMI subspace, the trajectory of the best-fitting scalar transformation qualitatively matched that from the 60D RNN, increasing across the delay preceding item *n + 1*, then (after holding a constant value across the time interval used to define the subspace) decreasing for the remainder of the trial. Unlike the RNN data, however, the scalar transform never reversed sign (Figure 2.6B, UMI row). Relative to the PMI subspace, the trajectory for the EEG data started with a steady increase across the delay preceding item *n + 1*, reaching its maximum value while item *n + 1* was on the screen (i.e., 2 sec. prior to the beginning of the time interval used to define the PMI subspace), then remaining unchanged for the remainder of the trial (Figure 2.6B, PMI row). This indicates that stimulus representations begin transforming toward their configuration in the PMI subspace, fully achieve it by epoch *n + 1* (at which time they have UMI status), and then maintain this end-state configuration for the

remainder of the trial. This trajectory differs markedly from the 60D RNN, for which the configuration relative to the PMI subspace was unchanging until after delay *1:2*, then rapidly changing across the second half of the trial. Also different from the RNN, the scalar transform for EEG did not reverse sign. These results indicate that representational reversals are not systematically present in the EEG data as they were in the RNN data. Anecdotally, inspection of the EEG data gave the impression of considerably more heterogeneity of representational geometry across subjects (in these first two UMI/PMI dPCs) than we saw across independently trained RNNs.

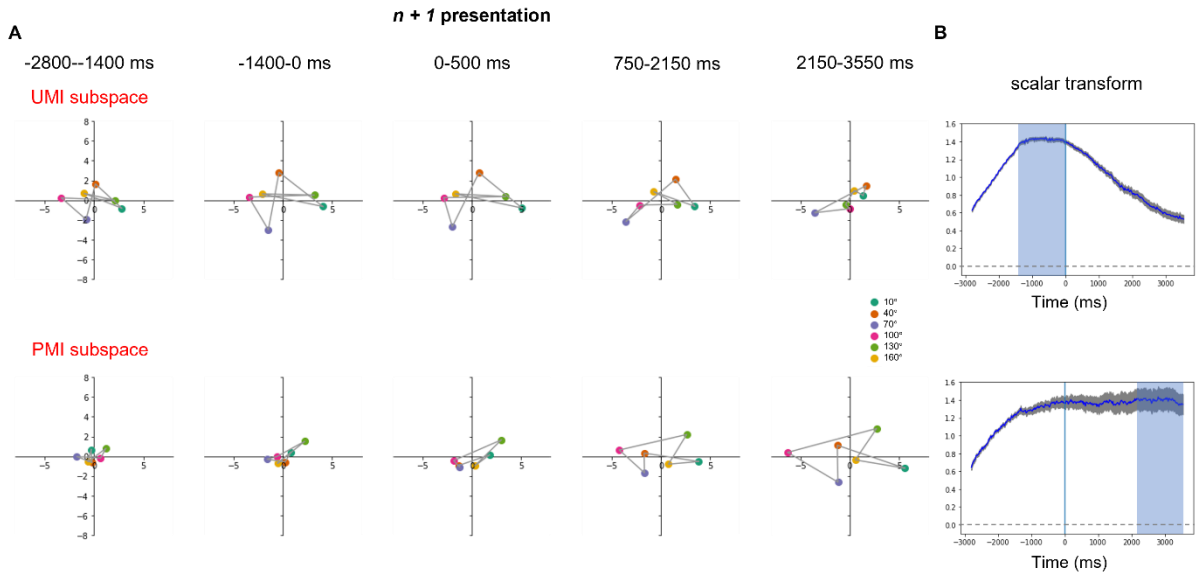


Figure 2.6. Stimulus trial-averages of EEG signal projected into UMI and PMI subspaces over a 2-delay time course (-2800ms to 3550ms relative to stimulus *n + 1* onset). (A) Results from an example subject. Dot color indicates stimulus *n*'s orientation angle. Data points of adjacent stimulus orientation angles are connected by a gray line. (B) Group-average time course of scalar transform over the course of a trial (N = 42). Blue vertical line indicates the onset of stimulus *n + 1*. Light blue shading shows the time windows that were used to identify the dPCs. The gray shading around the curve shows standard error of the mean.

• *Differential alignment of UMI and PMI subspaces with the decision subspace*: we found that, in the EEG data, the UMI and PMI subspaces were both largely orthogonal with the decision subspace (81.80° ± 6.69° SD and 82.74° ± 8.74° SD, respectively; Figure 2.7). In other words, we did not observe that the UMI representational subspace had a different geometric relationship to the decision subspace than the PMI subspace. The UMI and PMI subspaces were separated by an angle of 76.87° (SD = 12.33°).

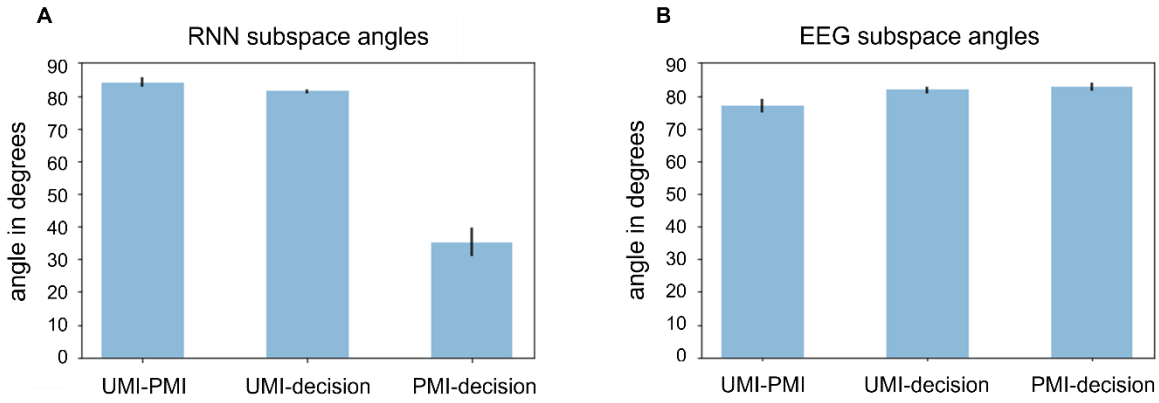


Figure 2.7. Angles between UMI, PMI and decision subspaces for (A) 60D RNN and (B) EEG data. Black bars indicate standard error of the mean.

Tracking the disappearance of information from WM using dPCA

As an exploratory endeavor, we have developed an approach using dPCA to track, at the level of an individual neural population (artificial or not), the disappearance of stimulus information when it is no longer relevant for a WM task. This was motivated by the observation that, for the three timesteps that follow *n + 2*, stimulus averages projected into the PMI subspace appeared to collapse (Supplementary Materials S2.4) – what one would expect as information "fades out" of WM at the end of a trial. To

quantify this intuition, we start with the three timesteps preceding the onset of stimulus *n* in the time course of stimulus averages projected into the PMI subspace (see Supplementary Materials S2.4 for an example RNN), reasoning that these will not have any information about stimulus *n*. The dispersion of stimulus averages for these timepoints can serve as an empirically derived baseline of discriminability when a stimulus is not in WM. Then, the timesteps immediately following stimulus *n + 2* (when *n* is no longer relevant) can be compared against the pre-trial baseline. To apply this procedure, we would calculate how dispersed the PMI stimuli are from each other at any given timestep and use the resultant metric as a proxy for stimulus discriminability.

We use a bootstrapping procedure in which, for each individual network, we resample with replacement the number of trials from each stimulus condition and perform dPCA on the resampled data as done in the section "*WM-specific dimensionality reduction via dPCA*" (PMI subspace). We then compute the dispersion (i.e., the variance) of the six stimulus trial-averages projected into the PMI subspace, for each timepoint, and repeat the procedure 10,000 times to construct the baseline (null) distribution of dispersion values. To be conservative in rejecting a timepoint as having WM information, we choose, in each iteration, the maximum dispersion value over the values calculated for the 3 timepoints preceding stimulus *n* to construct this distribution. We can then compare any timestep of interest with this distribution to

determine whether the dispersion value at the timestep lies beyond the 95th percentile (one-tailed test) of the baseline distribution: if so, we reject the null hypothesis and conclude that this timestep does contain discriminable information about item *n*; if not, we fail to reject the null hypothesis and conclude that we are unable to detect discriminable information about item *n* in its PMI dPC subspace. When applied to the network illustrated in Figure 2.5C, for example, information about item *n* persisted for one timestep after it was no longer relevant, but then was no longer detectable (Figure S2.4). Unfortunately, this approach cannot be applied to the data from Wan et al., 2020, due to their varying lengths of delay. We leave the application of this method to human neural signals for future work.

Discussion

Results from previous neuroimaging studies have given rise to the idea that representations in working memory (WM) undergo a "priority-based remapping" when they obtain the status of UMI (van Loon et al., 2018; Wan et al., 2020; Yu, Teng & Postle, 2020), but the mechanism underlying this transformation was unknown. Here, using neural network modeling and dimensionality reduction techniques, we have identified a transition through representational space that may reflect a general solution to the computational problem of needing to hold information in an accessible state (i.e., "in

WM") but in a manner that won't influence ongoing behavior. However, noteworthy differences between the transformational dynamics observed with RNNs versus with human EEG suggest important differences in implementational specifics, highlighting important questions for future work.

The 2-back task requires information to transition through three distinct functional states: that of a probe requiring comparison with the mnemonic representation of item *n – 2* and an overt match/nonmatch report; unprioritized (a state that should minimize interference with the concurrent *n – 1* vs. *n + 1* comparison and report); and prioritized (in preparation for its comparison against *n + 2*). PCA of hidden-layer activity of RNNs underwent a smooth rotation through 180° of the 2D space defined by the first two PCs. dPCA of RNNs characterized distinct subspaces corresponding to these states, and the trajectories between the states.

The organization of these functional subspaces is reminiscent of recent findings from NHPs performing a retrocuing WM task. Subjects first encoded two stimuli – one above fixation and one below – into WM, then viewed a cue indicating which one to report. Prior to the cue, PCA indicated that the *above* and *below* items were represented in subspaces of neural activity separated by a median angle of 79.1°. After the cue, the selected item transitioned into a different subspace, and the selected-from-*above* and selected-from-*below* subspaces were closely aligned -- separated by only 20.1°. The

authors interpreted this as a transition of the selected item from a representational format that emphasized the distinction between the two items to a "template" format that abstracted over location (no longer a relevant parameter) and facilitated behavioral read-out (specifically, recall; Panichello & Buschman, 2021). In our 2-back task, the UMI-to-PMI transition can be understood as the implicit selection of the UMI that occurs after a response is made to item *n + 1*. An important difference between our 2-back task and the retrocuing task of Panichello and Buschman (2021), however, is that their task lacked a UMI state. Rather, after the retrocue, there was no possibility that the uncued item would be needed. Nonetheless, in the PFC, a representation of the uncued item persisted, and its uncued subspace was orthogonal to the template subspace. Therefore, one important question for future work is whether, and if so how, the transition to UMI differs from the transition to no-longer-needed (i.e., "dropping" an item from WM).

It is important to note that the RNN modeling that we carried out here is not intended to simulate EEG data, nor the human brain, which has vastly different structural and functional architecture from our RNNs. For example, because of the relative simplicity of the RNN architecture, and the absence of many sources of noise that are characteristic of EEG (e.g., physiological noise, uncontrolled mental activity, measurement noise), the variability and SNR of the two signals differ markedly. This

limits what can be interpreted from direct comparisons between the two sets of results. Here we summarize where the two approaches have yielded similar versus dissimilar outcomes, and briefly consider some implications.

Comparison of RNN vs. EEG results

Similarities

• Stimulus representation in both RNN and EEG data went through a priority-based transformation, occupying, in turn, two distinct subspaces (UMI and PMI). This indicates that the UMI was actively represented in both RNN and EEG (c.f., LaRocque et al., 2013; Lewis-Peacock et al., 2012; Rose et al., 2016). Importantly, it confirms that, at the algorithmic level, prioritization in WM is carried out, at least in part, by an operation of representational transformation.

• The representational trajectories of RNN and EEG data are indicative of an active transformation, and so cannot be accounted for by inhibition (c.f., Manohar et al., 2019).

• The angles between UMI and PMI subspaces (RNN: 84°, EEG: 77°), and between UMI and decision subspaces (RNN: 81°, EEG: 82°) were similar for RNN and EEG. These patterns are consistent with a process that might minimize the influence of the UMI on other concurrent operations, including the retention of the PMI and the processing of the probe.

Differences

• Unlike the RNN data, the EEG data did not show evidence of a sign reversal of the best-fitting scalar transform. It remains to be determined if this reflects a fundamental difference in how the human brain carries out priority-based remapping, or if it may reflect a limitation of extracranial EEG. (E.g., the dynamics of priority-based transformations are different in different brain areas of the NHP (Panichello & Buschman, 2021), but comparable inter-regional differences would be mixed in our whole-scalp EEG data.)

• In the EEG data, the group-average stimulus representation transformed into its final configuration in the PMI subspace earlier than in the RNN data (Figures 2.5 and 2.6). (Indeed, human subjects, on average, recoded item *n* into its UMI and PMI configurations simultaneously, and then later prepared for item *n + 2* by collapsing the UMI structure during *delay 2*, whereas the RNNs recoded the item around the time when the priority status of the stimulus changed.) It is also noteworthy that the rate of this transformation was highly variable across individual EEG datasets, but not across RNNs. An important question for future research is whether individual differences in this factor may relate to behavioral performance, as well as whether it is sensitive to such factors as strategy or reward contingency.

• For the EEG data, the angle between PMI and decision subspaces was 83°, whereas for the RNN it was 35°. This pattern in the RNN data is consistent with close alignment of these two subspaces that might facilitate comparison of the PMI and the probe. Similar to the point raised previously, future work is needed to determine whether this difference reflects an important difference in decision-making between human and RNNs, or if it is a consequence of poor spatial resolution of the EEG data. (E.g., the effects of selection on WM information are markedly stronger in the PFC than in the visual cortex of NHPs (Panichello & Buschman, 2021).)

Contributions and limitations of the current work

One important role for the RNN simulations presented here has been to establish the validity and interpretability of our approach with dPCA. This, in turn, allowed us to use dPCA to evaluate neural coding in an EEG data set, including during task epochs for which multivariate methods had failed to find evidence for an active representation of the PMI (Figure 2.2). This successful application of dPCA to an extant EEG dataset in this study suggests that this approach may also provide novel insights if applied to the data from studies that have previously been interpreted as evidence for activity-silent storage mechanisms (LaRocque et al., 2013; Lewis-Peacock et al., 2012; Rose et al., 2016). The fact that dPCA does not make assumptions about the representational

structure of stimuli means that it's possible that it could find evidence for stimulus representation where a model-based approach, such as IEM, has failed. (Indeed, this is what happened with the PMI from the EEG data set in this study – compare Figure 2.2 with Figure 2.6.)

It is also important to note that the RNNs we simulated have a simple architecture, with a homogeneous LSTM layer, which is, of course, very different from the brain with its heterogeneous patterns of connectivity between neurons with varied structural and functional properties. The RNN simulations of Masse et al. (2019), employing different cell types and explicitly simulating factors like receptor time constants and presynaptic depletion of neurotransmitter, offer one promising example for developing more biologically plausible models. Also missing from our RNN architecture is an explicit source of control, such as that exerted by prefrontal and posterior parietal circuits in the mammalian brain. Through extensive training, our RNNs gradually learned to adjust their connection weights so as to achieve a high level of performance, but this was only possible because each item presented to the network always followed the same functional trajectory (probe, then UMI, then PMI). A hallmark of WM in the real world is the ability to flexibly respond to unpredictable changes in environmental exigencies. Thus, an important future goal will be to extend the present work to a network with separate modules with different connectivity patterns and governed by different

learning rules (e.g., Kruijne et al., 2020; O'Reilly & Frank, 2006), and to a task that requires truly flexible behavior.

Our work complements extant models of attentional prioritization in WM. First, it sheds light on the prioritization mechanisms of a continuous-performance WM task (2-back), a design that has recently received less attention than tasks employing retrocuing. Second, compared with the aforementioned computational accounts (Lorenc et al., 2020; Manohar et al., 2019), our use of dPCA provides a data-driven dimensionality reduction approach that does not make assumptions about the representational structure of stimuli. This allows one to examine the unmodeled structure of stimuli in the representational space. Third, our dPCA analyses were applied on a subject-by-subject basis, without assuming that the same representational and/or computational scheme is employed across individuals. Indeed, recent research has shown that representational biases of stimulus features vary among individuals in higher-order brain areas (Gong & Liu, 2020). Therefore, this approach may be helpful for explaining individual differences across many types of cognition.

To conclude, we used ANN simulations to validate the idea, at the level of representational codes, that shifts of priority status trigger the transformation of stimulus representations in WM. Applying dimensionality reduction to LSTM hidden layer activity in RNNs revealed the organization of functionally specific subspaces, and

the trajectories between different functional states. This approach translated to EEG data from subjects performing the same task, revealing similarities and differences between human and machine, and highlighting fruitful directions for future research.

Chapter 3

Representing context and priority in working memory

Quan Wan, Adel Ardalan, Jacqueline M. Fulvio, Bradley R. Postle

Submitted to *Journal of Cognitive Neuroscience*.

Abstract

The ability to prioritize among contents in working memory (WM) is critical for successful control of thought and behavior. Recent work has demonstrated that prioritization in WM can be implemented by representing different states of priority in different representational formats. Here, we explored the mechanisms underlying WM prioritization by simulating the double serial retrocuing (DSR) task with recurrent neural networks (RNNs). Visualization of stimulus representational dynamics using principal component analysis (PCA) revealed that the network represented trial context (order of presentation) and priority via different mechanisms. Ordinal context, a stable property lasting the duration of the trial, was accomplished by *segregating* representations into orthogonal subspaces. Priority, which changed multiple times during a trial, was accomplished by *separating* representations into different manifolds within each subspace. We assessed the generality of these mechanisms by applying dimensionality reduction and multiclass decoding to fMRI and EEG datasets and found that priority and context are represented differently along the dorsal visual stream, and that behavioral performance is sensitive to trial-by-trial efficacy of priority coding, but not context coding.

One of the hallmarks of working memory (WM) is its ability to flexibly prioritize among its contents in the service of the current behavioral goal. For example, say that you've just completed a talk at a conference, and you see two people simultaneously approaching each of two microphones to ask a question. You turn to the moderator and wait for them to indicate who will ask the first question, and based on this your shift of gaze is guided by your memory of the location of the cued microphone. To study prioritization in WM, one line of work has made extensive use of the double serial retrocuing (DSR) task, in which two sample items are initially presented and "remembered", followed by a blank "no-action" delay, then a retrocue indicating which of the two memorized items will be tested by an impending memory probe (see Figure 3.1A for an example). This item is said to take on the status of prioritized memory item (PMI). Because the item that was not cued may be tested later in the trial, however, it cannot be dropped from memory (i.e., "forgotten"), so it takes on the status of unprioritized memory item (UMI) until the PMI is tested. Subsequently, a second retrocue indicates, unpredictably, which item will be tested by a second memory probe; thus, either item can take on the status of PMI during the second half of the trial. An initial set of studies applying multivariate pattern analysis (MVPA) decoding to fMRI and EEG data from subjects performing the DSR task failed to find evidence for an active representation of the UMI, giving rise to the idea that it might be held in an "activity-

silent" state (Larocque et al., 2014; LaRocque et al., 2017; Lewis-Peacock et al., 2011b; Rose et al., 2016). More recently, however, studies using variants of the DSR task (with fMRI; van Loon et al., 2018; Yu, Teng & Postle, 2020) and the 2-back WM task (with EEG; Wan et al., 2020) have provided evidence for an active trace of the UMI that undergoes a transformation relative to the representational format of the PMI. Specifically, the UMI can produce significantly below-baseline MVPA decoding (van Loon et al., 2018) and "opposite" reconstruction with multivariate inverted encoding modeling (IEM; Wan et al., 2020; Yu, Teng & Postle, 2020).

As an initial step toward better understanding the priority-based representational transformations observed in neuroimaging data (van Loon et al., 2018; Wan et al., 2020; Yu, Teng & Postle, 2020), we had trained recurrent neural networks (RNNs) with a long short-term memory (LSTM) architecture to perform the 2-back WM task (Wan et al., 2022). Visualization of LSTM hidden layer activity using principal component analysis (PCA) had confirmed that stimulus representations in RNNs also undergo representational transformations when transitioning between priority states. Specifically, demixed (d)PCA of these data had identified two representational trajectories, one within a UMI-specific subspace and the other a PMI-specific subspace, both undergoing a reversal of stimulus coding axes. Having thus observed similar priority-based transformational dynamics in the human brain and in RNNs, we

speculated that this type of transformation might be a computationally rational way to meet the competing demands of retaining information in WM while simultaneously preventing it from interfering with concurrent behavior (Wan et al., 2022).

Whereas in Wan et al. (2022) we simulated the 2-back task, the results presented here were prompted by results of RNN simulation of the DSR. This was important to do because although the n-back task has been important for the study of many aspects of WM, it is poorly suited for the study of the flexible control of behavior with WM. This is because the n-back is a continuous performance task in which each item follows the same functional trajectory. For the 2-back, for example, each item *n* first serves as a memory probe against which to compare one's memory for item *n - 2*, then transitions to UMI (while *n + 1* is compared with the memory of *n – 1*), then transitions to PMI (for its comparison with item *n + 2*), then becomes no-longer-relevant and can be dropped from WM. The DSR, in contrast, does require online, flexible control, because the identity of the two retrocues can't be predicted prior to their onset. Unexpectedly, however, it is a different property of the DSR task that motivated the work presented here.

At the beginning of each trial of DSR, sample items can either be presented simultaneously or serially. When items are presented simultaneously, they necessarily each appear at a different location, and it is an item's unique location that is used by the

retrocue to designate it the PMI. Thus, the location at which an item appears serves as critical trial-specific context. When items are presented serially, they can appear at the same or different locations, and when they appear at the same location (as they did in Yu, Teng & Postle, 2020) the retrocue must designate the prioritized item by referring to the order in which it was presented (i.e., "first" or "second;" the item's ordinal context). When we simulated the DSR from Yu, Teng and Postle (2020), the unexpected finding was that the representation of the first sample item underwent a dramatic transformation upon the onset of the second item (i.e., prior to the designation of priority, which would be indicated by the retrocue). Specifically, whereas it had been represented in a subspace defined by the first two principal components of a PCA applied to the hidden layer or the RNN, it was displaced from this subspace by the representation of the second item, and shunted to a new subspace defined by the third and fourth principal components of the PCA. This finding caused us to reconsider our interpretation of the transformational dynamics observed in the 2-back task (Wan et al., 2022), because an item's functional trajectory during that task confounds priority with context. That is, while an item has the status of UMI it also has the contextual status of item-that-was-presented-most-recently (i.e., "1-back"), and when it then transitions to the status of PMI its context simultaneously transitions to item-that-was-presented-2-back. The aims of this report, therefore, are two-fold. One is to explore, at the

algorithmic level, how context-based representational transformations may differ from priority-based transformations. This will be carried out via RNN simulations. The second is to assess how these two properties, context and priority, might differ in the way they influence behavior, and in the way they are represented in the brain.

Methods

The data presented here derive from three sources: RNN simulations of a double serial retrocuing (DSR) task; reanalysis of data from an EEG study of DSR; and reanalysis of an fMRI study of DSR.

Participants

EEG

The EEG data set is from 12 healthy young adults (5 females, average age = 21.7 ± 3.2 years, all right-handed), as described in detail in Fulvio and Postle (2020). This *N* was double that of a previous EEG study for which MVPA decoding results yielded informative prioritization effects (Rose et al., 2016), and so was deemed satisfactory for the analyses to be carried out here.

fMRI

The fMRI data set is from 13 healthy young subjects (10 females, average age = 21.1 ± 4.5 years, all right-handed), as described in detail by Yu, Teng, and Postle (2020). Because IEM analyses of these fMRI data had yielded informative prioritization effects, this *N* was deemed satisfactory for the analyses to be carried out here.

Behavioral tasks

Recurrent Neural Network (RNN) models

The training task was modeled after the fMRI task (Yu, Teng & Postle, 2020; Figure 3.1B). Stimuli were randomly drawn from a pool of oriented gratings that covered the continuous range from [0°, 180°) interval (*Sample 1*: φ and *Sample 2*: θ). Stimulus location was not simulated, and it was possible for φ and θ to take on the same orientation. Each trial began with the presentation of *Sample 1* (50 timesteps) followed by an interstimulus interval (ISI, i.e., blank delay; 50 timesteps) followed by *Sample 2* (50 timesteps) followed by *Delay 1.1* (50 timesteps) followed by *Cue 1/Response 1* (50 timesteps; the response window was the duration of *Cue 1*). Next came another ISI (50 timesteps) followed by *Cue 2/Response 2* (50 timesteps). *Cue 2* matched ("stay") or did not match ("switch") *Cue 1*, unpredictably, and equal number of times.

fMRI: DSR with ordinal and location context, and recall probes

Stimuli were drawn from a pool of 9 oriented gratings that evenly covered the range from 0° to 179°, and could be presented at one of 9 locations that, each at a distance of 8° of visual angle from central fixation, evenly covered the range of possible locations from 0° to 359° of polar angle. Each trial began with the presentation *Sample 1* (.75 sec) followed by an ISI (.5 sec), followed by *Sample 2* (.75 sec), followed by *Delay 1.1* (8 sec), followed by a centrally presented digit ("1" or "2," *Cue 1*; .75 sec). After the ensuing *Delay 1.2* (8 sec), a recall dial appeared at the location that had been occupied by the PMI, and the subject had 4 sec to rotate it to match their memory of that item's orientation. Subsequently, after a brief unfilled interval (.5 sec), a second centrally presented digit ("1" or "2," *Cue 1*; .75 sec) indicated the item to be tested, after *Delay 2* (2 sec), at *Recall 2* (4 sec). *Cue 2* matched ("stay") or did not match ("switch") *Cue 1*, unpredictably, an equal number of times (Figure 3.1A).

Because the location of the recall dial indicated the item to be recalled, a possible strategy would be to ignore the cues and simply behave based on the location of the recall dial. However, this strategy was discouraged due to an important detail of the procedure. On each trial, the orientation and the location of each stimulus were selected at random (with replacement), and independently. Thus, on each trial there was a p = .11 chance that the second sample would have the same orientation as the first and, independently, a p = .11 chance that the second sample would appear at the same

location as had the first. These contingencies encouraged subjects to not wait for the onset of the recall dial to recall the orientation of the PMI and, indeed, patterns of priority-related transformation of the UMI during *Delay 1.2*, as assessed by IEM, confirmed that subjects used the ordinal cue to guide their behavior (Yu, Teng, and Postle, 2020).

EEG: DSR with location context and recognition probes

Each trial began with the simultaneous presentation of two sample items, one drawn from each of two out of three possible categories (faces, direction of dot motion, and words), one appearing above and one below central fixation (2 sec; Figure 3.1C). The samples were replaced by a central fixation symbol ("+") during an initial delay (*Delay 1.1*; 5 sec), followed by a dashed line appearing at one of the two sample locations (.5 sec), indicating that that item would be the first to be tested (*Cue 1*). After a second delay (*Delay 1.2;* 4.5 sec), during which the cued item had the status of prioritized memory item (PMI) and the uncued item the status of unprioritized memory item (UMI), an image serving as a recognition probe appeared centrally, and was either identical to the PMI ("match," p = .5), drawn from the same category but a different exemplar than the PMI ("nonmatch," p = .3), or identical to the UMI (also "nonmatch", p = .2; *Probe 1;* 1 sec). *Probe 1* was replaced by the fixation symbol (*Response 1*; 1 sec), and

a response was required during the 2 sec spanning Probe 1 and Response 1. Next a dashed line appeared at one of the two sample locations (*Cue 2*; .5 sec), thereby designating the PMI for the following *Delay 2* (4.5 sec), then *Probe 2* (1 sec) then *Response 2* (*1 sec*). ITI varied from 2-4 sec.

Data were collected during three sessions, each on a separate day, with each session comprising eight 30-trial blocks, alternating between blocks of DSR and a single retrocue task (results from single retrocue task not presented here). During each block *Cue 1* appeared unpredictably at the "up" or "down" location an equal number of times and, orthogonal to *Cue 1* location, *Cue 2* appeared, unpredictably, at the same ("Stay," Figure 3.1C, top row) or opposite ("Switch," Figure 3.1C, bottom row) location as had *Cue 1* an equal number of times. Balanced across cue conditions, spTMS was delivered 2-3 sec after the offset of *Cue 1* on 50% of trials and, orthogonally, after the offset of *Cue 2* on 50% of trials. Note that the EEG data used for the "transformation efficacy analyses" (see the "Analysis procedures" section below) include both epochs with and without spTMS.

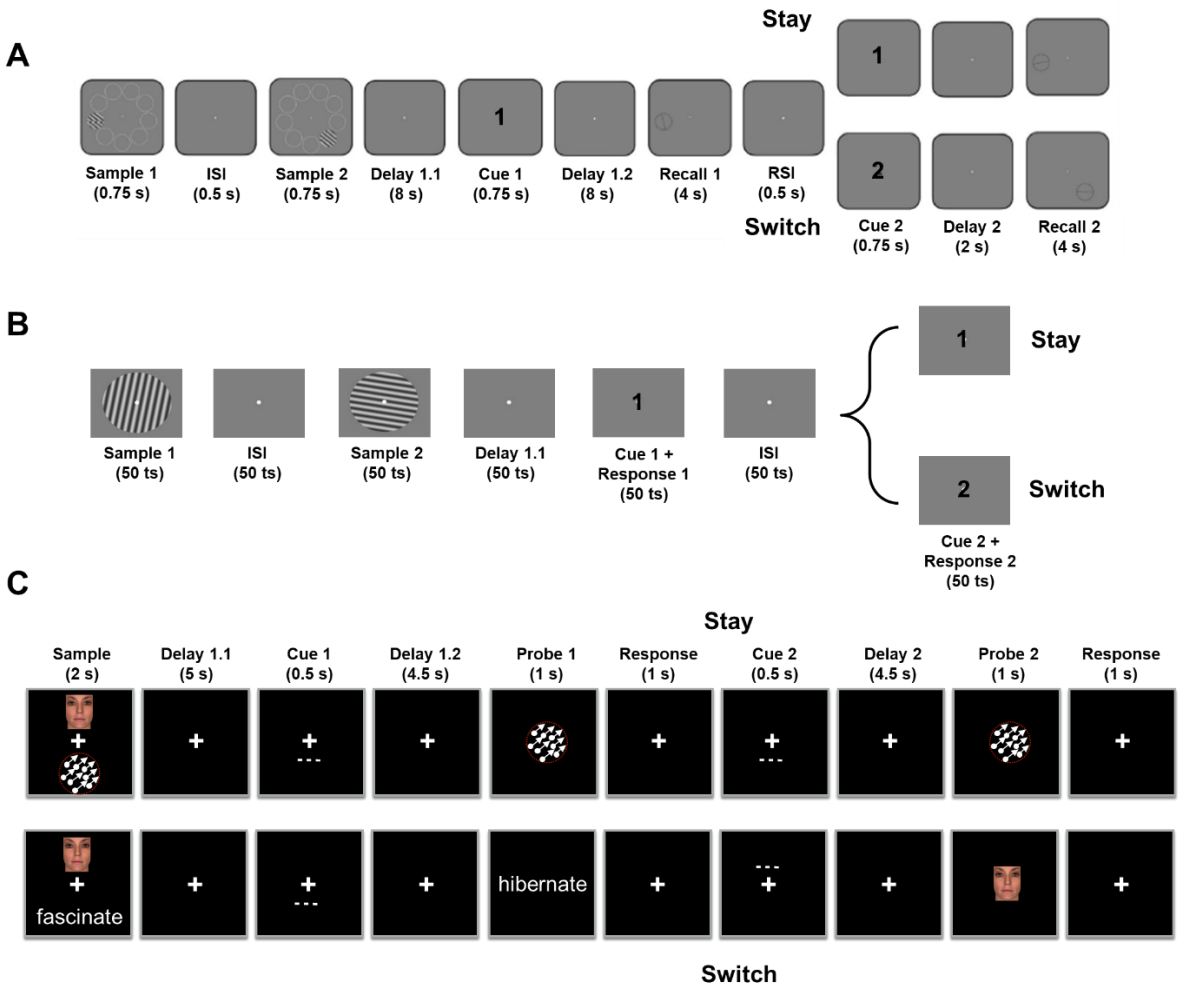


Figure 3.1. Experimental procedure for (A) The fMRI task, (B) the RNN task and (C) the EEG task. Figures adapted, with permission, from Yu, Teng and Postle (2020; panel A), Fulvio and Postle (2020; panel C).

Experimental procedures

RNN

Stimulus orientations were fed into the network via 32 orientation-tuned input units whose preferred orientations spanned the full 180° range, and whose response properties were based on V1 orientation-selective neurons (Teich & Qian, 2003; Figure 3.2). A 33rd input unit was used for retrocue, inputting "0" on each non-cue timestep and a "1" or "-1" (indicating "1st" or "2nd," respectively) during each cue timestep. The two

output units were trained to produce $\cos(2x)$ and $\sin(2x)$ (where x was either θ or φ depending on the cue) so that the 0° orientation had the same output as the 180° orientation (Figure 3.2).

Our network had 100 fully-connected recurrent units and the dynamics $u_i(t)$ of each recurrent unit were governed by the following standard continuous-time RNN equations:

$$\tau \frac{dx_i(t)}{dt} = -x_i(t) + \sum_{j=1}^{N^{rec}} W_{ij}^{rec}\, u_j(t) + \sum_{k=1}^{N^{in}} W_{ik}^{in} I_k(t) + b_i$$

$$u_i(t) = f\big(x_i(t)\big) + \xi_i(t)$$

for i = 1, …, N^{rec}. We introduced nonlinearity using the rectified linear unit (ReLU) function $f(x) = \max(0, x)$. Each recurrent unit received input from other units via recurrent connections with weights specified by the matrix W^{rec}, initialized orthogonally (Saxe et al., 2013). In addition, these units received external input $I(t)$ to the RNN via weights specified by the matrix W^{in}. Each unit carried two sources of bias: (1) b_i, learned during training, and (2) $\xi_i(t)$, which represented intrinsic noise in the network and took the form of white Gaussian (sampled independently at each timestep) with zero mean. We simulated the approximate network dynamics using the Euler method for T = 350 timesteps, each having a duration $\tau/10$ (Mante et al., 2013). We chose dt/τ = 0.1 similar to (Cueva et al., 2021); e.g., dt = 10 ms and τ = 100 ms, which would make the time scale of our simulations close to that of the fMRI experiment. The

outputs $y_j(t)$ were then generated by combining the activities of the recurrent units based on:

$$y_j(t) = g\left(\sum_{i=1}^{N^{rec}} W_{ji}^{out}\, u_i(t)\right)$$

where g is the tanh activation function.

We optimized the network parameters W^{in}, W^{rec}, b and W^{out} to minimize the mean squared error between the target outputs and the network outputs:

$$E = \frac{1}{MTN^{out}} \sum_{m,t,j=1}^{M,T,N^{out}} (y_j(t,m) - y_j^{target}(t,m))^2.$$

Parameters were updated with the Adam stochastic gradient descent (SGD) algorithm (Kingma & Ba, 2014) and each network was trained for 10,000 epochs.

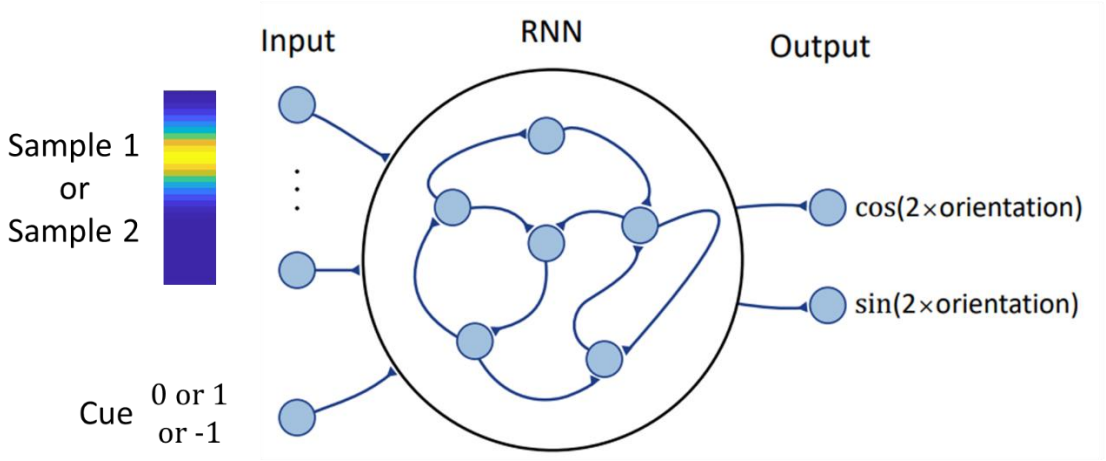


Figure 3.2. RNN input and architecture. Top left illustrates input of a stimulus with an angular value corresponding to the peak magnitude of this 32-dimensional vector; bottom left illustrates that at each timestep the value of the input to the cue input unit was *0*, *1*, or *-1*.

fMRI

For each subject regions of interest (ROIs) were defined, both anatomically and functionally, for eight regions: early visual cortex (EVC, V1 and V2 merged), IPS0-through-IPS5 (6 ROIs), and FEF (all ROIs cover both hemispheres). First, anatomical ROIs were defined by extracting masks from the probabilistic atlas of Wang and colleagues (2015) and warping them to each subject's structural scan in native space. To identify task-related activity, we modeled each epoch of the task with 6 boxcar regressors in a general linear model (GLM) – *Sample* (2 sec), *Delay 1.1* (8 sec), *Delay 1.2* (8 sec), *Recall 1* (4 sec), *Delay 2* (2 sec), and *Recall 2* (4 sec) convolved with a canonical hemodynamic response function and we also included covariates to control for motion. We proceeded to create anatomically constrained functional ROI for bilateral EVC by selecting the 500 voxels inside the V1-2 anatomical ROI with the strongest loading on the *Sample* regressor and for bilateral IPS0-5 and FEF by separately selecting the 500 voxels inside each of IPS0-5 and FEF anatomical ROIs, with highest loading on the *Delay 1.2* regressor.

EEG data collection

The experimental procedure from the experiment reported by Fulvio and Postle (2020) entailed recording the EEG with concurrent delivery of single pulses of transcranial magnetic stimulation (spTMS) on half of the delay periods of the DSR task.

However, because the original report only included behavioral results (with and without spTMS), here we detail the EEG methods.

EEG was recorded with a 60-channel cap and TMS-compatible amplifier, equipped with a sample-and-hold circuit that held amplifier output constant from 100 μs before stimulation to 2 ms after stimulation (NexStim eXimia, Helsinki, Finland). Electrode impedance was kept below 5 kΩ. The reference electrode was placed superior to the supraorbital ridge. Eye movements were recorded with two additional electrodes, one placed near the outer canthus of the right eye, and one underneath the right eye. The EEG was recorded between 0.1 and 350 Hz at a sampling rate of 1450 Hz with 16-bit resolution.

Data were processed offline using EEGLAB (Delorme & Makeig, 2004) with the TMS-EEG signal analyzer (TESA) open-source EEGLAB extension (Mutanen et al., 2020; Rogasch et al., 2017) and Fieldtrip (Oostenveld et al., 2010) toolboxes in MATLAB. The pipeline followed the TMS-EEG analysis pipeline (http://nigelrogasch.github.io/TESA/). Then, electrodes exhibiting excessive noise were removed and the data were epoched to -12 s to 8 s around the first spTMS event tag (*Delay 1.2*) and -4.5 s to 4.5 s around the second spTMS event tag (*Delay 2*). The data were downsampled to 500 Hz. In order to minimize the TMS artifact in the EEG signal, the data were interpolated using a cubic function from -2 to 30 ms around the TMS pulse, and this interpolation was also carried

out on delay periods on which TMS was not delivered. (For delay periods for which no spTMS was delivered ("spTMS-absent"), a dummy spTMS event tag was added at a latency that matched the most recent spTMS-present delay period.) The data were bandpass filtered between 1 and 100 Hz with a notch filter centered at 60 Hz. Independent component analysis (ICA) was used to identify and remove components reflecting residual muscle activity, eye movements, blink-related activity, residual electrode artifacts, and residual TMS-related artifacts. A spherical spline interpolation was applied to electrodes exhibiting excessive noise. Finally, the data were re-referenced to the average of all electrodes that were included in the ICA.

The present analyses included EEG data from all delay periods (i.e., averaging data from spTMS-present and spTMS-absent trials and ignoring this factor).

Analysis procedures

PCA visualization of the RNN hidden layer activity

We extracted from each network the activity of the 100 recurrent units from all 1000 testing trials and used PCA to project these 100-dimensional activity patterns onto the four dimensions accounting for the most variance across all training trials separately for each timestep. We then visualized the representations of each *Sample 1* and *Sample 2* by plotting the dimensionality-reduced activity across the 350-timestep

time course of a trial, and coloring the activity patterns according to stimulus identity, separately, in three 2D plots (PC1-2, PC2-3 and PC3-4).

In addition, we plotted the effective dimensionality (ED) of the data at each timepoint, which is the equivalent number of orthogonal dimensions that would produce the same overall pattern of covariation (Del Giudice, 2021). It is calculated using the following formula:

$$ED = \frac{\left(\sum_{i=1}^{N} \lambda_i\right)^2}{\sum_{i=1}^{N} \lambda_i{}^2}$$

where λ_is are the eigenvalues of the covariance matrix of the N recurrent units' activities at a certain time point.

Transformational efficacy analyses on EEG and fMRI data

The PCA visualizations of RNN activity revealed representational dynamics, such that stimulus information was represented differently as a function of context (1st or 2nd) and as a function of cue identity (essentially priority (PMI or UMI/IMI) for each stimulus). To assess the functional relevance of these two coding schemes for human behavior, we assessed trial-by-trial variation in the efficacy of context-based and priority-based transformations, and determined for each whether this variability related to trial-by-trial variation in behavior.

For the representation of context, we first calculated a template stimulus representational format for each subject by averaging the neural activity for each context status ("1st" or "2nd" for fMRI; "up" or "down" for EEG) over a time window corresponding to *Delay 1.1*, across all trials. (For the remainder of this section, for simplicity, we will only refer to ordinal context.) To these two windowed averages we applied demixed principal component analysis (dPCA; refer to Wan et al. 2022 for methodological details) to derive the first two demixed principal components (PCs), thereby constructing a *Sample 1* template subspace and a *Sample 2* template subspace. We then projected individual trial activity from the same time window into the template subspaces and calculated the "transformational efficacy index" (TEI) for that trial's representational transformation into the *Sample 1* subspace and its representational transformation into the *Sample 2* subspace. TEI was defined as the Euclidean distance between that trial's representation in the subspace and the template representation, normalized by the distance between the two template representations in that subspace. (E.g., for trial *n*, the TEI for the *Sample 1* subspace would be the Euclidean distance between the trial representation projected into the *Sample 1* subspace and the *Sample 1* template (projected into the *Sample 1* subspace), divided by the distance between the *Sample 2* template projected into the *Sample 1* subspace and the *Sample 1* template (projected into the *Sample 1* subspace). A lower average TEI for a context-based

transformation (e.g., the *Sample 1* transformation) corresponded to lower variability, for that subject, of that transformation, which we interpreted as higher "transformational efficacy." For the fMRI data, we used TR 5-7 to define the *Delay 1.1* subspaces, and for the for the EEG data we used the entirety of *Delay 1.1*.

For priority-based transformations, we followed the same procedures, but used TR 9-11 to define the *Delay 1.2* subspaces and the entirety of *Delay 1.2* period for the EEG data, and labeled the data according to priority status (i.e., PMI and UMI).

If the efficacy of a context-based transformation is important for behavior, smaller TEIs should be associated with superior performance. To assess this in the fMRI data, for each subject we sorted responses, separately for *Recall 1* and for *Recall 2*, by median split of angular error, then calculated, for each response, the average TEI for each type of transformation (e.g., "what was the average TEI for the transformation to *Sample 1* for low-error vs. high error responses to *Recall 1*?"). Then we performed paired-samples *t*-tests between group-average high-error and low-error TEIs, separately for each subspace, each brain region, and each response (*Recall 1* and *Recall 2*). The analysis procedure was similar for the EEG data except that the comparison was between incorrect and correct responses.

To test how the TEI for UMI and PMI covary, we ran two-sided Spearman's rank correlations between the two metrics across all trials for each subject and counted the number of subjects with correlations reaching the significance level of $\alpha = .05$.

Within- and cross-label decoding of RNN and fMRI data

To assess where in the brain context and priority are represented, we carried out a series of decoding analyses applying the following logic. If a region represents context, any given stimulus item will be represented differently when it has the status of, for example, *Sample 1* versus when it has the status of *Sample 2*. If a decoder applied to data from this region can be successfully trained to classify stimulus identity when the data are labeled as *Sample 1* (successful "within-label" decoding), it should fail to decode stimulus identity when the data are relabeled as *Sample 2* (unsuccessful cross-label decoding). If a region that does not represent context, in contrast, any given item's representational format will not differ as a function of its context status, and so a decoder that can be successfully trained on the data labeled as *Sample 1* should succeed at decoding stimulus identity when the data are relabeled as *Sample 2* (successful cross-label decoding). Before carrying out these analyses, we assumed the results with RNNs will have demonstrated that they can be trained to perform the DSR task. Therefore, it would necessarily be true that they represented both context and priority, and so

applying these analyses to the RNN data would act as a sanity check for this logic. To be consistent with the fMRI dataset, RNN data were generated by testing the trained network on 324 trials of 9 possible orientations (counterbalanced across the identities of *Sample 1*, *Sample 2*, *Cue 1* and *Cue 2*, to be analogous to the Yu, Teng and Postle (2020) task), and subsequently extracting the RNN hidden layer activity. For the RNN data we decoded orientation and for the fMRI data we decoded location. (Decoding item location is generally more sensitive than decoding item orientation, and so demonstrations of failures of cross-label decoding of item location would provide stronger evidence for the encoding of the stimulus property of interest.)

For the RNN data and for the fMRI data from each ROI, we trained linear Support Vector Machine (SVM) multiclass classifiers to decode stimulus identity with a *k*-fold cross-validation procedure and a 'one vs one' coding design (see Supplementary Materials S3.1 for comparisons with results from other decoding methods). For context-based decoding, for each subject and at each timepoint, we trained a classifier with the data labeled as *Sample 1* then tested it on the data labeled as *Sample 1* (within-label decoding) and with the data labeled as *Sample 2* (cross-label decoding). We then repeated this process by training on *Sample 2*, and with fMRI data, for simplicity, we averaged the results to generate the overall accuracies for within-label decoding and cross-label decoding. For priority-based decoding, we used the same procedure except

that the labels were *PMI* and *UMI*, instead of *Sample 1* and *Sample 2*, the *PMI/UMI* label reassigned at timestep 301 (for RNN) or TR 15 (for fMRI) to reflect identity of *Cue 2* (i.e., to account for the fact that priority status changed partway through "switch" trials). For the fMRI data, to evaluate the significance of decoding accuracy against chance level (1/9), we performed one-tailed one-sample *t*-tests against 1/9 on decoding accuracies across all subjects, and corrected for multiple comparisons using the false discovery rate (FDR) method.

Results

RNN

PCA visualization of hidden layer activity

PCA was carried out on the RNN hidden layer activity across all timepoints from 1000 withheld testing trials with *Sample 1* and *Sample 2* spanning the [0°, 180°) angular range and the resultant dimension-reduced activity projected into 3 subspaces that were spanned by PC1-PC2, PC2-PC3 and PC3-PC4, respectively, on a timepoint-to-timepoint basis. We trained three RNNs using the same training regime, used the PCA visualization of hidden layer activity from the first two for hypothesis generation and validation, and report results from the third network. The dynamical representational

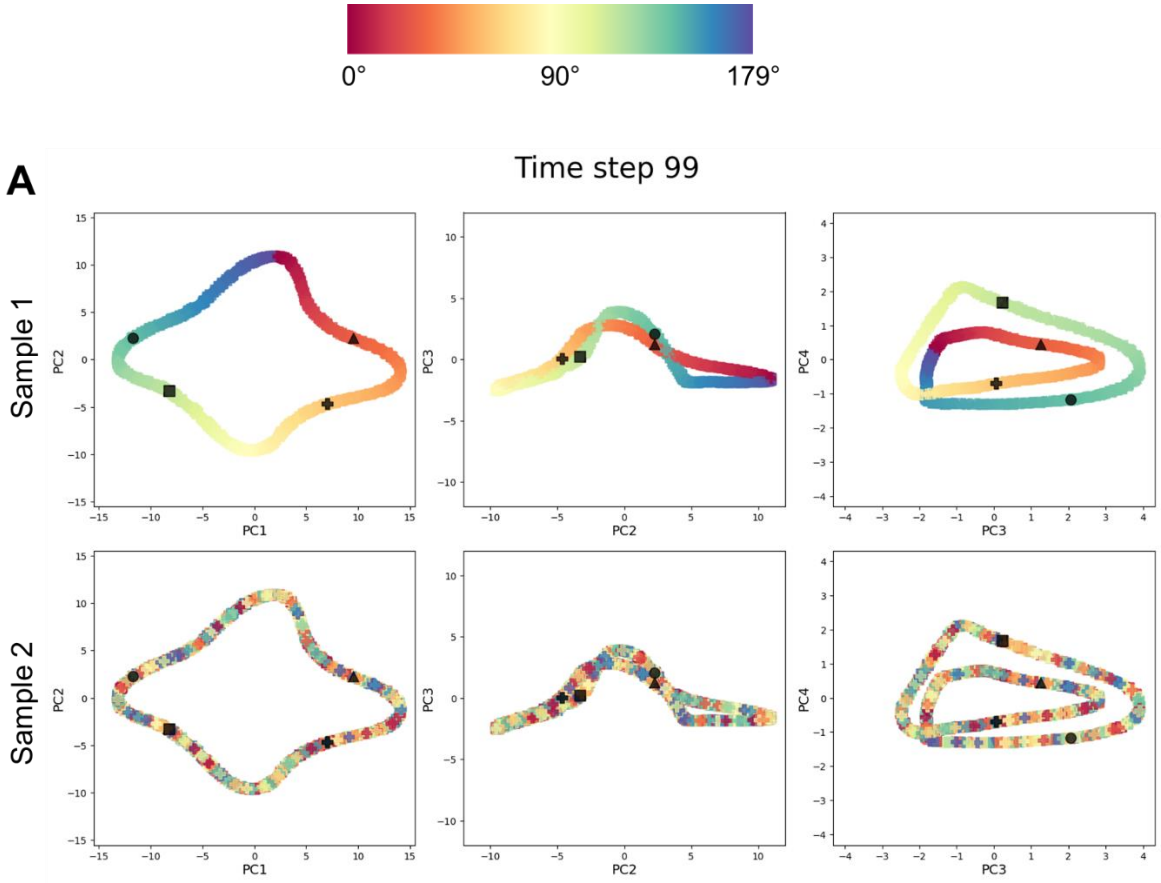
patterns observed in all 3 networks were highly consistent (See Supplementary Movie 1).

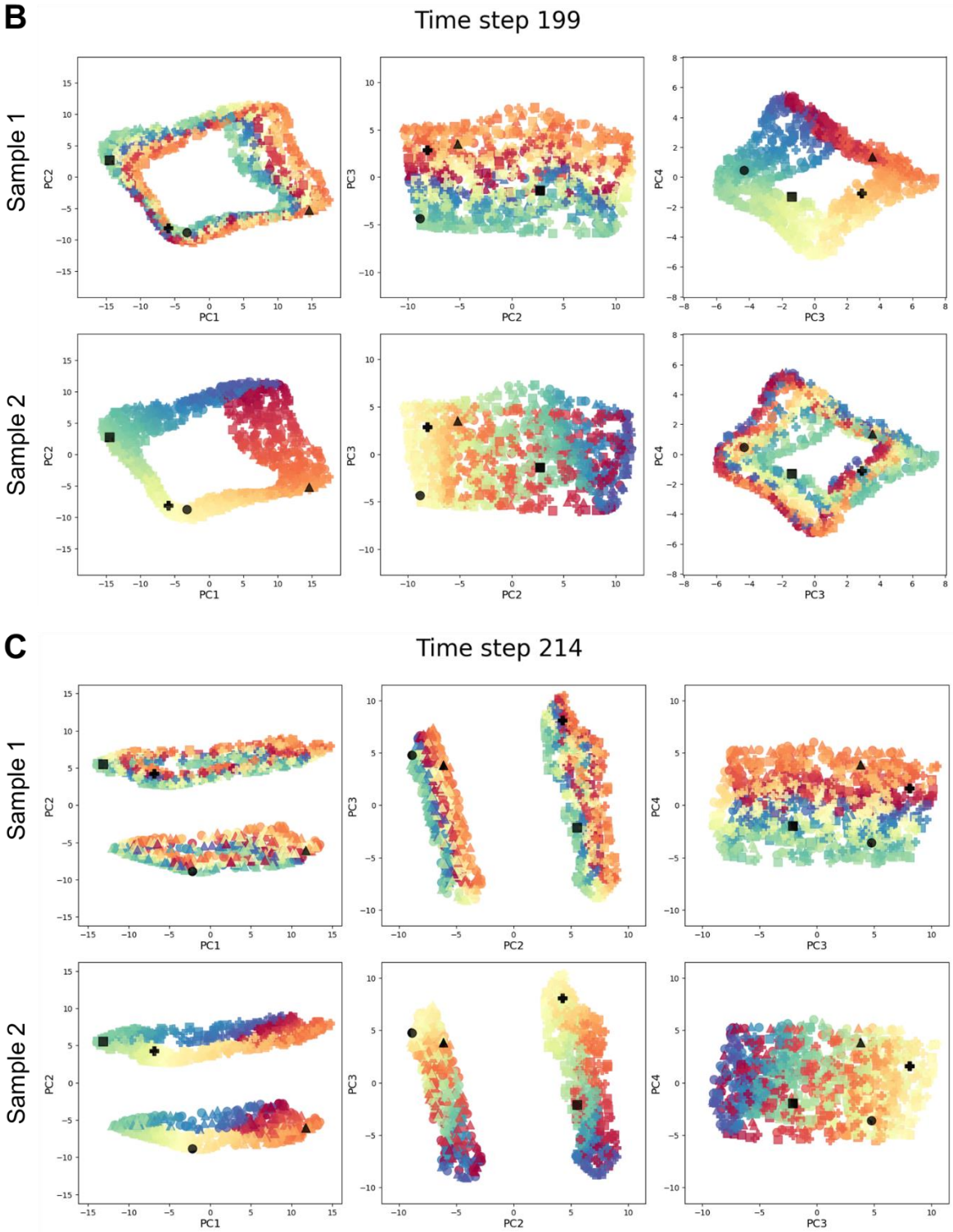
Upon the presentation of *Sample 1*, its representation formed a ring in the subspace spanned by the first 2 PCs, with relative distances between stimulus values preserved (as shown by the smooth color gradient of the ring; Figure 3.3A, top left panel), such that stimulus value can easily be read out from this subspace. Although there are also smooth color gradients in the other two subspaces, their geometry is more complex, making it less clear if they would support readout. The ring structure in the PC1-PC2 subspace was maintained across the ensuing ISI (see Figure 3.3A and Supplementary Movie 1). After the presentation of *Sample 2*, *Sample 2*'s identity was represented in the PC1-PC2 subspace, also in the form of a ring with a smooth color gradient (although the ring was somewhat "stretched out" relative to timestep 99; Figure 3.3B, bottom left panel). In parallel, information about *Sample 1* emerged in the subspace spanned by PC3 and PC4, in the shape of a ring with a smooth (albeit "stretched out") color gradient (Figure 3.3B, top right panel). In effect, whereas *Sample 1* was represented in the PC1-PC2 subspace when it was the only item in WM, it was shunted to the PC3-PC4 subspace with the presentation of *Sample 2*, which replaced *Sample 1* in the PC1-PC2 subspace. Thus, prior to cuing, the RNN encoded the ordinal context of *Sample 1* and *Sample 2* by segregating them in orthogonal subspaces.

Upon the presentation of *Cue 1* (at timestep 201), the stimulus representations within each subspace separated into two clusters that were defined by priority status. For example, Figure 3.3C illustrates that in the PC1-PC2 subspace, at timestep 214, trials for which *Sample 1* was cued (denoted by triangle and circle symbols) separated from trials for which *Sample 2* was first cued (square and plus-sign symbols). Throughout the *Cue 1* epoch the axis along which this separation occurred rotated in multidimensional space over time. Thus, whereas timestep 214 was selected for Figure 3.3C because it clearly shows this separation-by-priority status in the PC1-PC2 subspace; the separation was visible in the PC3-PC4 earlier during this epoch, at timestep 207 (see Supplementary Movie 1). Thus, the RNN encoded priority status via separation within each subspace.

During the delay between *Cue 1* and *Cue 2* (timesteps 251-300), the prioritization clusters merged such that, prior to the presentation of *Cue 2*, information about *Sample 1* and *Sample 2* was again clearly observed in the PC3-PC4 subspace and in the PC1-PC2 subspace, respectively (Figure 3.3D). Finally, upon the presentation of *Cue 2*, the network representation once again separated into two priority-defined clusters, this time based on *Cue 2*'s identity, (i.e., trials for which *Sample 1* was cued (denoted by circle and square symbols) and trials for which *Sample 2* was cued (triangle and plus-sign symbols) separated into two clusters; Figure 3.3E). Thus, visualization of the

representational of the RNN recurrent unit activities revealed that context and priority were represented via different transformational mechanisms, the former via the *segregation* of stimuli to orthogonal subspaces, and the latter via *separation* within each subspace.




B
Time step 199
Sample 1
Sample 2
PC1
PC2
PC3
PC4
C
Time step 214
Sample 1
Sample 2
PC1
PC2
PC3
PC4

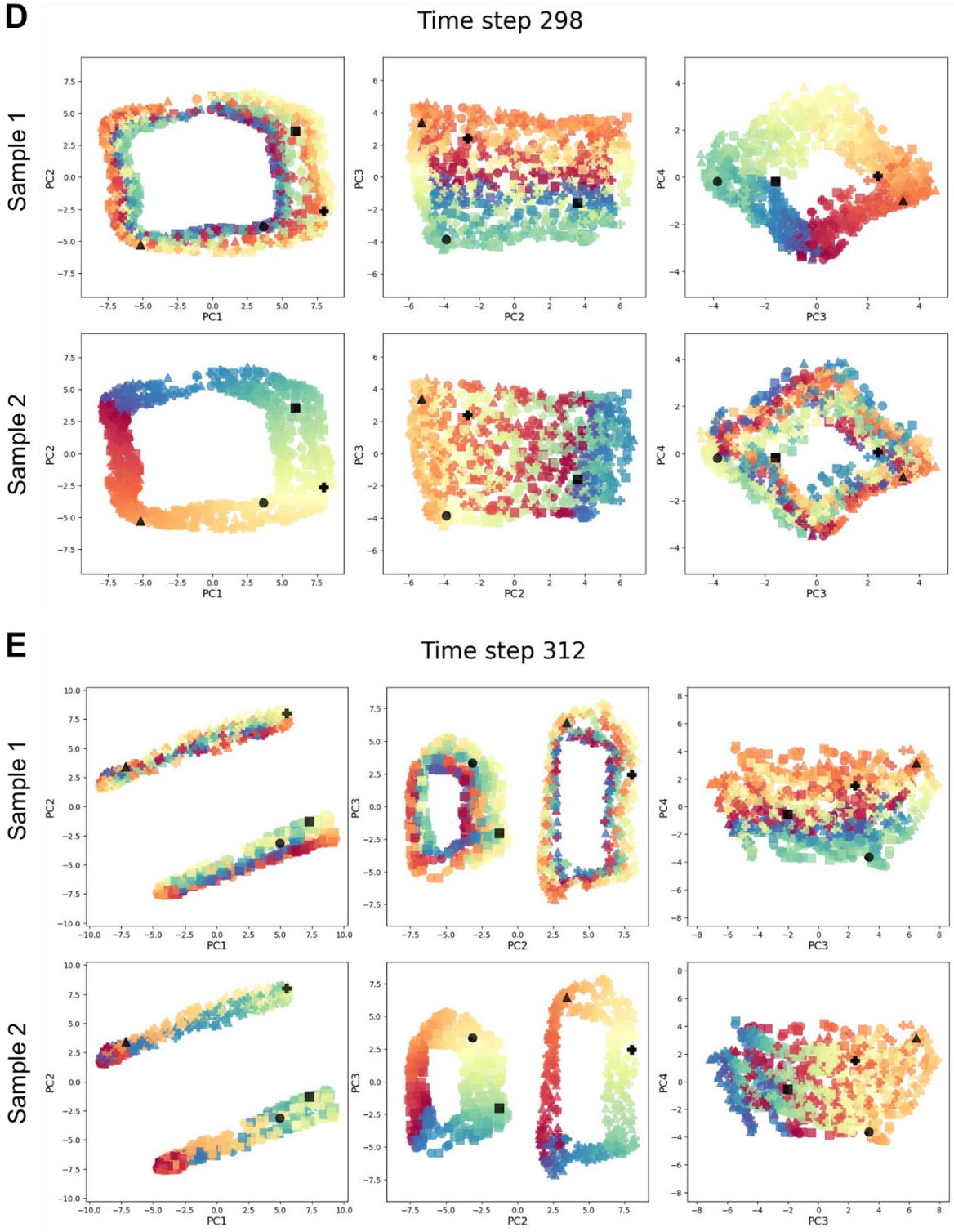


Figure 3.3. Visualization of representational dynamics embedded in RNN hidden layer at each of five representative timesteps across the DSR task. Each plot contains 1000 data points, one corresponding to each simulated trial, and the symbols indicating that trial's cue configuration: *Cue 1 -> Sample 1*, *Cue 2 -> Sample 1*(●); *Cue 1 -> Sample 1*, *Cue 2 -> Sample 2* (▲); *Cue 1 -> Sample 2*, *Cue 2 -> Sample 1* (■); *Cue 1 -> Sample 2*, *Cue 2 -> Sample 2* (+). In each plot, an example trial of each cue configuration is colored black for better visualization. For each of the

five timesteps the same data are illustrated in six ways: the top row with the data labeled as *Sample 1* and the bottom row with the data labeled as *Sample 2*, and for each they are projected into three subspaces. A. After the presentation of *Sample 1* (Timestep 99). Note that because *Sample 2* has not yet been presented, the stimulus values are haphazard. B. After the presentation of *Sample 2* (Timestep 199). With both items in WM, but prior to cuing, *Sample 1* is now represented in the PC3-PC4 subspace and *Sample 2* in the PC1-PC2 subspace. C. During presentation of *Cue 1* and generation of *Response 1* (Timestep 214), illustrating a separation-by-priority status in the PC1-PC2 subspace. (A comparable priority-based separation was visible in the PC3-PC4 subspace earlier during this same epoch (not shown).) D. During the delay between *Cue 1* and *Cue 2* (Timestep 298). E. During presentation of *Cue 2* and generation of *Response 2* (Timestep 312), again illustrating a separation-by-priority status in the PC1-PC2 subspace but now based on *Cue 2*. (As with the *Cue 1* epoch, a comparable priority-based separation was visible in the PC3-PC4 subspace earlier during this *Cue 2* epoch (not shown).)

Effective Dimensionality

During the processing of *Sample 1*, effective dimensionality (ED) initially rose to a value between 3 and 4 before declining to a value of ≈ 2 during the ensuing ISI (Figure 3.4). Upon the presentation of *Sample 2*, ED rose precipitously to a value close to 6 before declining steadily for the remainder of this epoch and the ensuing *Delay 1.1* to a value just below 3, which corresponds well to the encoding of a new stimulus and the segregation of subspaces to represent the ordinal context. The three remaining trial epochs were characterized by an initial increase of ED to a value of roughly 4 followed by a decline back to roughly 3. Particularly noteworthy in these results is the increase in ED following the offset of *Cue 1*. Note that because a similar increase in ED was not observed upon the offsets of the *Stimulus 1* or *Stimulus 2* epochs, this effect cannot be simply due to a transition from one epoch to the next. Rather, this effect closely

resembled those time-locked to the onset of *Cue 1* and to the onset of *Cue 2*, events that each prompted the separation of stimuli into priority-defined clusters (Figure 3.3C and 3.3E). Therefore, it may be that the operation of removing from the network the encoding of no-longer-relevant information about priority status related to *Cue 1* -- corresponding to the merging of priority-defined clusters that was observed in the PCA visualization – is also an operation that entails a transient increase in ED.

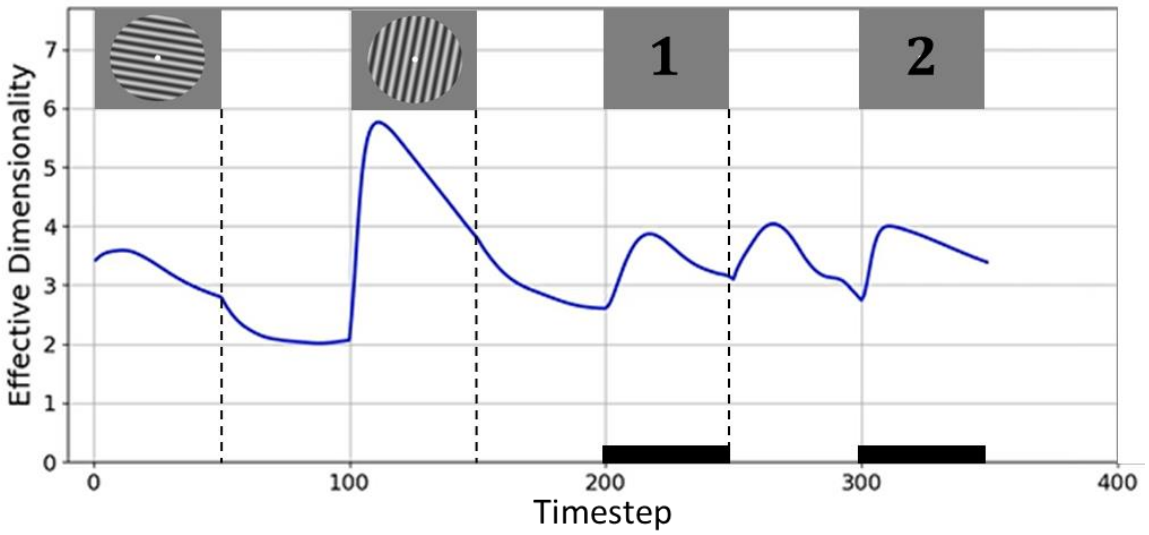


Figure 3.4. The time course of effective dimensionality (ED) of the RNN hidden layer stimulus representations. The rectangular images above the curve denote corresponding task events. The black rectangles along the *x*-axis indicate time periods when a response was being made.

Interim Discussion

We trained RNNs to perform the DSR task and applied dimensionality reduction to the internal representations of the network. Visualization of the representational dynamics yielded several important insights. First, prior to the first prioritization cue, information corresponding to the two sample stimuli was represented in orthogonal

subspaces (Panichello & Buschman, 2021). This may serve not only to individuate the two, but to encode the distinct ordinal context that the network needed to correctly interpret the cues. Second, priority status was represented by separating stimulus representations to distinct regions within each subspace, according to the cued "context". The first observation is important because it emphasizes the importance of encoding trial-specific context in WM, and previous empirical studies of prioritization have largely overlooked this operation. The second observation is important because it indicates that the representation of priority status may be implemented in a different way, via separation within a subspace, than is ordinal context, via segregation of stimulus information to distinct subspaces. This difference is all the more interesting when one considers that to the RNN, ordinal position and priority may be just two dimensions of task context that play out on different time scales during a single trial. In this variant of DSR, one dimension of an item's context is the order in which it was presented. This can be considered the "first-order" context because it uniquely individuates an item for the duration of a trial, and it does not change for the duration of the trial. (It is to first-order context that Oberauer and Lin (2017) refer when they state that the binding of context to a stimulus is fundamental to that stimulus being in the state of being "in WM".) A second dimension of context is priority status, and this differs from first-order context because its status for an item varies within the trial between

"not applicable," "prioritized," and "unprioritized" (indicated by values of 0, 1, and -1, respectively, being input by the cue unit). Thus, priority serves as a "second-order" level of context, one that indicates an item's in-the-moment status with respect to the rules of the task, and that cannot be interpreted in the absence of information about first-order context. These considerations highlight that to fully understand the flexible control of WM we need to understand how first-order context is coded in the brain and how it interfaces with higher-order context to guide thought and action.

Recent empirical studies that have manipulated demands on first-order context in WM have implicated regions of frontal cortex and the intraparietal sulcus (IPS) (for ordinal context, see Fulvio et al., 2023; Gosseries et al., 2018; for location context, see Cai et al., 2020; Fulvio et al., 2023). In Yu, Teng & Postle (2020), a study that also manipulated priority, the location context of differently prioritized orientation stimuli was found to be preferentially coded in IPS, and not early visual cortex, even though location information was not directly tested by the task. More recently, Teng and Postle (Stage 1-accepted registered report) used the same stimuli and procedure, but flipped the roles of context and content, making orientation the first-order context used to cue memory of an item's location. "Context load" was manipulated via the similarity of orientation of the two sample stimuli, and individual differences in context-load sensitivity of activity in IPS (but not early visual cortex) predicted behavioral sensitivity

to this factor. Generalizing across these studies suggests that first-order context in WM may be represented more prominently in areas associated with cognitive control than in areas associated with stimulus representation. The same may not be true for second-order context, because prioritization effects are prominent in early visual cortex (Yu, Teng & Postle, 2020; Teng & Postle, Stage 1-accepted registered report).

These considerations, prompted by the results from the RNNs, highlighted for us the importance of understanding the encoding of first-order context in WM, and of understanding similarities and differences of neural and behavioral correlates of first-order versus second-order context. What follows are initial attempts to do so, via reanalyses of an extant fMRI and an extant EEG dataset from two previous studies of the DSR task.

Analyses of fMRI and EEG data

The fMRI study used a DSR procedure that was most closely matched to that used with the RNN, including the fact that it used stimulus order as first-order context. The fMRI data would also allow for assessment of possible regional differences in the representation of the two types of context. The task used in the EEG study used location as the dimension of first-order context, and so would allow an assessment of generalization of what has been observed for ordinal context (with the RNN and fMRI) to location context. (For ease of exposition, in the results that follow we will refer to

first-order context as “context” and second-order context as “priority,” because priority is the only dimension of second-order context that is relevant in the DSR task.)

Transformational efficacy

One way to compare the neural representation of context versus priority is to assess their influence on behavior. To do this, we took an individual differences approach, using the variability of trial-by-trial encoding of context and of priority as proxies of the efficacy with which these operations were carried out. (I.e., a subject for whom context-based or priority-based transformations were more variable from trial-to-trial might be expected to perform worse on the task.)

For context, results failed to show evidence that behavior was sensitive to transformational efficacy. For the fMRI data (ordinal context), transformational efficacy indices (TEI) did not differ for low- vs. high-error trials, for *Recall 1* or *Recall 2*, in any of the 3 ROIs (early visual cortex, IPS 0-5, FEF; all $t(12) < 1.74$, *n.s.*). For the EEG data (location context), TEI did not differ for correct vs. incorrect trials ($t(11) < 1.37$, *n.s.*).

For priority, there was considerable evidence that behavior was sensitive to transformational efficacy. For the fMRI data, in early visual cortex, TEI was lower for low-error than high-error trials for the UMI subspace for *Recall 1* ($t(12) = 1.81$, $p < .05$) and for the PMI subspace for *Recall 2* ($t(12) = 2.06$, $p < .05$). For IPS0-5, TEI for the PMI

subspace was lower for low-error than high-error trials for *Recall 1* ($t(12) = 2.04$, $p < .05$), and was lower for low-error than high-error trials for both UMI ($t(12) = 3.04$, $p < .01$) and PMI ($t(12) = 3.00$, $p < .01$) subspaces for *Recall 2*. All other comparisons, including all for FEF, failed to achieve significance (all $t(12) < 1.57$, *n.s.*). For the EEG data, TEI was lower for correct trials than incorrect trials for *Recall 1* in both UMI ($t(11) = 2.17$, $p < .05$) and PMI ($t(11) = 4.28$, $p < .001$) subspaces.

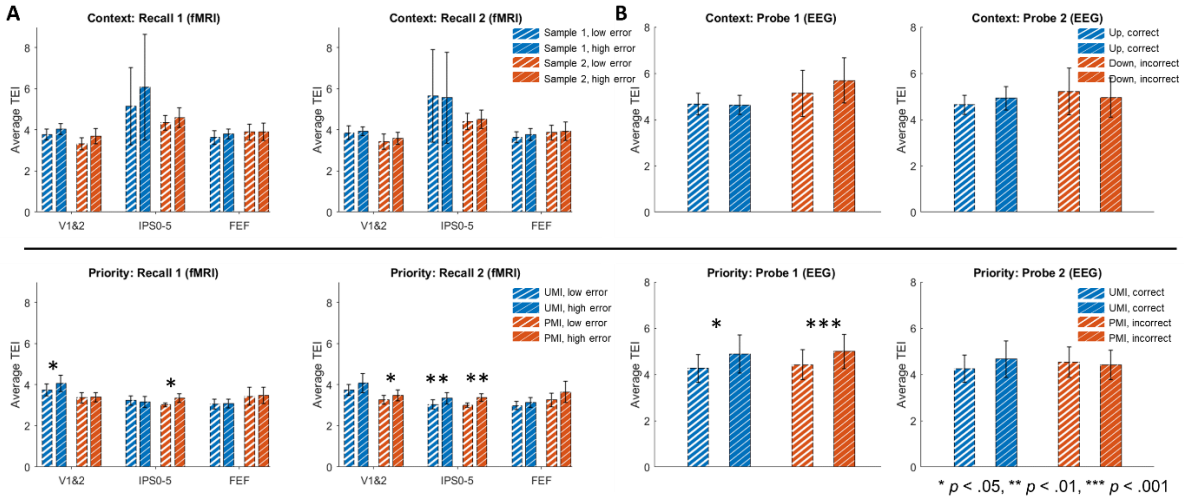


Figure 3.5. Transformational efficacy analysis results on fMRI (Yu, Teng & Postle, 2020) and EEG (Fulvio & Postle, 2020) data. (A) Comparisons between average TEI for high-error and low-error trials across subjects from the fMRI dataset. (B) Comparisons between average TEI for incorrect and correct trials across subjects from the EEG dataset. Top row: priority-based decoding; bottom row: context-based decoding. The subspace from which the TEI is calculated is indicated in the legends. Asterisks above bars of the same color indicate the significance level of the paired-sample *t* tests comparing the average TEI between each two groups.

The TEI also offered a metric with which to begin exploring whether the transformation to PMI and the transformation to UMI may share a common component that acts on the two simultaneously (c.f., Panichello & Buschman, 2021). Specifically, we

correlated trial-by-trial TEI for the PMI with trial-by-trial TEI for the UMI (two-sided), reasoning that evidence of correlation would be expected if the two do share an underlying mechanism. For the fMRI data, in early visual cortex, this correlation was significant at $p < .05$ for 12 out of 13 subjects, in IPS 0-5 for 11 subjects, and in FEF for 10 subjects. For the EEG data, TEIs for PMI and UMI were significantly correlated for 10 out of 12 subjects. All correlations were positive.

Within- and cross-label decoding of RNN and fMRI data

RNN data. To assess the representation of both context and priority in the RNN, we performed within- and cross-label decoding analyses on the RNN recurrent unit activities across all 350 timesteps from 324 novel, counterbalanced trials of 9 different orientations using a linear SVM classifier (Figure 3.6).

For context-based decoding we obtained close to perfect decoding accuracy when training and testing on the labels of the same sample throughout the task (note that for train S2, test S2 decoder performance was at chance prior to timestep 101, due to the absence of information about *Sample 2* at those time steps). For cross-label decoding, however, accuracy was at chance level for the duration of the trial. For priority-based decoding, within-label decoding accuracy for both PMI and UMI was close to chance level prior to *Cue 1*. With the onset of *Cue 1*, for both PMI and UMI, decoder performance

rose to close-to-perfect for the remainder of the trial. For cross-label decoding, whereas decoding accuracy for both PMI and UMI was above chance level prior to *Cue 1*, for both it dropped to chance level with the onset of *Cue 1*, and remained there for the remainder of the trial. Both of these sets of results validated the reasoning that a system that represents context and priority would not support cross-label decoding for either factor.

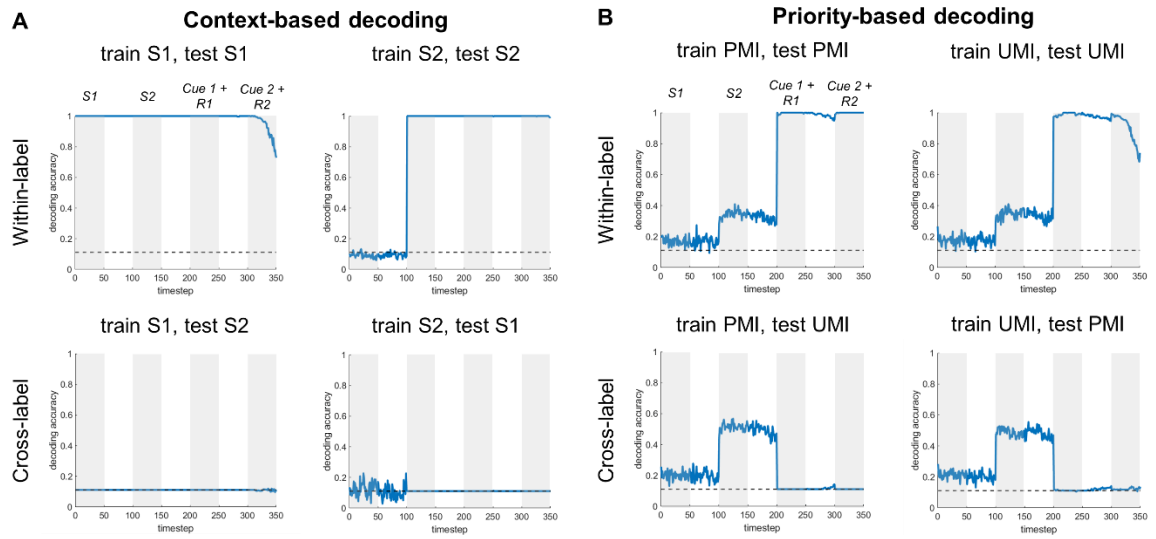


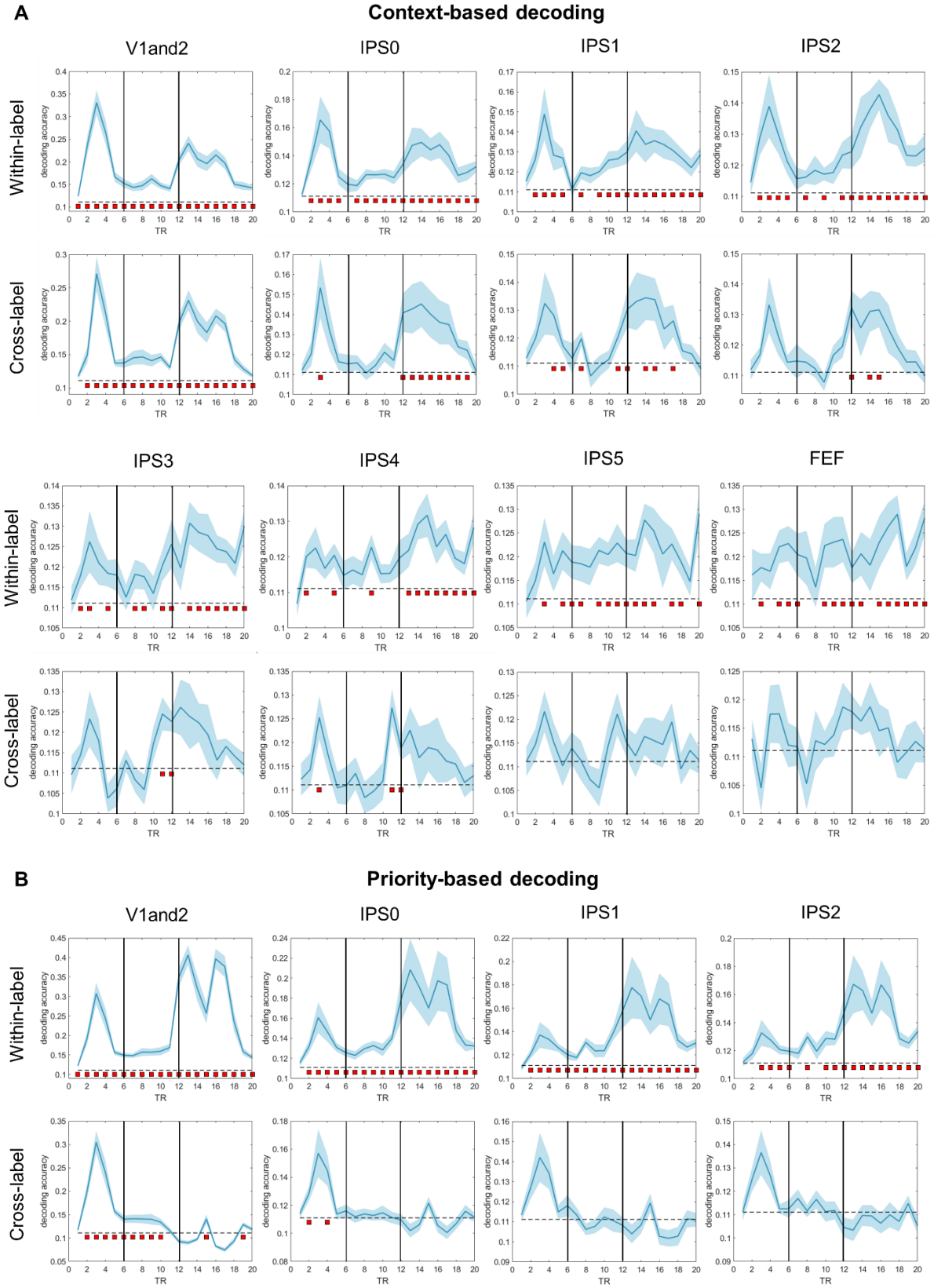
Figure 3.6. Within- and cross-label decoding of RNN data across the DSR trial. (A) Context-based decoding. Classifiers were trained on *Sample 1/2*, then tested on *Sample 1/2* (within-label), or tested on *Sample 2/1* (cross-label). (B) Priority-based decoding. Classifiers were trained on PMI/UMI, and tested on PMI/UMI (within-label) or tested on UMI/PMI (cross-label). S1: *Sample 1*, S2: *Sample 2*; R1: *Response 1*; R2: *Response 2*.

fMRI data. We investigated the anatomical distribution of the representation of context and priority during the DSR task by carrying out a series of multiclass decoding analyses on the fMRI dataset. In general (and unlike for the RNNs) decoder performance was far from ceiling, and tended to be superior for time points corresponding to trial

epochs when stimuli were on the screen. Importantly, however, we were generally able to decode the stimulus identity across the whole time course with above-chance accuracy in every ROI, especially in the time period between *Cue 1* and *Cue 2*, where one stimulus is prioritized over the other in working memory (within-label rows of Figure 3.7). (The one exception was in IPS4 with context-based decoding; the reason for this is unclear.)

For context, cross-label decoding revealed a marked posterior-to-anterior gradient: it was successful for the entirety of the trial in V1-V2 (indicating that changing the context does not affect the decoding accuracy, and hence, context information is likely ignored); successful for *Cue 2* and *Delay 2* epochs for IPS0 and for a smaller number of timepoints for IPS1 and IPS2; successful only for late *Delay 1.2* for IPS3 and IPS4, and entirely at chance for IPS5 and FEF (Figure 3.7A). This indicates that context was not represented at the earliest stations of the visual system and become progressively more prominent at progressively higher levels of the dorsal stream.

For priority, cross-label decoding for V1-V2 was successful for the beginning of the trial through late *Delay 1.2*, after which it dropped to baseline. For the remainder of the ROIs cross-label decoding was at baseline for the entirety of the trial. This suggests that priority is represented in every ROI that we investigated, albeit taking longer to manifest in V1-V2 (Figure 3.7B).


A
Context-based decoding
V1and2
IPS0
IPS1
IPS2
IPS3
IPS4
IPS5
FEF
Within-label
Cross-label
decoding accuracy
TR
B
Priority-based decoding
V1and2
IPS0
IPS1
IPS2
Within-label
Cross-label
decoding accuracy
TR

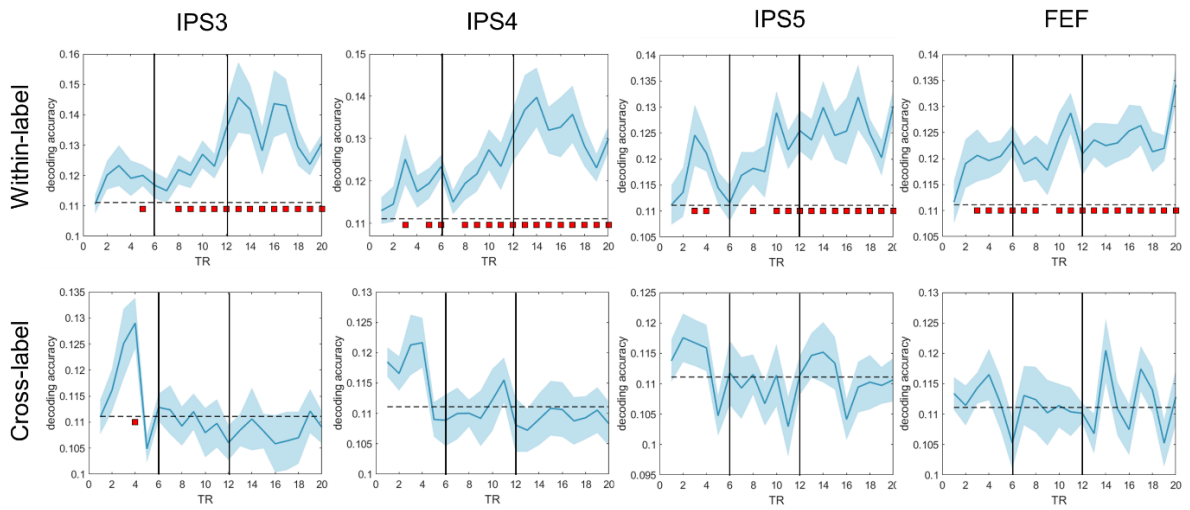


Figure 3.7. Within- and cross-label decoding analyses from the fMRI dataset. (A) Context-based decoding. (B) Priority-based decoding. In each graph, the two vertical solid black lines indicate *Cue 1* and *Cue 2*, respectively. The blue shading around each curve shows standard error of the mean. The horizontal dashed line indicates the chance-level decoding accuracy of 0.11. Red squares below the dashed line indicate time points with significant above-chance decoding accuracy ($p < .05$, FDR-corrected across all time points). Note that the range of the *y*-axis varies from graph to graph.

Discussion

In this study we initially set out to investigate the mechanisms underlying prioritization on a task in which changes of priority were not predictable – the double serial retrocuing (DSR) task – via visualization of representational dynamics of an RNN trained to perform the task. Unexpectedly, results from the RNN called to our attention the importance of also understanding the representation of an additional dimension of trial-specific information, the first-order context that uniquely individuates each item during the trial. Across model training, validation and testing, we saw that the encoding of first-order context was accomplished via the segregation into orthogonal subspaces of the representation of the first and second item to be presented. Unlike first-order

context, higher-order context can change within a trial, a property that is often manipulated with instructional cues. In the DSR, priority status is the second-order context, and it is specified, then removed, then specified a second time, during the course of each trial. The encoding of priority was accomplished via separation within each context-encoding subspace of prioritized from unprioritized items. Thus, the RNN indicated that first- and second-order contexts are encoded via distinct mechanisms, *segregation* to orthogonal subspaces versus *separation* within a subspace, respectively. Furthermore, an effective dimensionality analysis suggested that the operation of resetting second-order context (as happens during the ISI separating *Cue 1* and *Cue 2* in the RNN version of the task) may make computational demands that are comparable (in terms of requiring additional dimensions) to those needed to establish it.

Consistent with the distinct dynamics observed with RNNs, reanalyses of an fMRI and an EEG dataset established that the processing of first- and second-order contexts has distinct behavioral and neural profiles for humans performing the DSR. To assess relations to behavior, dimensionality reduction was applied to the neural data and transformational efficacy indices (TEI) derived for each subject for each of the two levels of first-order context and for each the two levels of second-order context. Correlations with behavior failed to show any evidence that performance is sensitive to variation in TEI for first-order context, whether first-or-second-to-be-presented (fMRI

study) or top-or-bottom-location-of-presentation (EEG study). For priority (i.e., second-order context), in contrast, there was considerable evidence that larger TEIs (indicating higher trial-to-trial variability) corresponded to poorer performance. In the fMRI dataset, the anatomical distribution of the representation of order and priority also differed, with the former absent from early visual cortex and becoming progressively more robust in more rostral ROIs, whereas the latter was evident in every ROI that we investigated. Thus, our results suggest that not only are representational transformations corresponding to first-order versus second-order context implemented via different mechanisms, they also differ according to their influence on behavior and to their distribution in the brain.

These results share some similarities and some differences with a recent study that recorded neuronal activity from several brain areas of nonhuman primates (NHP) performing a single-retrocue working memory task (Panichello & Buschman, 2021). In that study, dimensionality reduction revealed that, prior to the retrocue, the two stimuli were represented in orthogonal subspaces that corresponded to the location at which each had been presented (i.e., first-order context). Upon cuing, stimulus information transformed into different "post-selection" subspaces that retained first-order context and now also represented selection status (selected/non-selected; i.e. second-order context). Notably, the representations of "selected upper" and "selected lower" items

were no longer orthogonal. The degree of cue-triggered representational transformation was highest in dorsolateral PFC and progressively weaker in more posterior regions, weakest in extrastriate visual area V4. One similarity of those results to those reported here is the encoding of first-order context into orthogonal subspaces. A notable difference between the two is the nature of the post-cue transformations.

In the DSR, the representational transformation of one item into a PMI and the other into a UMI are prompted by the same cue, a design feature that allows for direct comparison of the two processes. For the majority of subjects in the EEG study, and in the majority of ROIs in the majority of subjects in the fMRI study, trial-by-trial variation in the TEIs for the transformation to PMI and for the transformation to UMI were correlated, a result consistent with the idea that a common factor underlies both. There are at least two possible accounts for this pattern of results that will require future research to adjudicate. One is a parallel mechanism whereby a single signal is "split" so as to trigger the simultaneous output gating of one item into the PMI state and of the other item into the UMI state. A second is a serial process akin to biased competition (c.f., Desimone & Duncan, 1995) whereby a control signal first selects the cued item, and a consequence of this item's transformation to PMI is that it "pushes" the other item into the UMI state. Importantly, the correlation of TEIs reported here rules out what had been a third possibility, which was a "passive" account of the transformation to UMI

whereby the withdrawal of attention would allow the relaxation of the representation into a default state such that the relaxation process would not be influenced by the active PMI transformation. Along with the application of second-order context that is prompted by the prioritization cue, the time course of effective dimensionality of the RNN suggests that the resetting of second-order context partway through the trial may be a process that requires as much active control as does its initial application.

Chapter 4

General Discussion

Chapters 2 and 3 present two pieces of research using RNNs that simulate WM tasks to shed light on the neural mechanisms of WM prioritization on an algorithmic level. In Chapter 2, in a 2-back task, the RNNs revealed representational reversals between prioritized and unprioritized memory items in specific subspaces defined by priority. This differs from representational transformations in humans (as shown by EEG) in important ways. In Chapter 3, I trained RNNs to perform the DSR task, which is a classic task in investigations of WM prioritization. Unlike the 2-back which has a fixed and predictable trajectory of prioritization, DSR features flexible online control using unpredictable retrocues. The RNN simulations yielded the important insight that first-order context (e.g., spatial location or temporal order) plays a significant role in WM prioritization, and that context and priority are potentially represented via different mechanisms: context through the *segregation* of items into orthogonal subspaces and priority via the *separation* of representations within each subspace. Context and priority have been shown to differentially predict behavioral performance: larger variability of priority-based transformations, as a proxy for their efficacy, is associated with more/larger performance errors, but it is not the case with context-based transformations. Context and priority are shown to be differentially represented in the brain: context was represented in frontoparietal areas but not early visual cortex whereas priority was coded across the hierarchy along the dorsal visual stream.

The significance of context in WM prioritization

An abundance of previous neuroimaging research focuses on the difference in the representations of differentially prioritized memory information, either in terms of spatial distribution (e.g., Christophel et al., 2018) or representational format (e.g., Yu, Teng & Postle., 2020). The motivation for this line of inquiry makes intuitive sense as priority can be conceptualized as a level of context that is immediately relevant to behavior. However, as we see in Chapter 3, in WM tasks manipulating priority through explicit (retro-)cuing, priority would be meaningless without its reference to the first-order context. As shown in the RNN simulations, the stable representation of first-order context across the trial is crucial in accommodating the temporally fluctuating status of priority. Indeed, the prominence of contextual representation in frontoparietal areas is consistent with previous accounts (Cai et al., 2020; Fulvio et al., 2023; Gosseries et al., 2018; Panichello & Buschman, 2021).

In fact, context representations are essential to quite a few formulations of working memory. In some computational models (Burgess & Hitch, 1999; Lewandowsky & Farrell, 2008; Oberauer et al., 2012), working memory is implemented in neural networks that have distinct layers of units that separately represent the WM content (e.g., colors and letters) and context, which can serve as retrieval cues to access the

memoranda when necessary. During the encoding phase, representations of WM content are bound to their contexts via quick changes in connection strengths between the two layers of units. For example, in serial recall of lists, words are bound to their serial positions in the list as they get sequentially encoded, and at test, these positions are activated to cue the retrieval of the words in order. In visual object recall tasks, stimuli can be bound to various feature dimensions such as spatial location, luminance, color, orientation, which serve as contextual retrieval cues during recall of other features that are bound to the cued dimensions. Furthermore, it has been argued that maintaining context representations to individuate WM contents is metabolically efficient (Beukers et al., 2021).

As mentioned in the Introduction, in the embedded-component model, the region of direct access of WM is composed of the subset of activated LTM that is bound to a context, thus relating to each other. According to another influential model (Oberauer & Lin, 2017), memory items can be said to be meaningfully in working memory only when they are bound to the contexts that individuate them. Based on this formulation, the limited capacity of working memory partially originates from the confusability of memory contents that are bound to similar contexts. In fact, “swap errors” observed in working memory tasks can be attributed to the failure in such bindings (Schneegans & Bays, 2017a).

A remaining question is how context interacts with priority to generate successful behavior. How this is implemented in the RNN is fathomable given the two aforementioned representational mechanisms: it is likely that depending on the cue identity, representations will fall into one of two clusters (*separation*), each of which only allows information from a distinct subspace (*segregation*) to be read out. However, whether this is indeed the readout mechanism, and how the two contextual representations interact in the human brain is yet unclear.

How WM prioritization is controlled

Control processes are central to working memory and cognitive control has important parallels with motor control. As a concrete example, a cup of black coffee and a bowl of sugar afford two actions, and if one likes one's coffee sweet, one should prioritize actions related to preparing coffee over actions related to drinking it. Action selection has been shown to be mediated by fronto-striatal circuitry, by gating effects of two classes of medium spiny neurons of the striatum: Go cells and NoGo cells. Go cells, which express D1 receptors (and are part of the direct pathway), 'open the gate', and release the inhibition of the thalamus, which has an excitatory effect on the motor cortex, hence allowing an action to be executed; conversely, NoGo cells (part of the indirect pathway) express D2 receptors and act to 'close the gate', thus blocking the

thalamo-cortical influence, preventing the action to be executed (Figure 4.1). Many studies have found that such a mechanism in action selection also applies to the cognitive domain, and specifically it has been suggested that this mechanism of output gating implements selection of one among multiple items held in WM (Chatham et al., 2014; Chatham & Badre, 2015).

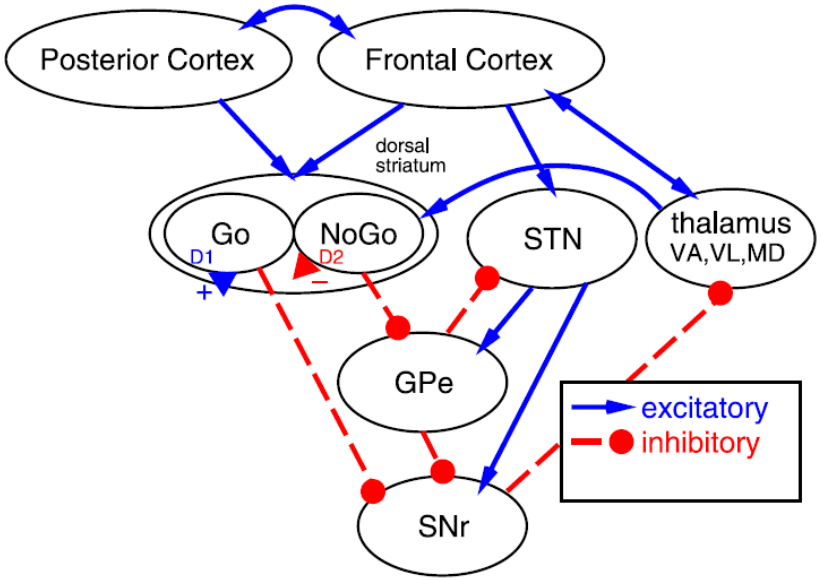


Figure 4.1. Neural basis of WM gating. The BG are interconnected with frontal cortex through a series of parallel loops as shown above. The thalamus is bidirectionally associated with frontal cortex through excitatory connections, and the SNr is tonically active and inhibiting this excitatory circuit. When direct pathway "Go" neurons in dorsal striatum fire, they inhibit the SNr, and thus disinhibit frontal cortex, generating a gating-like operation that permits PFC representations in or out of working memory. The indirect pathway "NoGo" cells of dorsal striatum have the opposite effect: they inhibit the GPe that inhibits SNr. The STN generates an extra dynamic background of inhibition (NoGo) by exciting the SNr. SNr, substantia nigra pars reticulata; GPe, globus pallidus external segment; STN, subthalamic nucleus. Figure and caption adapted from Hazy et al. (2006).

To examine the neural bases of WM output gating, Chatham and colleagues (2014) designed a WM task with three sequentially presented stimuli: two "item" stimuli and a

"context" stimulus that designates one of the two stimuli as relevant for responses. The context stimulus essentially triggers WM gates: when the context is presented first, it drives an *input* gate that selectively lets information into storage; when the context is presented last, however, it opens up an *output* gate that only permits the relevant information out of the storage to guide behavior. They found that during output gating, the dorsal pre-premotor cortex (pre-PMd) becomes more activated and its functional coupling with the caudate is increased, which is consistent with BG-mediated output gating models. Furthermore, brain-behavior correlations were also consistent with these models: (1) bilateral pre-PMd activations predicted mean response efficiency (indexed by mean RT) during output gating; (2) pre-PMd's bilateral functional coupling with the caudate predicted response variability, which aligns with a stochastic BG-mediated output gate.

Van Schouwenburg and colleagues (2010), using dynamic causal modelling (DCM), demonstrated that BG modulates the functional connectivity between prefrontal cortex and higher-level visual cortex during a task requiring attentional switching between stimuli of two categories: scenes and faces. In a follow-up study (van Schouwenburg et al., 2015), the same group showed that such attentional gating is selective: BG increased the functional connectivity between the prefrontal cortex and visual cortex

contralateral to the attended side, while inhibiting connection strength between PFC and visual cortex contralateral to the unattended side (Figure 4.2).

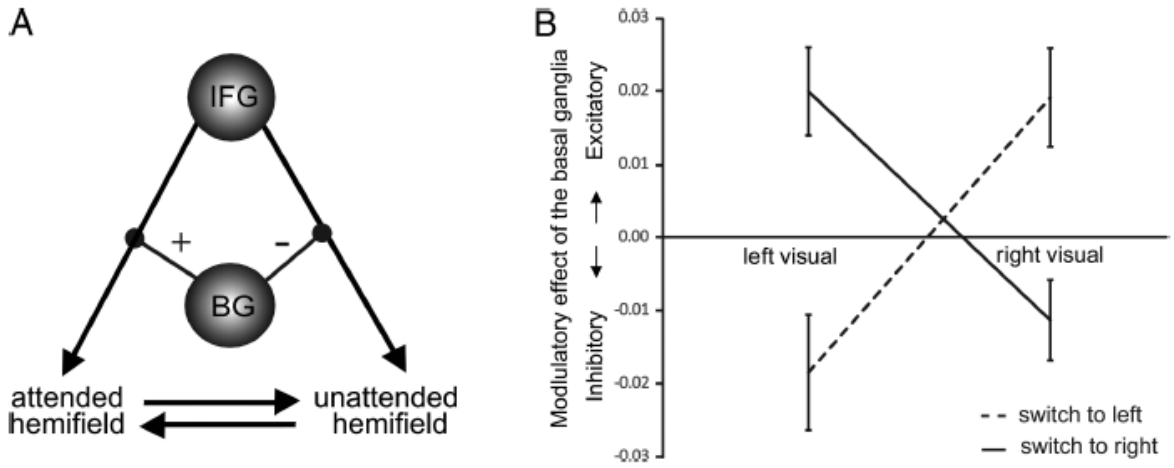


Figure 4.2. Results from van Schouwenberg (2015). (A) The average DCM model showed that BG both suppress previously attended visual information and enhance the newly attended visual information through modulation of top-down connections. (B) Consistent with this model, the BG inhibited the strength of the connection with the left visual cortex when subjects switched attention to the left visual hemifield, but enhanced connection strength with the left visual cortex when subjects switched attention to the right visual hemifield. The opposite pattern was demonstrated in the right visual cortex. Figure and caption adapted from van Schouwenburg et al., (2015).

Output gating, which involves selecting specific memorandum from WM, is strongly reminiscent of the operation of attentional prioritization in WM. Given the intertwined relationship between attention and WM (Oberauer, 2019), in parallel with attentional gating of perceptual information, cortico-striatally mediated gating serves as a potent candidate mechanism for WM memoranda to be differentially prioritized to guide behavior. As shown in both the EEG and fMRI datasets in Chapter 3, trial-by-trial transformational efficacy for PMI is closely correlated with that for UMI, indicative of a

common process underlying the transformations of both. As discussed at the end of Chapter 3, WM prioritization might be accounted for by a parallel mechanism where the signal from a single source, likely from the striatum, exerts differential gating effects on PMI and UMI, driving their transformations. Alternatively, via a serial process, striatally generated gating signal might output-gate one item into the PMI state, which consequently induces the other item into a UMI state by way of lateral inhibition. Future neuroimaging research and advanced analytic techniques are needed to adjudicate between these accounts.

How does output gating make use of representations of context and priority? To unite basal ganglia's role in WM control and motor control, Chatham and Badre (2015) proposes a framework that features a series of fronto-striatal loops that are nested rostro-caudally to support input and output gating functions of information of different levels of abstractions, with caudal loops gating information that is directly related to action. Based on the finding that context and priority representations are widely available throughout the fronto-parietal network presented in Chapter 3, it is fathomable that priority status (also represented in early sensory cortex, unlike context) is utilized to gate immediately action-relevant information involving caudal PFC and BG, with spatial/ordinal context, which is more abstracted from behavior,

implemented by rostral loops. Apparently, these speculative ideas await exploration in future research.

References

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. (2007). Fractionating the central executive. In A. Baddeley (Ed.), *Working Memory, Thought, and Action* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198528012.003.0007

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Baddeley, A. D., & Hitch, G. J. (2019). The phonological loop as a buffer store: An update. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *112*, 91–106. https://doi.org/10.1016/j.cortex.2018.05.015

Baddeley, A., & Hitch, G. (2007). Working memory: Past, present…and future? In N. Osaka, R. H. Logie, & M. D'Esposito (Eds.), *The Cognitive Neuroscience of Working Memory* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198570394.003.0001

Barak, O. (2017). Recurrent neural networks as versatile tools of neuroscience research. *Current Opinion in Neurobiology*, *46*, 1–6. https://doi.org/10.1016/j.conb.2017.06.003

Barak, O., & Tsodyks, M. (2014). Working models of working memory. *Current Opinion in Neurobiology*, *25*, 20–24. https://doi.org/10.1016/j.conb.2013.10.008

Beukers, A. O., Buschman, T. J., Cohen, J. D., & Norman, K. A. (2021). Is Activity Silent Working Memory Simply Episodic Memory? *Trends in Cognitive Sciences*, S136466132100005X. https://doi.org/10.1016/j.tics.2021.01.003

Braver, T. S., Gray, J. R., & Burgess, G. C. (2008). Explaining the Many Varieties of Working Memory Variation: Dual Mechanisms of Cognitive Control. In A. Conway, C. Jarrold, M. Kane, A. Miyake, & J. Towse (Eds.), *Variation in Working Memory* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195168648.003.0004

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, *106*(3), 551–581. https://doi.org/10.1037/0033-295X.106.3.551

Cai, Y., Fulvio, J. M., Yu, Q., Sheldon, A. D., & Postle, B. R. (2020). The Role of Location-Context Binding in Nonspatial Visual Working Memory. *eNeuro*, *7*(6). https://doi.org/10.1523/ENEURO.0430-20.2020

Chatham, C. H., & Badre, D. (2015). Multiple gates on working memory. *Current Opinion*

in Behavioral Sciences, *1*, 23–31. https://doi.org/10.1016/j.cobeha.2014.08.001

Chatham, C. H., Frank, M. J., & Badre, D. (2014). Corticostriatal Output Gating during Selection from Working Memory. *Neuron*, *81*(4), 930–942. https://doi.org/10.1016/j.neuron.2014.01.002

Christophel, T. B., Hebart, M. N., & Haynes, J.-D. (2012). Decoding the Contents of Visual Short-Term Memory from Human Visual and Parietal Cortex. *Journal of Neuroscience*, *32*(38), 12983–12989. https://doi.org/10.1523/JNEUROSCI.0184-12.2012

Christophel, T. B., Iamshchinina, P., Yan, C., Allefeld, C., & Haynes, J.-D. (2018). Cortical specialization for attended versus unattended working memory. *Nature Neuroscience*, *21*(4), 494–496. https://doi.org/10.1038/s41593-018-0094-4

Collette, F., Van der Linden, M., Laureys, S., Delfiore, G., Degueldre, C., Luxen, A., & Salmon, E. (2005). Exploring the unity and diversity of the neural substrates of executive functioning. *Human Brain Mapping*, *25*(4), 409–423. https://doi.org/10.1002/hbm.20118

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, *104*(2), 163–191. https://doi.org/10.1037/0033-2909.104.2.163

Cowan, N. (1995). *Attention and memory: An integrated framework* (pp. xv, 321). Oxford University Press.

Cowan, N. (1999). An Embedded-Processes Model of Working Memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (pp. 62–101). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.006

Cowan, N., Fristoe, N. M., Elliott, E. M., Brunner, R. P., & Saults, J. S. (2006). Scope of Attention, Control of Attention, and Intelligence in Children and Adults. *Memory & Cognition*, *34*(8), 1754–1768.

Cueva, C. J., Ardalan, A., Tsodyks, M., & Qian, N. (2021). *Recurrent neural network models for working memory of continuous variables: Activity manifolds, connectivity patterns, and dynamic codes* (arXiv:2111.01275). arXiv. http://arxiv.org/abs/2111.01275

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, *7*(9), 415–423. https://doi.org/10.1016/s1364-6613(03)00197-9

Del Giudice, M. (2021). Effective Dimensionality: A Tutorial. *Multivariate Behavioral Research*, *56*(3), 527–542. https://doi.org/10.1080/00273171.2020.1743631

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-

trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, *18*(1), 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

D'Esposito, M., & Postle, B. R. (2015). The Cognitive Neuroscience of Working Memory. *Annual Review of Psychology*, *66*(1), 115–142. https://doi.org/10.1146/annurev-psych-010814-015031

Devinsky, O., M.D, D'Esposito, M., & M.D. (2003). *Neurology of Cognitive and Behavioral Disorders*. Oxford University Press.

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, *2*(11), 820–829. https://doi.org/10.1038/35097575

Emrich, S. M., Riggall, A. C., LaRocque, J. J., & Postle, B. R. (2013). Distributed Patterns of Activity in Sensory Cortex Reflect the Precision of Multiple Items Maintained in Visual Short-Term Memory. *Journal of Neuroscience*, *33*(15), 6516–6523. https://doi.org/10.1523/JNEUROSCI.5732-12.2013

Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity, not quality: The relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review*, *17*(5), 673–679. https://doi.org/10.3758/17.5.673

Fulvio, J. M., & Postle, B. R. (2020). Cognitive Control, Not Time, Determines the Status of Items in Working Memory. *Journal of Cognition*, *3*(1), 8. https://doi.org/10.5334/joc.98

Fulvio, J. M., Yu, Q., & Postle, B. R. (2023). Strategic control of location and ordinal context in visual working memory. *Cerebral Cortex*, *33*(13), 8821–8834. https://doi.org/10.1093/cercor/bhad164

Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science (New York, N.Y.)*, *173*(3997), 652–654. https://doi.org/10.1126/science.173.3997.652

Gathercole, S. E., & Pickering, S. J. (2000). Working memory deficits in children with low achievements in the national curriculum at 7 years of age. *The British Journal of Educational Psychology*, *70 (Pt 2)*, 177–194. https://doi.org/10.1348/000709900158047

Gold, J. M., Barch, D. M., Feuerstahler, L. M., Carter, C. S., MacDonald, A. W., Ragland, J. D., Silverstein, S. M., Strauss, M. E., & Luck, S. J. (2019). Working Memory Impairment Across Psychotic disorders. *Schizophrenia Bulletin*, *45*(4), 804–812. https://doi.org/10.1093/schbul/sby134

Gong, M., & Liu, T. (2020). Biased Neural Representation of Feature-Based Attention in the Human Frontoparietal Network. *The Journal of Neuroscience*, *40*(43), 8386–8395. https://doi.org/10.1523/JNEUROSCI.0690-20.2020

Gosseries, O., Yu, Q., LaRocque, J. J., Starrett, M. J., Rose, N. S., Cowan, N., & Postle, B. R. (2018). Parietal-Occipital Interactions Underlying Control- and Representation-Related Processes in Working Memory for Nonspatial Visual Features. *The Journal of Neuroscience*, *38*(18), 4357–4366. https://doi.org/10.1523/JNEUROSCI.2747-17.2018

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947

Han, X., Berg, A. C., Oh, H., Samaras, D., & Leung, H.-C. (2013). Multi-voxel pattern analysis of selective representation of visual working memory in ventral temporal and occipital regions. *NeuroImage*, *73*, 8–15. https://doi.org/10.1016/j.neuroimage.2013.01.055

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), Article 7238. https://doi.org/10.1038/nature07832

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2006). Banishing the homunculus: Making working memory work. *Neuroscience*, *139*(1), 105–118. https://doi.org/10.1016/j.neuroscience.2005.04.067

Hinton, G., & Sejnowski, T. (1983). *Optimal perceptual inference*. 448–453.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558. https://doi.org/10.1073/pnas.79.8.2554

Kell, A. J., & McDermott, J. H. (2019). Deep neural network models of sensory systems: Windows onto the role of task constraints. *Current Opinion in Neurobiology*, *55*, 121–132. https://doi.org/10.1016/j.conb.2019.02.003

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [Cs]*. http://arxiv.org/abs/1412.6980

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C. E., Kepecs, A., Mainen, Z. F., Qi, X.-L., Romo, R., Uchida, N., & Machens, C. K. (2016). Demixed principal component analysis of neural population data. *eLife*, *5*, e10989. https://doi.org/10.7554/eLife.10989

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science (New York, N.Y.)*, *302*(5648), 1181–1185.

https://doi.org/10.1126/science.1088545

Koenigs, M., Barbey, A. K., Postle, B. R., & Grafman, J. (2009). Superior Parietal Cortex Is Critical for the Manipulation of Information in Working Memory. *The Journal of Neuroscience*, *29*(47), 14980–14986. https://doi.org/10.1523/JNEUROSCI.3706-09.2009

Kruijne, W., Bohte, S. M., Roelfsema, P. R., & Olivers, C. N. L. (2020). Flexible Working Memory Through Selective Gating and Attentional Tagging. *Neural Computation*, *33*(1), 1–40. https://doi.org/10.1162/neco_a_01339

Kubota, K., & Niki, H. (1971). Prefrontal cortical unit activity and delayed alternation performance in monkeys. *Journal of Neurophysiology*, *34*(3), 337–347. https://doi.org/10.1152/jn.1971.34.3.337

LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2013). Decoding Attended Information in Short-term Memory: An EEG Study. *Journal of Cognitive Neuroscience*, *25*(1), 127–142. https://doi.org/10.1162/jocn_a_00305

Larocque, J. J., Lewis-Peacock, J. A., & Postle, B. R. (2014). Multiple neural states of representation in short-term memory? It's a matter of attention. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00005

LaRocque, J. J., Riggall, A. C., Emrich, S. M., & Postle, B. R. (2017). Within-Category Decoding of Information in Different Attentional States in Short-Term Memory. *Cerebral Cortex*, *27*(10), 4881–4890. https://doi.org/10.1093/cercor/bhw283

Lee, S.-H., Kravitz, D. J., & Baker, C. I. (2013). Goal-dependent dissociation of visual and prefrontal cortices during working memory. *Nature Neuroscience*, *16*(8), Article 8. https://doi.org/10.1038/nn.3452

Lewandowsky, S., & Farrell, S. (2008). Short-Term Memory: New Data and a Model. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 49, pp. 1–48). Academic Press. https://doi.org/10.1016/S0079-7421(08)00001-7

Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2011a). Neural Evidence for a Distinction between Short-term Memory and the Focus of Attention. *Journal of Cognitive Neuroscience*, *24*(1), 61–79. https://doi.org/10.1162/jocn_a_00140

Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2011b). Neural Evidence for a Distinction between Short-term Memory and the Focus of Attention. *Journal of Cognitive Neuroscience*, *24*(1), 61–79. https://doi.org/10.1162/jocn_a_00140

Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2012). Neural Evidence for a Distinction between Short-term Memory and the Focus of Attention. *Journal of Cognitive Neuroscience*, *24*(1), 61–79. https://doi.org/10.1162/jocn_a_00140

Libby, A., & Buschman, T. J. (2021). Rotational dynamics reduce interference between sensory and memory representations. *Nature Neuroscience*, 1–12.

https://doi.org/10.1038/s41593-021-00821-9

Logie, R. H. (2003). Spatial and Visual Working Memory: A Mental Workspace. In *Psychology of Learning and Motivation* (Vol. 42, pp. 37–78). Academic Press. https://doi.org/10.1016/S0079-7421(03)01002-8

Lorenc, E. S., Vandenbroucke, A. R. E., Nee, D. E., de Lange, F. P., & D'Esposito, M. (2020). Dissociable neural mechanisms underlie currently-relevant, future-relevant, and discarded working memory representations. *Scientific Reports*, *10*(1), 11195. https://doi.org/10.1038/s41598-020-67634-x

Manohar, S. G., Zokaei, N., Fallon, S. J., Vogels, T. P., & Husain, M. (2019). Neural mechanisms of attending to items in working memory. *Neuroscience & Biobehavioral Reviews*, *101*, 1–12. https://doi.org/10.1016/j.neubiorev.2019.03.017

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), Article 7474. https://doi.org/10.1038/nature12742

Masse, N. Y., Yang, G. R., Song, H. F., Wang, X.-J., & Freedman, D. J. (2019). Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature Neuroscience*, *22*(7), 1159. https://doi.org/10.1038/s41593-019-0414-3

Merrikhi, Y., Clark, K., Albarran, E., Parsa, M., Zirnsak, M., Moore, T., & Noudoost, B. (2017). Spatial working memory alters the efficacy of input to visual cortex. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/ncomms15041

Mutanen, T. P., Biabani, M., Sarvas, J., Ilmoniemi, R. J., & Rogasch, N. C. (2020). Source-based artifact-rejection techniques available in TESA, an open-source TMS–EEG toolbox. *Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation*, *13*(5), 1349–1351. https://doi.org/10.1016/j.brs.2020.06.079

Myers, N. E., Stokes, M. G., & Nobre, A. C. (2017). Prioritizing Information during Working Memory: Beyond Sustained Internal Attention. *Trends in Cognitive Sciences*, *21*(6), 449–461. https://doi.org/10.1016/j.tics.2017.03.010

Nelissen, N., Stokes, M., Nobre, A. C., & Rushworth, M. F. S. (2013). Frontal and Parietal Cortical Interactions with Distributed Visual Representations during Selective Attention and Action Selection. *Journal of Neuroscience*, *33*(42), 16443–16458. https://doi.org/10.1523/JNEUROSCI.2625-13.2013

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *28*(3), 411–421.

Oberauer, K. (2005). Control of the contents of working memory—A comparison of two paradigms and two age groups. *Journal of Experimental Psychology. Learning,*

Memory, and Cognition, *31*(4), 714–728. https://doi.org/10.1037/0278-7393.31.4.714

Oberauer, K. (2013). The focus of attention in working memory—From metaphors to mechanisms. *Frontiers in Human Neuroscience*, *7*. https://www.frontiersin.org/articles/10.3389/fnhum.2013.00673

Oberauer, K. (2019). Working Memory and Attention – A Conceptual Analysis and Review. *Journal of Cognition*, *2*(1), 36. https://doi.org/10.5334/joc.58

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin & Review*, *19*(5), 779–819. https://doi.org/10.3758/s13423-012-0272-4

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*(1), 21–59. https://doi.org/10.1037/rev0000044

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2010). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience*, *2011*, e156869. https://doi.org/10.1155/2011/156869

O'Reilly, R. C., & Frank, M. J. (2006). Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia. *Neural Computation*, *18*(2), 283–328. https://doi.org/10.1162/089976606775093909

Panichello, M. F., & Buschman, T. J. (2021). Shared mechanisms underlie the control of working memory and attention. *Nature*. https://doi.org/10.1038/s41586-021-03390-w

Parthasarathy, A., Herikstad, R., Bong, J. H., Medina, F. S., Libedinsky, C., & Yen, S.-C. (2017). Mixed selectivity morphs population codes in prefrontal cortex. *Nature Neuroscience*, *20*(12), 1770. https://doi.org/10.1038/s41593-017-0003-2

Postle, B. R. (2006). Working Memory as an Emergent Property of the Mind and Brain. *Neuroscience*, *139*(1), 23–38. https://doi.org/10.1016/j.neuroscience.2005.06.005

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., Berker, A. de, Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., … Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), Article 11. https://doi.org/10.1038/s41593-019-0520-2

Riggall, A. C., & Postle, B. R. (2012). The Relationship between Working Memory Storage and Elevated Activity as Measured with Functional Magnetic Resonance Imaging. *Journal of Neuroscience*, *32*(38), 12990–12998. https://doi.org/10.1523/JNEUROSCI.1892-12.2012

Rogasch, N. C., Sullivan, C., Thomson, R. H., Rose, N. S., Bailey, N. W., Fitzgerald, P. B.,

Farzan, F., & Hernandez-Pavon, J. C. (2017). Analysing concurrent transcranial magnetic stimulation and electroencephalographic data: A review and introduction to the open-source TESA software. *NeuroImage*, *147*, 934–951. https://doi.org/10.1016/j.neuroimage.2016.10.031

Romo, R., Brody, C. D., Hernández, A., & Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature*, *399*(6735), Article 6735. https://doi.org/10.1038/20939

Rose, N. S., LaRocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., & Postle, B. R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. *Science*, *354*(6316), 1136–1139. https://doi.org/10.1126/science.aah7011

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2013, December 20). *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks*. arXiv.Org. https://arxiv.org/abs/1312.6120v3

Schneegans, S., & Bays, P. M. (2017a). Neural Architecture for Feature Binding in Visual Working Memory. *The Journal of Neuroscience*, *37*(14), 3913. https://doi.org/10.1523/JNEUROSCI.3493-16.2017

Schneegans, S., & Bays, P. M. (2017b). Restoration of fMRI Decodability Does Not Imply Latent Working Memory States. *Journal of Cognitive Neuroscience*, *29*(12), 1977–1994. https://doi.org/10.1162/jocn_a_01180

Schouwenburg, M. R. van, Ouden, H. E. M. den, & Cools, R. (2010). The Human Basal Ganglia Modulate Frontal-Posterior Connectivity during Attention Shifting. *Journal of Neuroscience*, *30*(29), 9910–9918. https://doi.org/10.1523/JNEUROSCI.1111-10.2010

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science*, *20*(2), 207–214. https://doi.org/10.1111/j.1467-9280.2009.02276.x

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *298*(1089), 199–209. https://doi.org/10.1098/rstb.1982.0082

Shipstead, Z., & Engle, R. W. (2013). Interference within the focus of attention: Working memory tasks reflect more than temporary maintenance. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(1), 277–289. https://doi.org/10.1037/a0028467

Shipstead, Z., Redick, T. S., Hicks, K. L., & Engle, R. W. (2012). The scope and control of attention as separate aspects of working memory. *Memory (Hove, England)*, *20*(6), 608–628. https://doi.org/10.1080/09658211.2012.691519

Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring Latent Visual Working

Memory Representations in Human Cortex. *Neuron*, *91*(3), 694–707. https://doi.org/10.1016/j.neuron.2016.07.006

Sreenivasan, K. K., Vytlacil, J., & D'Esposito, M. (2014). Distributed and Dynamic Storage of Working Memory Stimulus Information in Extrastriate Cortex. *Journal of Cognitive Neuroscience*, *26*(5), 1141–1153. https://doi.org/10.1162/jocn_a_00556

Stokes, M. G. (2015). 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, *19*(7), 394–405. https://doi.org/10.1016/j.tics.2015.05.004

Stokes, M. G., Muhle-Karbe, P. S., & Myers, N. E. (2020). Theoretical distinction between functional states in working memory and their corresponding neural states. *Visual Cognition*, *28*(5–8), 420–432. https://doi.org/10.1080/13506285.2020.1825141

Sussillo, D. (2014). Neural circuits as computational dynamical systems. *Current Opinion in Neurobiology*, *25*, 156–163. https://doi.org/10.1016/j.conb.2014.01.008

Sussillo, D., Churchland, M. M., Kaufman, M. T., & Shenoy, K. V. (2015). A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience*, *18*(7), Article 7. https://doi.org/10.1038/nn.4042

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *arXiv:1409.3215 [Cs]*. http://arxiv.org/abs/1409.3215

Teich, A. F., & Qian, N. (2003). Learning and adaptation in a recurrent model of V1 orientation selectivity. *Journal of Neurophysiology*, *89*(4), 2086–2100. https://doi.org/10.1152/jn.00970.2002

Teng, C., & Postle, B.R. (2020). Investigating the roles of visual and parietal cortex in representing content versus context in visual working memory. Stage 1-accepted Registered Report; *eNeuro*.

van Loon, A. M., Olmos-Solis, K., Fahrenfort, J. J., & Olivers, C. N. (2018). Current and future goals are represented in opposite patterns in object-selective cortex. *eLife*, *7*, e38677. https://doi.org/10.7554/eLife.38677

van Schouwenburg, M. R., den Ouden, H. E. M., & Cools, R. (2015). Selective Attentional Enhancement and Inhibition of Fronto-Posterior Connectivity by the Basal Ganglia During Attention Switching. *Cerebral Cortex*, *25*(6), 1527–1534. https://doi.org/10.1093/cercor/bht345

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, *438*(7067), 500–503. https://doi.org/10.1038/nature04171

Wallis, J. D., Anderson, K. C., & Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature*, *411*(6840), Article 6840.

https://doi.org/10.1038/35082081

Wan, Q., Cai, Y., Samaha, J., & Postle, B. R. (2020). Tracking stimulus representation across a 2-back visual working memory task. *Royal Society Open Science*, *7*(8), 190228. https://doi.org/10.1098/rsos.190228

Wan, Q., Menendez, J. A., & Postle, B. R. (2022). Priority-based transformations of stimulus representation in visual working memory. *PLOS Computational Biology*, *18*(6), e1009062. https://doi.org/10.1371/journal.pcbi.1009062

Wang, L., Mruczek, R. E. B., Arcaro, M. J., & Kastner, S. (2015). Probabilistic Maps of Visual Topography in Human Cortex. *Cerebral Cortex*, *25*(10), 3911–3931. https://doi.org/10.1093/cercor/bhu277

Warden, M. R., & Miller, E. K. (2010). Task-Dependent Changes in Short-Term Memory in the Prefrontal Cortex. *Journal of Neuroscience*, *30*(47), 15801–15810. https://doi.org/10.1523/JNEUROSCI.1569-10.2010

Wilson, H. R., & Cowan, J. D. (1972). Excitatory and Inhibitory Interactions in Localized Populations of Model Neurons. *Biophysical Journal*, *12*(1), 1–24. https://doi.org/10.1016/S0006-3495(72)86068-5

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X.-J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, *22*(2), 297–306. https://doi.org/10.1038/s41593-018-0310-2

Yang, G. R., & Molano-Mazón, M. (2021). Towards the next generation of recurrent network models for cognitive neuroscience. *Current Opinion in Neurobiology*, *70*, 182–192. https://doi.org/10.1016/j.conb.2021.10.015

Yu, Q., Teng, C., & Postle, B. R. (2020). Different states of priority recruit different neural representations in visual working memory. *PLOS Biology*, *18*(6), e3000769. https://doi.org/10.1371/journal.pbio.3000769

Appendices

Appendix 1

Chapter 2 Supplementary Materials

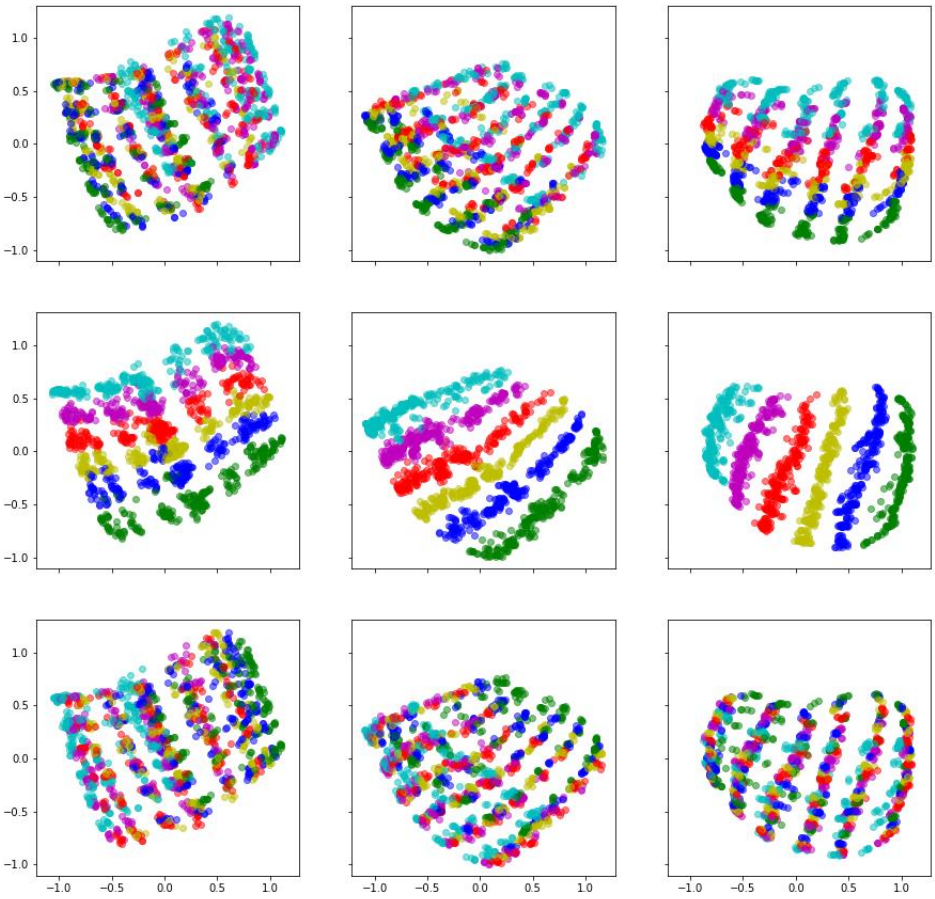


Figure S2.1. Example 7-hidden-unit RNN trained with input following the basis function used to build IEMs in Wan et al. (2020). Shown is the 2D visualization of the LSTM hidden layer activity of this RNN. The network architecture and training procedure are identical to the 7D RNNs reported in the main text with the exception that the inputs are not one-hot vectors; instead, they are specified by the IEM basis function: $R = \sin^6(x)$ (e.g., for stimulus #3, input vector is [0.0156, 0.4219, 1, 0.4219, 0.0156, 0]). Note that these results are qualitatively similar to RNNs reported in the main text (Figure 2.4).

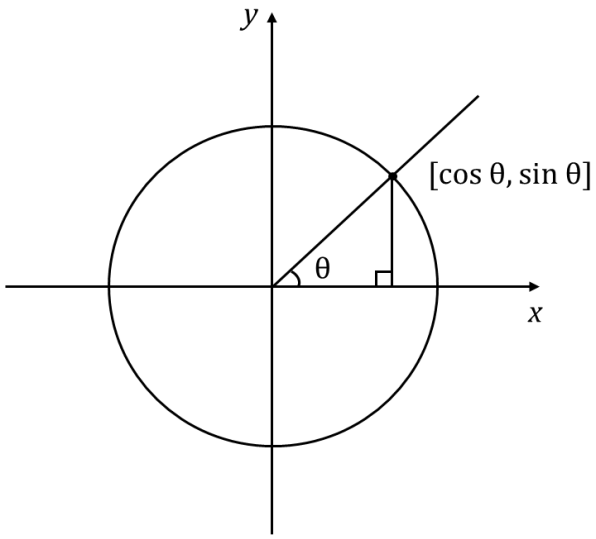


Figure S2.2. Generating circular input for 60-hidden-unit RNNs. Each point on the circle can be characterized by an angle relative to the easternmost point of the circle. The coordinates of these points within the 2D space on which this circle lives are given by [cos θ, sin θ]. To construct input vectors used in our RNN model, we mapped each stimulus orientation θ to the corresponding point on the circle at 2 * θ. The multiplication by 2 is necessary to match the periodicity of the input vectors to the periodicity of the oriented grating stimuli, which have a period of 180° (i.e., the stimulus at θ is equivalent to the stimulus at θ + 180°).

7D RNN

	subspace	D1	D2	D3	D4	D5
Stimulus PEV	**UMI**	93.10%	97.41%	98.08%	98.19%	98.21%
	PMI	97.33%	99.77%	99.95%	99.96%	99.96%
Global PEV	**UMI**	93.10%	97.41%	98.08%	98.19%	98.21%
	PMI	97.33%	99.77%	99.95%	99.96%	99.96%

60D RNN

	subspace	D1	D2	D3	D4	D5
Stimulus PEV	**UMI**	69.10%	98.75%	99.49%	99.74%	99.86%
	PMI	71.30%	98.95%	99.55%	99.81%	99.96%
Global PEV	**UMI**	69.10%	98.75%	99.49%	99.74%	99.86%
	PMI	71.30%	98.95%	99.55%	99.81%	99.96%

EEG

	subspace	D1	D2	D3	D4	D5
Stimulus PEV	**UMI**	45.41%	69.06%	83.54%	93.37%	99.97%
	PMI	46.18%	69.39%	83.89%	93.61%	99.97%
Global PEV	**UMI**	19.61%	29.79%	36.10%	40.40%	43.32%
	PMI	18.38%	27.62%	33.51%	37.54%	40.21%

Figure S2.3. Cumulative percent variance explained (PEV) by top dPCs of the UMI and PMI subspaces for 7D RNN, 60D RNN and EEG data. The percentages of both stimulus and global variance explained are shown.

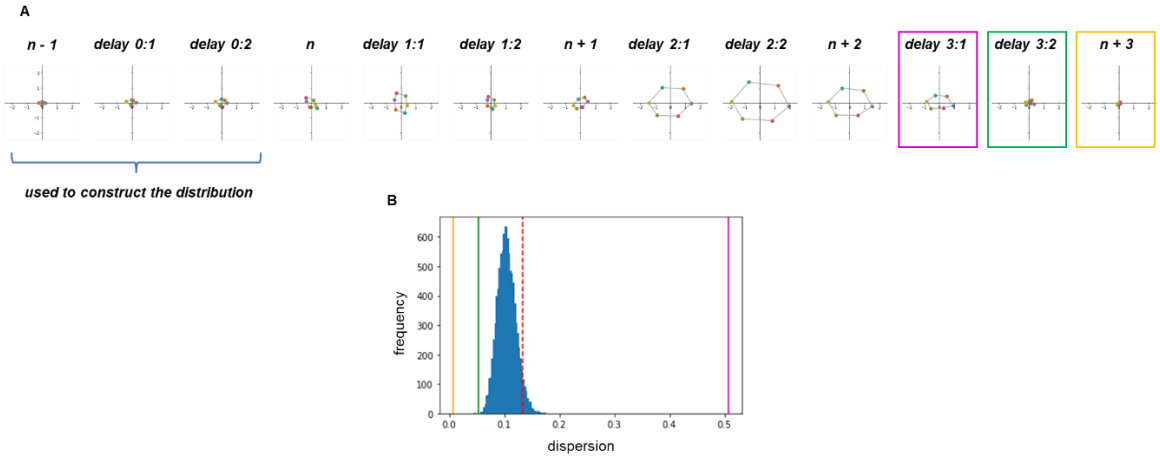


Figure S2.4. Empirical test for presence of stimulus information in WM. (A) Time course of stimulus averages projected into the PMI subspace from an example 60D RNN. Data points are colored based on item *n*'s identity. We used the 3 timesteps prior to the presentation of *n* to construct a baseline distribution of dispersion values using a bootstrapping procedure. Visually, one can see stimulus information collapsing in the PMI subspace across the three timesteps that follow timestep *n + 2*, colored squares added to identify them for panel B. (B) The baseline distribution of dispersion values, with red dashed line indicating the 95th percentile criterion. Magenta, green and orange lines indicate the dispersion values from timesteps *delay 3:1*, *delay 3:2*, and *n + 3,* respectively.

Appendix 2

Chapter 3 Supplementary Materials

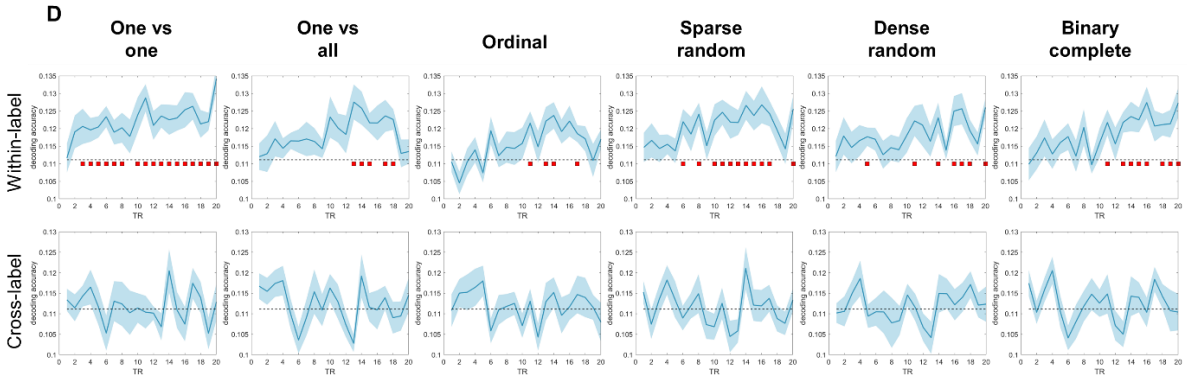


Figure S3.1. Comparisons of within- and cross-label decoding from the fMRI dataset across various SVM coding designs in MATLAB. (A) Context-based decoding for V1-2. (B) Context-based decoding for FEF. (C) Priority-based decoding for V1-2. (D) Priority-based decoding for FEF. In each graph, the blue shading around each curve shows standard error of the mean. The horizontal dashed line indicates the chance-level decoding accuracy of 0.11. Red squares below the dashed line indicate time points with significant above-chance decoding accuracy ($p < .05$, FDR-corrected across all time points).