PREFRONTAL MECHANISMS OF TRAINING IN WORKING MEMORY

BY

RYE G. JAFFE

A Dissertation Submitted to the Graduate Faculty of

WAKE FOREST UNIVERSITY GRADUATE SCHOOL OF ARTS AND SCIENCES

In Partial Fulfilment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Neuroscience Program

May 2026

Winston-Salem, North Carolina

Approved By:

Christos Constantinidis, Ph.D., Advisor

Robert Hampson, Ph.D., Chair

Terrence Stanford, Ph.D.

Emilio Salinas, Ph.D.

Benjamin Rowland, Ph.D.

ACKOWLEDGEMENTS

The author would like to acknowledge my advisor and mentor Christos Constantinidis for his support and guidance, as well as current and previous lab members Wenhao Dang, Chrissy Suell, Junda Zhu, Tony Gao, Hua Tang, Zhengyang Wang, Xue-Lian Qi, Peiyu Chen, Balbir Singh, Rana Mozumder, Sihai Li, Shusen Pu, Skylar Lambert, Daniel Woods, Kayla Yetman, Katelyn Clemencich, Kris Clifft, Anna Machado, and Gracie Hilber with assistance in experiments and data analysis. The author would also like to acknowledge Robert Hampson, Emilio Salinas, Benjamin Rowland, and Terrence Stanford for their support and guidance. Finally, the author would also like to acknowledge and thank family members Helen, David, Spencer, Laurel and Rachel Jaffe, and Ali, Julian and Ruby Ramis for their support and love. This dissertation is dedicated to the improvement of the human condition.

Table of Contents

List of Illustrations and Tables.....4

List of Abbreviations.....7

Abstract 8

Chapter 1: Introduction.....10

Chapter 2: Emergence of Nonlinear Mixed Selectivity in Prefrontal Cortex after Training.....127

Chapter 3: Prefrontal Mechanisms of Rule Learning.....190

Chapter 4: General Discussion.....267

References.....287

Curriculum Vitae.....369

List of Illustrations and Tables

Chapter 1:

1. Schematic diagrams of different models of working memory.....109
2. Schematic illustration of bump attractor via single neurons and populations.....111
3. Schematic illustration of types of dynamic coding.....112
4. Schematic demonstrations of activity silent mechanisms.....113
5. Schematic illustration of the division of labor model.....115
6. Stimulus selectivity along the dorso-ventral and anterior-posterior axes of the PFC before and after training.....116
7. Local organization of the PFC.....117
8. Schematic illustration of a cross-modal WM task.....118
9. Properties of working memory capacity in human subjects.....119
10. Neural activity in capacity tasks.....120
11. Drift and decay processes in the bump attractor model.....121
12. Patterns of activation of mouse working memory as well as the behavioral effects of inactivating mPFC and MD terminals.....123
13. Behavioral improvement and PFC activity changes in working memory tasks across adolescent development.....124
14. Effects of training on the activity of different PFC areas.....125

Chapter 2

1. Task structure and stimuli used for Chapter 2 experiments.....172

2. Exemplar neural responses from the spatial task for classical, linear mixed, and nonlinear mixed selectivity 173
3. Proportions of selective neurons before and after training.....174
4. F-scores and dimensionality of population responses before and after training.....177
5. Training effects on the cell selectivity of different PFC areas.....179
6. Cell selectivity in tasks with different levels of complexity.....180
7. Cell selectivity in tasks with the same sensory input but different behavioral requirements.....182
8. The relative contribution of linearly versus nonlinearly selective cells to encoding new task information.....184
9. Coding dynamics of linearly versus nonlinearly selective cells.....186
10. Comparison of F-scores and dimensionality across correct and error trials.....187
11. The role of nonlinear mixed selectivity in saccade direction decoding from correct versus error trials.....188

Chapter 3

1. Tasks and behavioral analyses used for Chapter 3 experiments.....246
2. Nonlinear training effects on firing rate.....248
3. The effects of training on unexplained variance.....249
4. The effects of training on decoding different task factors.....251
5. Dynamics of populations responses in PFC.....253

6. Multi-phase spatial match-nonmatch training paradigm.....255

7. Expanded tasks and behavioral analyses.....257

8. Chronic recording setup and neural activity.....259

9. Expanded effects of training on unexplained variance.....261

10. Partial correlation of task variables with firing rate across training.....262

11. Expanded effects of training on decoding different task factors.....263

12. Expanded dynamics of populations responses in PFC.....264

List of Abbreviations

ANOVA.....Analysis of Variance

CS.....Classical Selectivity

EEG.....Electroencephalography

fMRI.....functional Magnetic Resonance Imaging

IT.....Inferior Temporal

LMS.....Linear Mixed Selectivity

LTM.....Long Term Memory

MD.....Mediodorsal Nucleus

NMS.....Nonlinear Mixed Selectivity

ODR.....Oculomotor Delayed Response

PCA.....Principal Component Analysis

PFC.....Prefrontal Cortex

PPC.....Posterior Parietal Cortex

RL.....Rule Learning

SVM.....Support Vector Machine

TDR.....Targeted Dimensionality Reduction

WM.....Working Memory

ABSTRACT

Rye G. Jaffe

PREFRONTAL MECHANISMS OF TRAINING IN WORKING MEMORY

Dissertation under the direction of

Christos Constantinidis, Ph.D., Professor of Neurobiology and Anatomy

Working memory is the ability to maintain and manipulate information in the conscious mind over a timescale of seconds. This is considered to be centered on the prefrontal cortex—as evidenced by neurophysiological recordings in non-human primates—and further studies have demonstrated that training in related tasks may cause a variety of changes in prefrontal activity. For example, neurons in the prefrontal cortex are typically activated by different stimuli and abstract variables. A single neuron's selectivity for a given stimulus dimension often changes depending on its context in a task, and this phenomenon is known as nonlinear mixed selectivity, which has previously been hypothesized to emerge as a result of training to perform tasks in different contexts. We therefore tested this hypothesis directly by examining the neuronal responses of different prefrontal areas before and after male monkeys had been trained to perform different working memory tasks involving visual stimulus locations and/or shapes. This revealed a modest increase in nonlinear mixed selectivity over the course of training in spatial—but not

shape—working memory tasks. Moreover, to explore more deeply into the field of prefrontal training effects, we also investigated how prefrontal activity predicts rule learning across parallel spatial and object memory tasks. There is a significant gap regarding how related mechanisms may reflect behavioral improvements in different contexts, especially when examining across different tasks and modalities. This was confronted by recording single units from chronic electrode arrays implanted in the prefrontal cortex of four monkeys as they were trained to perform parallel spatial and object memory tasks with the goal of assessing the resulting activity changes that would be induced by rule learning. Progression of training thus allowed behavioral improvements to be correlated with a variety of neural effects. Some of these were unique for spatial and object learning tasks, including firing rate changes decoding information about match and nonmatch stimuli. Others, however, were shared between tasks, including a decrease in the variance explained by the reward and separation of task variables in principal component space as training progressed. Such generalizable mechanisms thus reveal that the object-spatial divide is not necessarily absolute. Altogether, our studies present both novel expansions and limitations on how changes in prefrontal activity may underlie training in working memory, highlighting specific mechanisms that may be targeted for the treatment of related impairments, the enhancement of healthy populations, and an improved understanding of human cognition.

Chapter 1: Introduction— Working Memory: From Neural Activity to the Sentient Mind

The following chapter was adapted from a published article: Jaffe, R. J. and C. Constantinidis (2021). "Working Memory: From Neural Activity to the Sentient Mind." Compr Physiol 11(4): 2547–2587. Any difference in stylistic variations is due to the requirement of the journal.

Introduction

Working memory (WM)—the ability to maintain and manipulate information in conscious mind over a timescale of seconds—is a critical cognitive function in the ability to learn, make decisions, and function in daily life (Baddeley 1992, Baddeley 2012). This ability can be disrupted after brain injury, most importantly in the prefrontal cortex (PFC) (Morris and Baddeley 1988, Curtis and D'Esposito 2004, Martinussen, Hayden et al. 2005, Rossi, Bichot et al. 2007, Forbes, Carrick et al. 2009). Neural correlates of WM have been identified in the activity of neurons in cortical and subcortical areas. As such, working memory represents a prototypical case of a mental phenomenon that can be explained in terms of underlying brain activity. The precise mechanisms that underlie WM remain subject to debate (Suchow, Fougnie et al. 2014). In this article, we review the current state of knowledge and open questions, examining prior models and proposing a road forward.

We begin by reviewing conceptual models of WM as general frameworks that can be used to bind experimental findings into a unified theory. This discussion will also demonstrate how some of the controversies surrounding the field can be accounted for by the different methodologies and levels of analysis that different studies have undertaken. We proceed by reviewing the alternative accounts of the neural basis of WM, including those based on persistent activity generated by cortical neurons, those based on rhythmic discharges and those that do not depend on discharge rates modulation (activity-silent models). The debate between these

accounts also provides an opportunity to examine the localization of WM in the brain. Traditional models, such as the original bump attractor model, have placed the PFC as the seat of WM, though alternative accounts suggest that the sensory cortex is the primary site of WM, with the PFC playing a supervisory role instead.

Although the emphasis of the article is on visual WM, neural correlates of other modalities have also been described and their review is instructive to defining the underlying mechanisms of WM. Human WM is notoriously limited in terms of its capacity and duration; we examine the neural basis of these limitations, and the competing models that have been proposed to account for them. We then discuss neural correlates of WM in non-primates (rodent and avian species) and what they reveal about the evolution of working memory maintenance. We move on to examine the enhancement of working memory abilities through the course of childhood development and through training, in adulthood. We also discuss individual differences between people, the relationship between working memory and other constructs such as attention and intelligence, and disorders of working memory. The review ends with conclusions and open questions for future investigation.

Conceptual Models of Working Memory

The definition of working memory has undergone a series of revisions since its introduction to the scientific vernacular in the 1960's (Miller 1960), better relating this mental phenomenon to well-defined neuroscience concepts in order to alleviate the indeterminacy that often relates to philosophical constructs (Buzsaki 2020, Poeppel and Adolfi 2020). WM is therefore conventionally defined at present as a fundamental cognitive system that facilitates the temporary storage and manipulation of information in the immediately conscious mind (Awh, Vogel et al. 2006, Fukuda, Awh et al. 2010, Baddeley 2012, Oberauer 2019). This is a critical cognitive function in the ability to learn, make decisions, and function in daily life, and its underlying mechanisms are often examined through the life disruptions that are suffered in the case of brain injuries or conditions such as schizophrenia and ADHD (Morris and Baddeley 1988, Curtis and D'Esposito 2004, Martinussen, Hayden et al. 2005, Rossi, Bichot et al. 2007, Forbes, Carrick et al. 2009).

The classic Atkinson–Shiffrin multi-store model (Fig. 1A), also known as the modal model of memory, distinguished between three memory stores: a sensory register, a short-term store, and a long-term memory store (Malmberg, Raaijmakers et al. 2019). The sensory register buffers information from sensory modalities, and in the case of vision ("iconic" memory) it has very rapid decay, in the order of half a second or less. The duration of the sensory register for auditory information ("echoic" memory) is longer in the order of 5 seconds or so. As we will discuss in the later sections regarding the neural basis of WM, certain aspects of this model have

a strong empirical grounding in current neuroscience research. The activation of neurons in the visual cortex persists for several hundred milliseconds after the offset of a stimulus, thus providing a neural correlate for the sensory register (Zaksas and Pasternak 2006). The PFC and its connected areas then maintain specific patterns of activity throughout the entire period in which a stimulus is remembered, providing a neural correlate for the short-term store. Atkinson and Shiffrin also referred to the short-term store as working memory, using a term introduced by Miller and colleagues (Miller 1960, Atkinson, Berry et al. 2018, Malmberg, Raaijmakers et al. 2019). Despite its appeal, the Atkinson–Shiffrin model was not without shortcomings. It assumed a single short-term memory system for all memory items and it proposed rehearsal as the main mechanism for the transfer of information from short-term to long-term memory, without which, information would decay (Plancher and Barrouillet 2020).

Spurred by experimental observations that contradicted these tenets, Baddeley and Hitch put forth a model of working memory (Fig. 1B), the most recent version of which distinguishes between a central executive and three subsidiary systems, the phonological loop, the visuospatial sketchpad, and episodic buffer (Baddeley 2000). Auditory information, spoken and written words, and any other phonological information is thought to be stored in the phonological loop, which retains a limited store of objects through the process of real time vocal and subvocal rehearsal. The visual-spatial sketchpad plays a similar role for visual and spatial information, with limited interference from items maintained in the

phonological loop. The most recent addition to the Baddeley model is the episodic buffer, which is thought to be a passive storage system that would link the different memory domains (Baddeley 2012). This is a highly important function, as the integration of objects from different types of WM would allow the creation of entirely new objects from our imagination. Moreover, given the limits of working memory, the episodic buffer would serve the critical purpose of allowing a collection of associated features to be stored as a single complex object, to increase the efficiency of WM capacity (Cowan 2001, Eng, Chen et al. 2005). In addition, the episodic buffer also plays an important role in linking WM to long term memory (LTM), with retrieval occurring through conscious awareness (Baddeley and Andrade 2000). As a result, the episodic buffer would be the component of WM which makes objects consciously available. Arguably, the most important component of the Baddeley model is the central executive, which is responsible for the control and regulation of the other components. This is achieved through focusing attention, which allows specific objects to be selected in WM, or dividing attention simultaneously between multiple targets. Another executive function is the ability to switch between tasks. Finally, the central executive assists in connecting WM with LTM, both in recalling object that were previously stored in LTM, such as an individual attempting to remember a prior event, as well as selecting objects to transfer to from WM to LTM (Gardiner, Gawlik et al. 1994, Norris 2019).

The Baddeley model has been tremendously influential, however it has also been the target of criticism. For example, the function of task switching in the central executive has proven to be more complex than originally thought, with multiple stages proposed, and different modalities requiring different cognitive resources (Rubinstein, Meyer et al. 2001, Sandhu and Dyson 2013). As a result, the prospect of a centralized task switching function appears increasingly remote (Parkin 1998, Miyake, Friedman et al. 2000, Whitney, Arnett et al. 2001). The correspondence between the model's conceptual systems and neural structures and processes is also not straightforward. One class of neuroscience models of working memory places the executive role of the model on the prefrontal cortex, whereas the subsidiary systems maintain the contents of working memory at the sensory cortices (e.g. visual, auditory) (D'Esposito and Postle 2015). However, other models dispute this division on the basis of strong evidence for sensory information being maintained within the PFC, by the same neurons that implement top-down control, thus supporting the idea of the PFC being the anatomical seat of both executive and subsidiary systems of WM (Riley and Constantinidis 2016).

In view of this criticism, and to provide an alternative framework for future studies, we propose a model of working memory inspired from the anatomical organization of the brain circuits involved in working memory maintenance and the integration of the executive and storage WM functions into the same neuronal circuits (Fig. 1C). Areas involved in working memory maintenance are represented in color in all of the models of Fig. 1. Our model places prefrontal cortex in the

center of the working memory circuit but recognizes that the sensory pathways that transmit information to prefrontal cortex are interconnected through reciprocal loops with the prefrontal cortex. In this sense, afferent cortical areas are not sufficient for the maintenance of working memory by themselves, but necessary by virtue of their connections that allow persistent activity to reverberate (and are drawn in intermediate color saturation). This proposed model supports the relative independence of different modalities, as these would activate different ensembles of neurons, though it emphasizes that this separation is not absolute within the prefrontal cortex; neurons with non-linear interaction of stimulus properties, or "mixed selectivity" have been described between domains (Rigotti, Barak et al. 2013). Our new model therefore represents a new starting point for the investigation of WM, building upon the models of the past.

Some recent studies of working memory, in an attempt to clarify ambiguities in vocabulary, have begun using the term "WM" to refer specifically to the process of manipulation for complex information, in contrast to "short term memory", which is then used exclusively to denote the memory of simple stimuli (e.g. colored squares) that are maintained without any further transformation (Todd and Marois 2004). Although the idea is far from unreasonable, it has created considerable confusion in the literature, as the use of "working memory" with its original meaning has persisted, in parallel.

Another unresolved conceptual debate centers on the relationship between memory and attention. At one extreme of the spectrum, WM and attention are

synonymous processes: we become consciously aware of objects when we attend them, and the application of attention, such as through rehearsal, is also used to maintain objects even when they are no longer present (Deubel and Schneider 1996, Kuo, Stokes et al. 2014, Sprague, Ester et al. 2014, Sprague, Ester et al. 2016). There is also strong evidence of overlap between the neural apparatuses of attention and WM (Constantinidis and Klingberg 2016). At the other extreme, attention and memory are readily dissociable, implying the possibility of attending a stimulus without maintaining it in memory (Lebedev, Messinger et al. 2004) or storing an item in memory without even awareness of it, let alone attention (Soto, Mantyla et al. 2011, Soto and Silvanto 2014). In this review, we take the position in favor of the likely dissociation of attention and WM, even if one does not take this concept to the extreme. For example, the concept of attention is meaningful for examining stimuli that are physically present; we tend to attend specific visual stimuli even under conditions that place no memory demand. The neural correlates of attention are also evident in the processing of sensory stimuli in the early sensory cortex, despite the fact that these neurons do not maintain discharges while WM is in use (Chelazzi, Miller et al. 1993, Treue and Maunsell 1996, Leavitt, Mendoza-Halliday et al. 2017). Conversely, a complete disregard for attention in WM does not seem to be possible, as attention plays a variety of roles, including the direction of focus to one of multiple stimuli held in memory (Rose, LaRocque et al. 2016).

It is also instructive to contrast WM with LTM. Classic neuropsychological studies suggested that patients with anterograde amnesia, most famously, patient

H.M., had intact WM (Squire and Wixted 2011). The basis of LTM storage was thus thought to be entirely separate from WM, being mediated by long term potentiation in the hippocampus rather than persistent discharges in the PFC (Ranganath, Cohen et al. 2005). This dichotomy has also been revisited in recent years as recent studies have recognized that the hippocampus is active during WM (Hampson, Simeral et al. 1999). Moreover a class of models suggest that synaptic mechanisms make up the primary mechanism of WM, just as they do for LTM (Mongillo, Barak et al. 2008). Neural correlates of intermediate scales in the continuum between short-term and LTM have also been recognized (Zhang and Williams 2015). However, a variety of qualitative differences still remain unchallenged. Specifically, LTM does not seem to have an upper limit to its duration, lacks an upper limit to storage, and is not impaired by the addition of new items, regardless of how many were already in storage. We therefore favor their conceptual separation.

Neural Basis of Working Memory

Early neurophysiological experiments in non-human primates identified neurons that not only respond to sensory stimuli, but remain active during a period after the stimuli were no longer present; this "persistent activity" therefore provided a neural correlate of working memory (Fuster and Alexander 1971, Funahashi, Bruce et al. 1989). Persistent activity has since been demonstrated in human intracranial recordings, as well (Kaminski, Sullivan et al. 2017). Visuo-spatial working memory has been a particularly fruitful model since spatial location can be varied parametrically and the activity of neurons representing each location can be studied systematically. Persistent activity in the prefrontal cortex has been shown to explain many aspects of behavioral performance in visuo-spatial working memory tasks, such that firing rate can predict whether the subject will recall the item correctly or not, or what location the subject will recall (Qi, Zhou et al. 2015). However, in recent years, alternative accounts for the neural correlates of working memory have been introduced. We will group these alternatives into two categories: first, those that rely on rhythmic activity and second, those that propose information encoding without changes in mean firing rate during the delay period, that is, "activity silent" models.

Persistent Discharges

A great deal of experimental work has been centered on the representation of spatial information in working memory and theoretical research has established a concrete framework in the context of the "bump attractor model" (Meyer, Qi et al. 2007, Wimmer, Nykamp et al. 2014, Riley and Constantinidis 2015). Models for the maintenance of object memory have been comparatively less established, considering that a near infinite number of objects can be stored in memory with no obvious network structure that can represent them in an equivalent, parametric fashion, and only a small percentage of neurons are active during maintenance of any object in memory (Meyer, Qi et al. 2007). In light also of the relative segregation of spatial and non-spatial information in the brain (Mishkin, Ungerleider et al. 1983, Wilson, Scalaidhe et al. 1993), we will henceforth consider these systems to be interlinked, but ultimately separate domains. In each case, we examine the evidence for persistent activity being the neural correlate of working memory and discuss arguments raised for and against it.

Spatial Working memory

Spatial working memory in both animal models and humans has been assessed through a variety of classical tasks, including the delayed response, delayed alternation, and match/nonmatch tasks. In the oculomotor delayed response task (ODR – also referred to as the memory guided saccade task) a brief visual stimulus is presented, and must be maintained in the subject's working memory throughout a delay period, after which it is reported via an eye movement (Funahashi, Bruce et

al. 1989, Rao, Williams et al. 1999, Constantinidis, Franowicz et al. 2001). Another common task, the delayed alternation task, also requires (hand or eye) movements to one of two locations, alternating in successive trials and therefore requiring the preceding location to be maintained in working memory (Kubota and Niki 1971, Niki 1974). Individual neurons exhibit persistent activity with selectivity for different spatial locations during the performance of these tasks (Fig. 2), thus allowing the remembered location to be decoded from the activity of the neuronal population (Meyers, Qi et al. 2012). The dorsolateral prefrontal cortex appears to be particularly specialized for spatial location, and recordings that sampled prefrontal neurons in a random, unbiased fashion (without isolating neurons based on their responses) found approximately 30% of neurons exhibit persistent discharges during spatial working memory tasks (Qi and Constantinidis 2013). This percentage varies between the anterior/posterior and dorsal/ventral subdivisions, with the mid-dorsal area exhibiting the highest proportion of neurons with persistent activity (Riley, Qi et al. 2018). This proportion may even be underestimated by the limited number of spatial locations typically sampled in the ODR task (e.g. eight spatial locations arranged on a ring of 10-15 degree eccentricity); experiments that used more extensive arrays of stimuli, encompassing 16 locations, reported persistent discharges from as many as 70% of all prefrontal neurons (Leavitt, Pieper et al. 2017).

Does persistent activity represent merely motor preparation? This question was once a common argument against persistent activity being a neural correlate of

working memory (Markowitz, Curtis et al. 2015, Lundqvist, Herman et al. 2018), stemming from how the location of the preceding stimulus in ODR tasks is confounded with the direction of the motor response. However, support for this argument has been critically undermined as more complex tasks have since revealed that only a minority of prefrontal neurons represent motor preparation when this factor is dissociated from stimulus properties. For example, when a task requires monkeys to make an eye movement to the location opposite to the location of the remembered visual stimulus (delayed anti-saccade task), or to a location rotated relatively to the stimulus location (rotational ODR task), the majority of prefrontal neurons that generate persistent activity represent the location of the preceding stimulus rather than the location of the impeding saccade (Funahashi, Chafee et al. 1993, Takeda and Funahashi 2002). Moreover, persistent activity tuned for the location of a stimulus appears in the prefrontal cortex even in tasks where the stimulus does not immediately allow planning of a movement. This can be observed in the spatial delayed-match-to-sample task, where subjects are required to release a lever or press a button when a stimulus appears at a previously cued location. Prefrontal neurons generate persistent activity following the presentation of the original stimulus that is tuned for its spatial location, and not the preparation of a motor response, which cannot be planned until after the end of the delay period (Qi, Katsuki et al. 2010, Qi, Meyer et al. 2011, Goodwin, Blackman et al. 2012, Masse, Hodnefield et al. 2017).

Is persistent activity merely an epiphenomenon of spatial working memory?

Strong evidence exists that persistent discharges are causally related to behavior. For example, working memory performance is significantly impaired when persistent activity is abolished via the reversible inactivation of the prefrontal cortex, e.g. through cooling (Chafee and Goldman-Rakic 2000). Under normal execution of the task, without such an intervention, trials in which persistent activity is diminished are also more likely to result in errors (Funahashi, Bruce et al. 1989, Zhou, Zhu et al. 2013). A near linear relationship between behavioral performance and persistent activity has been revealed in other tasks that parametrically modulate the difficulty of a working memory judgment (Constantinidis, Franowicz et al. 2001). Lower performance of the ODR task in adolescent and aged monkeys (Wang, Gamo et al. 2011) is also associated with lower levels of persistent activity compared to young adults. Choice probability analysis, comparing the distributions of firing rates in the delay period of correct and error trials, also reveals a strong relationship between prefrontal persistent activity and the behavioral outcome of each trial (Mendoza-Halliday, Torres et al. 2014).

Computational models provide mechanistic detail of how persistent activity influences working memory performance. Persistent activity is sustained in such models by virtue of re-entrant connections between neurons with similar tuning for stimulus properties, so that activation after afferent input is maintained in the network, which behaves as a continuous attractor (Seung, Lee et al. 2000). The bump (peak) of activity in the network determines the location recalled by the

subject (Fig. 2), hence the term bump attractor (Wimmer, Nykamp et al. 2014). Drifts in neuronal activity thus account for deviations of behavior: persistent activity recorded from trials in which monkeys make eye movements deviating clockwise vs. counterclockwise relative to the true location of the stimulus yields slightly different tuning curves, implying that the peak of activity at the end of the delay period determines the recalled location (Wimmer, Nykamp et al. 2014). Similarly, the variability of a neuron's delay period activity (estimated by the Fano factor of spike counts, i.e. the variance divided by the mean) is maximal for inaccurate saccades to locations at the flanks of the neuron's tuning curve but lower for locations in the peak or tail. This finding is also explained by small deviations in saccadic endpoint corresponding to the bump of activity shifting in one direction or another, with the most rapid changes in neuronal activity occurring if the bump traverses the flank of its tuning curve rather than its peak or tail. Finally, spike-count correlations of two simultaneously recorded neurons are lowest and negative for inaccurate saccades when the cue appears between the peaks of their tuning curves. This result is also consistent with the idea that working memory inaccuracies are caused by drifts of persistent activity in the delay period; when the bump attractor randomly varies around a location between the peaks of two neurons, it inevitably causes an increase in firing rate for one neuron, but a decrease for the other. Importantly, these findings do not hold for neurons that do not exhibit persistent discharges, despite these comprising the majority of the prefrontal population (Wimmer, Nykamp et al. 2014).

Is persistent activity an "artifact of averaging"? Some neurophysiological experiments have reported that individual neurons are only transiently representing information in the delay period (Zaksas and Pasternak 2006) and persistent activity can be highly variable during the course of a trial, and from trial to trial (Lundqvist, Rose et al. 2016). The stimulus properties can be maintained across the entire delay period only if one were to average activity across multiple trials and multiple neurons (Hussar and Pasternak 2012, Hussar and Pasternak 2013). This led to the idea that previous reports of persistent activity were an artifact of averaging. However, these findings are entirely consistent with computational models of persistent activity. Across the population of prefrontal neurons, only a small minority would be expected to be active during maintenance of any single stimulus in memory. Variability in discharge rate during the course of the trial would be expected even among neurons that are highly active at some time point, as the activity might drift in the population. In fact, increased spiking irregularity (quantified by the coefficient of variation of the inter-spike interval) has been observed in the delay period compared to the fixation period of the ODR task (Compte, Constantinidis et al. 2003). This otherwise puzzling finding is precisely predicted by the network models of persistent activity. Conversely, if working memory were characterized by short intermittent bursts of high-rate firing (Lundqvist, Rose et al. 2016), then across-trial spike-count variability (e.g., as quantified by the Fano factor) would increase dramatically during the delay periods, relative to the fixation period (Li, Constantinidis et al. 2020), which is inconsistent

with empirical measurements (Chang, Armstrong et al. 2012, Qi and Constantinidis 2012).

Are temporal dynamics of experimentally observed delay-period activity inconsistent with persistent activity? Persistent activity is not stationary during the delay interval (Hussar and Pasternak 2010, Qi, Katsuki et al. 2010, Hussar and Pasternak 2012, Hussar and Pasternak 2013), which does represent a contradiction of the simplest, bump attractor models, though fundamentally, models of persistent activity describe properties of a *population code*, rather than an individual neuron. Specifically, the working memory representations are encoded as a pattern of activations across a population of neurons that is not dependent upon any individual cell. Theoretical and empirical analyses have shown that stable population coding of working memory is consistent with time-varying neuronal activity (Machens, Romo et al. 2010, Druckmann and Chklovskii 2012, Murray, Bernacchia et al. 2017). Principal Component Analysis reveals a low-dimensional representation, where stimulus location evolves dynamically in time after the cue presentation, but different locations remain constrained in separable subspaces (Murray, Bernacchia et al. 2017).

Non-Spatial Working memory

Prefrontal neurons generate persistent discharges that represent the identity of objects held in memory, though smaller percentages of neurons are active during feature working memory than spatial working memory. Whether object working

memory is localized in a different subdivision of the prefrontal cortex than spatial working memory (ventral vs. dorsal) has been the matter of debate (Wilson, O Scalaidhe et al. 1993, Rao, Rainer et al. 1997). At least a quantitative difference seems to be evident, with spatial information more prevalent in the dorsolateral prefrontal cortex than the ventrolateral prefrontal cortex, and this dissociation is more pronounced in the posterior rather than the anterior prefrontal cortex (Meyer, Qi et al. 2011, Constantinidis and Qi 2018).

Stimulus-selective persistent activity has been described in working memory tasks that require the maintenance of stimulus identity or features, such as shape, color, or luminance (Quintana, Yajeya et al. 1988, Hoshi, Shima et al. 1998, Constantinidis, Franowicz et al. 2001, Sakagami, Tsutsui et al. 2001, Averbeck, Chafee et al. 2003, Inoue and Mikami 2006, Genovesio, Tsujimoto et al. 2009). Other experiments have reported stimulus-selective persistent activity it tasks that required subjects to remember complex images, such as real objects and faces, or abstract pictures (Wilson, O Scalaidhe et al. 1993, Miller, Erickson et al. 1996, O Scalaidhe, Wilson et al. 1997, Rao, Rainer et al. 1997, Rainer, Asaad et al. 1998, O Scalaidhe, Wilson et al. 1999, Rainer and Miller 2000, Freedman, Riesenhuber et al. 2001, Roy, Buschman et al. 2014). Robust persistent activity has been described for the direction of motion of a random-dot stimulus that is always presented at the same location (Zaksas and Pasternak 2006, Mendoza-Halliday, Torres et al. 2014). Persistent neuronal firing in prefrontal cortex has been observed even in the absence of performance of a task, or even learning of a task, while subjects view

stimuli passively, and we are thus able to recall encountered stimuli even when we are not prompted to remember them ahead of time (Qi, Zhou et al. 2015). Consistent with this finding, recordings during passive fixation reveal persistent discharges selective for faces in the ventrolateral prefrontal cortex (O Scalaidhe, Wilson et al. 1999).

However, the evaluation of these findings is complicated by the recent revelation that persistent activity in the prefrontal cortex also represents information beyond the characteristics of stimuli, including the abstract rules of the cognitive task subjects are required to perform (White and Wise 1999, Wallis, Anderson et al. 2001), categories (Freedman, Riesenhuber et al. 2001, Shima, Isoda et al. 2007), and numerical quantities (Nieder, Freedman et al. 2002). Persistent activity may be also represent perceptual decisions (Kim and Shadlen 1999, Barraclough, Conroy et al. 2004), reward expectation (Leon and Shadlen 1999), and sequences of events or actions (Averbeck, Chafee et al. 2002, Inoue and Mikami 2006, Sigala, Kusunoki et al. 2008, Berdyyeva and Olson 2010). Persistent firing may even represent different aspects of the same stimulus, depending on task instructions (Mante, Sussillo et al. 2013), and a subset of neurons can represent multiple stimulus features and task variables simultaneously, a phenomenon known as mixed selectivity (Rigotti, Barak et al. 2013, Parthasarathy, Herikstad et al. 2017, Dang, Jaffe et al. 2020).

The realization that prefrontal activity is modulated by task factors to such extent has led to a re-evaluation of the nature of information represented in

persistent activity (D'Esposito and Postle 2015). The most extreme possibility would suggest that all stimulus-selective information actually originates from task rules or categorical judgments between alternatives rather than representing the memoranda themselves. For example, in a study that required the maintenance of stimulus color in working memory, significantly more prefrontal neurons were selective to location than color, despite the fact that only color was task relevant (Lara and Wallis 2014). Another experiment, which required the maintenance of a sequence of stimuli, revealed a drop off in their linear classifier's ability to decode any stimuli that preceded the most recent from prefrontal activity (Konecky, Smith et al. 2017).

These negative findings must be interpreted cautiously. To explain the lack of color selectivity, the activation of only a small proportion of prefrontal neurons, in the order of 5-15% (Lara and Wallis 2014) may be sufficient for the representation of stimulus information. It is also possible that color-selective neurons—and their persistent activity patterns—are concentrated in specific prefrontal "patches" (Lafer-Sousa and Conway 2013) rather than be diffused across the entire prefrontal surface. Moreover, to explain the decreased decoding ability for less recent stimuli, information about multiple stimuli may be abstracted (Tang, Riley et al. 2017) so that activity representing a sequence may differ from the representation of each stimulus in the sequence, thus resulting in an apparent negative finding. The generation of stimulus-selective persistent activity in monkeys never trained to perform a task also argues against the idea that persistent prefrontal activity *only*

represents tasks and rules and that stimulus information is mediated in other brain areas or through other mechanisms. Prefrontal neurons also routinely represent stimulus features even when they are irrelevant for the task at hand (Constantinidis, Franowicz et al. 2001, Lauwereyns, Sakagami et al. 2001, Donahue and Lee 2015). For example, in a study that required the maintenance of stimulus color in WM, significantly more prefrontal neurons were selective to location than color, despite the fact that only color was task relevant (Lara and Wallis 2014)

Rhythmic Models

Rhythmic activity has long been implicated in hippocampal-dependent memory, and communication between the hippocampus and prefrontal cortex, in rodents (Buzsaki 2010). In the human literature, the frequency of oscillations evident through MEG, EEG and ECoG recordings has also been associated with distinct working memory processes (Roux and Uhlhaas 2014). Recent neurophysiological studies in nonhuman primates have begun to address more specifically what role rhythmic firing may play in working memory (Siegel, Warden et al. 2009, Buschman, Denovellis et al. 2012, Liebe, Hoerzer et al. 2012, Salazar, Dotson et al. 2012, Brincat and Miller 2015). The magnitude, frequency, and phase of oscillations within the PFC and between the PFC and other areas have been shown to be modulated depending on stimulus and task information (Buschman, Denovellis et al. 2012, Liebe, Hoerzer et al. 2012), thus allowing information about the stimulus held in memory or the task to be performed to be decoded based on these parameters. For

example, oscillatory synchronization between LFP signals recorded from different sites within the PFC has been shown to be modulated based on which of two task rules a monkey is performing (Buschman, Denovellis et al. 2012). The coherence in rhythmic synchronization between neurons in prefrontal and posterior parietal cortices has also been reported to be content dependent with neurons from both areas synchronizing their firing at specific frequencies, for different stimuli held in memory (Salazar, Dotson et al. 2012). The phase of rhythmic activity seems able to differentiate information representing two sequentially presented stimuli (Siegel, Warden et al. 2009).

Recent studies have specifically proposed that the rate of LFP bursting in the gamma frequency range, which correlates negatively with power in the beta frequency, underlies WM (Lundqvist, Rose et al. 2016, Bastos, Loonis et al. 2018, Lundqvist, Herman et al. 2018, Stanley, Roy et al. 2018, Wutz, Loonis et al. 2018). A corollary of this model is that gamma bursting pooled from error trials should be lower than that of correct trials. Unfortunately, no measures of behavior were shown to correlate with the purported neural basis of WM during the delay interval in any of these studies. Differences in gamma bursting between correct and error trials were reported in one study (Lundqvist, Herman et al. 2018). Critically, no differences were reported during the delay periods following the sample presentations in the WM task used. Instead, error and correct trials were differentiated by levels of gamma bursting only during the period when test stimuli were presented and the monkeys had to judge whether they matched or not stimuli

held in memory, and errors were characterized by generally higher (not lower) levels of gamma bursting (Lundqvist, Herman et al. 2018).

Gamma bursting may still be necessary for the bottom-up input of sensory information (Miller, Lundqvist et al. 2018). For example, in human MEG studies, gamma oscillations were demonstrated to follow visual information through the cortical hierarchy during processing into WM (Michalareas, Vezoli et al. 2016). Moreover, directing attention to sensory objects leads to sensory enhancements consistent with prior association between attention and synchronized prefrontal gamma oscillations (Desimone and Duncan 1995, Bauer, Oostenveld et al. 2006, Womelsdorf, Fries et al. 2006). In any case, oscillatory activity is not incompatible with persistent activity, but instead, might reflect the underlying persistent firing and its ramifications from a distance. For example, both robust persistent activity and gamma-band rhythmicity have been reported during the delay period of the ODR task (Pesaran, Pezaris et al. 2002, Lundqvist, Rose et al. 2016), as well as the two-item sequential WM task (Siegel, Warden et al. 2009). Furthermore, concurrent persistent activity and gamma-band rhythmicity are observed when recordings are performed in the cortical site that corresponds to task demands: both increased persistent firing (Warden and Miller 2007) and gamma-band activity (Lundqvist, Herman et al. 2018) were captured from more ventral recording sites during an object feature WM task. Similarly, the classic ODR spatial WM task that generates persistent firing in dorsolateral PFC was associated with pronounced gamma bursts from the same region (Lundqvist, Rose et al. 2016). Thus, although measures of

oscillatory activity allow the researcher to sample a broader range of neuronal activity than can be performed with single or multiple unit recordings, persistent firing appears to underlie the oscillatory events captured during WM.

Activity Silent Models

Another class of WM models suggests that information may be maintained in WM by neuronal population discharge patterns or synaptic mechanisms that are not evident through spiking at all. We refer to these as "dynamic encoding" and "synaptic" models, respectively, in the sections that follow.

Dynamic Encoding

Information may be represented dynamically in a neuronal population, even in the absence of rhythmicity. Specifically, the precise pattern of activation of an ensemble of neurons at each time point during a WM task can be used to decode the identity of the stimulus, even though overall activity during the delay period is not significantly elevated above the baseline (Stokes, Kusunoki et al. 2013). Furthermore, this pattern of neuronal activation may vary dynamically over time, with different neurons becoming active at different time points (Fig. 3). Stimuli that were decoded dynamically absent persistent activity are similar to those used in previous studies where persistent activity was observed (Miller, Erickson et al. 1996, Rao, Rainer et al. 1997, Rainer, Asaad et al. 1998). Thus, although persistent

activity may have been generated, the averaging of the population might have made it difficult to detect. Dynamic activity informative about stimulus identity and task rules has been observed even when informative persistent activity is also present in the population (Crowe, Averbeck et al. 2008, Meyers, Qi et al. 2012). Moreover, stimulus location evolves dynamically in time after the cue presentation, while different locations remain constrained in separable subspaces (Murray, Bernacchia et al. 2017, Spaak, Watanabe et al. 2017, Parthasarathy, Tang et al. 2019).

.

Synaptic Models of Working Memory

Activity elicited after repeated presentation of the same stimulus is typically reduced, a phenomenon termed repetition suppression (Grill-Spector, Henson et al. 2006). As a result, the level of response to a particular stimulus in the context of a WM task, such as the delayed match to sample task, can be informative about whether it was preceded by the same stimulus or not; match suppression may signal that the sample was the same as the match (Fig. 4A-B). This suppressed response to a matching stimulus does not require persistent activity and may be observed even when multiple seconds intervene between the sample and match (Miller, Li et al. 1991, Miller, Erickson et al. 1996). Match suppression (or enhancement, for some neurons) is observed for stimuli matching in shape, color, and form, in spatial location, or in direction of motion, in various cortical areas, including the prefrontal, posterior parietal, and inferior temporal cortices (Miller, Li

et al. 1991, Steinmetz, Connor et al. 1994, Miller, Erickson et al. 1996, Zaksas and Pasternak 2006, Woloszyn and Sheinberg 2009). Furthermore, the extent of response difference to matching and nonmatching stimuli has predictive power over behavioral performance, as it differs systematically in correct and error trials, thus providing compelling evidence that memory performance has access to this activity (Zaksas and Pasternak 2006, Qi, Meyer et al. 2012).

Computational models have been proposed that could account for such changes via mechanisms that depend on synaptic strength modifications instead of spike generation (Mongillo, Barak et al. 2008, Sugase-Miyamoto, Liu et al. 2008). Such mechanisms may be mediated by calcium availability at the presynaptic terminal (Fig. 4C), whose kinetics has a time constant in the scale of seconds (Mongillo, Barak et al. 2008). Another possible mechanism can be seen in the processes of long-term potentiation, which represents memories as changes in the synaptic architecture, rather than neuronal activity (Ranganath, Cohen et al. 2005). Importantly, the synaptic model for WM is not mutually exclusive with persistent activity, and cannot, by itself, account for WM performance in other tasks such as the ODR, delayed alternation, or free recall tasks, that do not depend on a comparison of a subsequent stimulus with a prior one.

In recent years, insights about the neural mechanisms of WM, and other cognitive functions, have been drawn from Artificial Neural Networks used to model brain processes (LeCun, Bengio et al. 2015, Sinz, Pitkow et al. 2019). Recurrent Neural Networks (RNN) in particular have been used very successfully to model

complex cognitive tasks, including those related to WM (Mante, Sussillo et al. 2013, Masse, Yang et al. 2019, Yang, Joglekar et al. 2019, Kim and Sejnowski 2021). These studies have demonstrated that although it was also possible to perform simple WM tasks by virtue of rapid changes in synaptic weights after appearance of a stimulus, mimicking the neuronal synaptic mechanisms discussed above, persistent activity also emerged spontaneously in the network (Masse, Yang et al. 2019, Yang, Joglekar et al. 2019). Importantly, however, even modestly complex tasks, such as the delayed-match-to-sample task, could not be performed by networks that did not generate persistent activity. Although not offering definitive evidence for or against the synaptic model, these findings are instructive in the limitations of synaptic mechanisms alone.

An Integrative Approach

Although we have pointed out shortcomings of activity-silent models in fully accounting for WM, we do not wish to suggest that the underlying phenomena are not true: phenomena such as repetition suppression are robust and ubiquitous and there may be a close relation between activity-based and activity-silent mechanisms. Indeed, modeling studies that have successfully implemented these activity-silent conditions invariably require the network to be configured close to the attractor network regime (Mongillo, Barak et al. 2008, Fiebig and Lansner 2017), the mechanism of persistent activity. This way, a nonspecific drive can take the network

to the attractor regime and reactivate a robust attractor response selected on the basis of the weak subthreshold trace. The attractor nonlinearity is necessary to increase the contrast of a subthreshold signal that is fading away. Moreover, there is also the possibility of subthreshold mechanisms playing a supportive role for persistent activity in attractor networks (Hansel and Mato 2013). As a result, we conclude that activity-based and activity-silent mechanisms may interact synergistically instead of serving as mutual alternatives.

Recent studies examined closely the relationship of persistent spiking and activity silent mechanisms in the context of serial biases in spatial WM (Barbosa, Stein et al. 2020, Stein, Barbosa et al. 2020). The location subjects would recall about the cue they had to remember is often biased into the direction of the cue in the previous trial (Fig. 4D), especially when the successive cues appear in close proximity (Fischer and Whitney 2014, Papadimitriou, Ferdoash et al. 2015). Serial biases span the inter-trial interval between successive trials when no persistent activity survives, suggesting they are mediated by activity-silent rather than spiking mechanisms (Papadimitriou, Ferdoash et al. 2015, Bliss and D'Esposito 2017, Bliss, Sun et al. 2017, Kilpatrick 2018). Indeed, between successive persistent activity mnemonic codes, an activity-silent code in the PFC carried stimulus information through inter-trial periods (Fig. 4E). Increased firing rates during the onset of the following trial's fixation period revealed this latent activation as the trace reactivated: firing rate shortly after the beginning of the fixation period was tuned for the location of the previous stimulus (Barbosa, Stein et al. 2020). This reactivation

interacted with the appearance of the following stimulus, attracting or repelling the bump generated by the appearance of the stimulus, thus creating the serial bias in memory (Barbosa, Stein et al. 2020, Stein, Barbosa et al. 2020). This interplay could be the basis of closely associated memory storage processes operating at different time scales and serving different behavioral purposes, possibly including volitional effortful memory or occasional reactivation of recent experiences from latent traces. The overarching principle that we draw from this discussion is that synaptic mechanisms are important to the extent that they influence spike generation during WM, which manifests itself as persistent activity.

Circuit Mechanisms of Persistent Activity Generation

The mechanisms through which persistent activity is generated and the reasons it is much more prevalent in the PFC can be traced to a number of circuit specializations. We will review those related to pyramidal neurons and interneurons, and to the neuromodulatory role of the monoamine neurotransmitter systems.

Pyramidal Neurons and Excitatory Connections

The classical view of pyramidal neurons has been that they are essentially uniform across the cortex. This idea has been challenged by experimental findings demonstrating a systematic difference across the cortical hierarchy, with the prefrontal pyramidal neurons exhibiting the most extensive dendritic trees and the largest number of spines among cortical neurons (Elston 2000, Elston 2003). Functional correlates of this anatomical specialization are reflected in the patterns of neuronal discharges at different areas. Prefrontal neurons receive a greater proportion of distal synaptic inputs compared to the neurons at other brain areas, with a substantial proportion of these inputs originating at distances greater than 1 mm. By contrast, the majority of inputs to posterior parietal neurons appear to originate from neurons at shorter distances, in the order of 0.2-0.5 mm (Katsuki, Qi et al. 2014, Hart and Huk 2020). Independent evidence for this finding is provided by anatomical studies of myelin. The MRI-based T1-weighted/T2-weighted ratio is

indicative of the extent of myelin presence within gray matter and provides a measure of convergence of axonal projections (Glasser and Van Essen 2011, Huntenburg, Bazin et al. 2017). The highest ratio is observed in the primary visual cortex and lowest (indicating most sparse connections) in PFC (Burt, Demirtas et al. 2018). It has long been speculated that prefrontal neurons with similar memory fields are grouped in clusters with reciprocal connections, often visualized in anatomical tracer studies (Goldman-Rakic 1984, Levitt, Lewis et al. 1993, Kritzer and Goldman-Rakic 1995, Pucak, Levitt et al. 1996), and more extensive networks of neurons in the PFC would explain the stability of prefrontal persistent activity.

NMDA Receptors

NMDA receptors are glutamate-gated cation channels, critical for the generation of persistent activity, as they are responsible for extending the duration of the postsynaptic depolarization through their relatively slow decay time constant (Wang 2001, Constantinidis and Wang 2004). Thus, a circuit with exclusively AMPA receptors—which produce synaptic currents with very fast decay time constants—would require unrealistically high firing rates to sustain neural activity during the delay period of a memory task (Wang 1999). Experimental results further support the role of NMDA receptors in the generation of persistent activity, as the systemic administration of ketamine, a nonspecific NMDA antagonist, seems to decrease the effective connectivity between prefrontal neurons, demonstrated by a decrease in the synchronous spiking between simultaneously recorded neurons (Zick,

Blackman et al. 2018). Persistent activity also seems to be degraded by the iontophoretic diffusion of NMDA antagonists through the cortex (Wang, Yang et al. 2013, Wang and Arnsten 2015), though interestingly, this result has failed to replicate in some studies (Rodermund, Westendorff et al. 2020). The area-specific expression of NMDA further underlies its role in facilitating the prevalence of persistent activity in the PFC. For example, GluN2B—the NMDAR subunit with the slowest decay time constant-is expressed in a gradient across the primate brain, with highest levels of expression observed in the PFC. NMDA receptor trafficking is regulated by activation of dopamine receptors (Wang, Wong et al. 2012) and dopamine innervation is concentrated in the frontal lobe (Levitt, Rakic et al. 1984). Thus, the NDMA receptor also represents one of the main mechanisms of dopaminergic influence on persistent activity in the PFC.

Interneuron Specialization

Prefrontal inhibitory neurons exhibit persistent activity that is thought to be essential for stimulus selectivity in WM (Rao, Williams et al. 1999, Rao, Williams et al. 2000, Constantinidis and Goldman-Rakic 2002). Prefrontal interneurons generally exhibit higher baseline firing rates and broader tuning than pyramidal neurons (Constantinidis and Goldman-Rakic 2002). Their action serves to “sculpt” the spatial and temporal tuning of prefrontal neurons (Constantinidis, Williams et al. 2002), without which stimulus-specific persistent activity is much less viable in computational models (Compte, Brunel et al. 2000).

Cortical interneurons are hypothesized to use multiple types of GABAergic neurons to form specialized networks (Figure 5) for the purpose of facilitating stimulus-specific persistent activity (Wang, Tegner et al. 2004). Three types account for most interneurons in the cortex, those expressing parvalbumin (PV), those expressing vasoactive Intestinal peptide which tends to co-localize with calretinin (Gabbott and Bacon 1997), and those expressing somatostatin (SST), which tends to co-localize with calbindin. PV interneurons target the cell bodies of pyramidal neurons and when activated by their preferred stimulus, they would tend to suppress the activation of pyramidal neurons with different spatial turning than their own and sharpen the tuning function of those with similar tuning (Li, Zhou et al. 2020). Without feedback inhibition, recurrent excitation may shift the excitatory/inhibitory balance to a hyper-excited state and bring the network into an unstable, hyper-excited state, which would also be deleterious for the maintenance of WM (Constantinidis and Wang 2004).

VIP/calretinin interneurons are thought to inhibit other types of interneurons including SST/calbindin ones (Meskenaite 1997, Melchitzky and Lewis 2008, Fish, Rocco et al. 2018). Furthermore, interneuron-targeting cells are more abundant in association cortices, particularly in the PFC, compared to the sensory cortex (Defelipe, Gonzalez-Albo et al. 1999, Elston and Gonzalez-Albo 2003). SST interneurons, on the other hand, are peridendritic-targeting cells and are thought to exhibit high spontaneous firing rates that may tonically inhibit all pyramidal neurons during the task baseline period, prior to any stimulus presentation. SST neurons

would lift inhibition on the pyramidal neurons that are excited by a stimulus maintained in WM, while other SST neurons, not recruited by the maintained stimulus, would continue to inhibit nonactivated pyramidal neurons, thus suppressing both background noise, and any potential activation by subsequent, distracting stimuli (Wang, Tegner et al. 2004).

Anatomical and physiological evidence supports the greater pronunciation of the disinhibiting circuit in the PFC compared to other areas. Calretinin-positive interneurons are more numerous in the PFC compared to visual cortical areas MT and MST (Torres-Gomez, Blonde et al. 2020). Moreover, interneurons with high baseline firing rate and inverted tuning (consistent with the profile of disinhibiting neurons) have also been found to be more numerous in the PFC than in the posterior parietal cortex (Zhou, Katsuki et al. 2012). The importance of inhibitory-to-inhibitory connections has been confirmed independently by neural-network modeling studies (Kim and Sejnowski 2021). Such connections emerge in the network as training of synaptic weights progresses and play a critical role in maintaining WM activity. The increased presence of these circuits thus underlies the prefrontal specialization toward persistent activity.

Neuromodulatory Systems

As noted, dopamine innervation is concentrated in the frontal lobe (Levitt, Rakic et al. 1984) and the D1 receptor has been implicated in the generation of persistent activity. Iontophoretic application of D1 receptor antagonists, at least in large

doses, compromise WM function and erode persistent activity in the ODR task (Sawaguchi and Goldman-Rakic 1994, Williams and Goldman-Rakic 1995). In contrast, D1 agonists increase activity for preferred stimuli and suppress nonpreferred responses (Vijayraghavan, Wang et al. 2007, Ott, Jacob et al. 2014). However, the effects of dopamine receptors are complex and depend on dosage (Williams and Goldman-Rakic 1995, Vijayraghavan, Wang et al. 2007), with differential effects on pyramidal neurons and interneurons (Jacob, Stalter et al. 2016). D2/D3 antagonists also suppress persistent activity, though their action primarily modulates motor responses of prefrontal neurons (Wang, Vijayraghavan et al. 2004). Computational and experimental studies suggest that the overall effect of dopamine is to improve the signal-to-noise ratio of persistent activity (Yang and Seamans 1996, Durstewitz, Seamans et al. 2000, Seamans, Durstewitz et al. 2001, Chen, Greengard et al. 2004).

Other monoamine neurotransmitter systems, although not exclusive to the frontal lobe, influence persistent activity. For example, blockade of α2a adrenergic receptors erodes spatially tuned persistent activity and α2a agonists enhance prefrontal persistent activity (Wang, Ramos et al. 2007). Similarly, cholinergic agents (both muscarinic and nicotinic) are known to influence persistent activity. Muscarinic antagonists impair WM performance (Bartus 2000) and suppress prefrontal persistent activity (Zhou, Qi et al. 2011). Nicotinic α7-nACh receptor agonists enhance and antagonists inhibit prefrontal persistent activity (Yang, Paspalas et al. 2013). Capitalizing on the known effects of the cholinergic system

some recent studies have relied on endogenous activation of the cholinergic system through electrical stimulation of the Nucleus Basalis of Meyenert to improve WM (Liu, Crawford et al. 2017, Qi, Liu et al. 2019).

Localization of Working Memory Activity in the Brain

Persistent discharges are not an exclusive property of the PFC. Neurons in premotor, parietal, cingulate, and temporal association areas also generate robust persistent activity, as do subcortical structures including the basal ganglia and the mediodorsal nucleus of the thalamus (Constantinidis and Procyk 2004, Pasternak and Greenlee 2005). Models of WM relying on activity-silent mechanisms have suggested that WM is localized in sensory areas, as early as the primary visual cortex (Harrison and Tong 2009). We will review the evidence for working memory representations across all of these brain areas.

Posterior Parietal (PPC) and Inferior Temporal (IT) Cortex

The posterior parietal and inferior temporal cortex represent two main cortical afferents to the PFC(Constantinidis and Procyk 2004). Neurons in posterior parietal cortex share many functional properties with neurons in the dorsolateral aspect of the PFC, to which they are interconnected (Rawley and Constantinidis 2009) and both regions are activated simultaneously in human imaging studies of spatial WM (Jonides, Smith et al. 1993, Courtney, Ungerleider et al. 1997, Owen, Stern et al. 1998, Ungerleider, Courtney et al. 1998, Marshuetz, Smith et al. 2000, Bunge, Ochsner et al. 2001, Stern, Sherman et al. 2001). Moreover, neurons in the posterior parietal cortex also generate persistent activity through a virtually identical percentage of neurons compared to the dorsolateral prefrontal areas during the

ODR task (Chafee and Goldman-Rakic 1998). The remembered locations of visual stimuli can be decoded from the persistent activity from either area, independent of a planned motor response (Gnadt and Andersen 1988, Constantinidis and Steinmetz 1996), thus implying that spatial WM representations may be either redundant or distributed across multiple areas at once.

Similarly, the IT cortex shares a number of physiological properties with the ventrolateral PFC and both exhibit persistent activity that encodes the features of remembered stimuli during the delay period of object WM tasks (Fuster and Jervey 1981, Fuster and Jervey 1982, Miyashita and Chang 1988, Miller, Li et al. 1993, Nakamura and Kubota 1995, Naya, Yoshida et al. 2001, Sigala and Logothetis 2002). As noted in previous sections, the simultaneous activation of the PFC and its interconnected areas during WM tasks has led to the implication that the PFC may not necessarily comprise the complete WM representations, but would only hold the focus of attention, or other top-down signals, while the actual contents of WM would be represented in the posterior parietal and inferior temporal cortex (Cowan 1988, Miller and Cohen 2001, Hazy 2006, Postle 2006, D'Esposito 2007).

One line of evidence suggesting that the PFC maintains stimulus information not present in its afferent areas comes from memory tasks that require maintenance in memory of an original item through sequential presentation of distracting stimuli, such as the object and spatial versions of the delayed match to sample task. In object delayed-match-to-sample task, persistent discharges of IT neurons are interrupted by nonmatching, distractor stimuli presented after the

sample (Miller, Li et al. 1993). Similarly, persistent posterior parietal discharges represent the most recent stimulus location and are disrupted by distracting stimuli in spatial delayed-match-to-sample tasks (Constantinidis and Steinmetz 1996). In contrast, responses in the PFC are able to represent the actively remembered object (Miller, Erickson et al. 1996) or spatial stimuli (di Pellegrino and Wise 1993, Qi, Katsuki et al. 2010, Suzuki and Gottlieb 2013).

More recent studies have somewhat qualified these findings, demonstrating that differences between IT/PPC and prefrontal neurons are quantitative rather than qualitative (Woloszyn and Sheinberg 2009, Qi, Katsuki et al. 2010), and that in some tasks, prefrontal neurons may display greater responses to distractors than actively remembered stimuli (Jacob and Nieder 2014, Qi, Elworthy et al. 2015). Nonetheless, performing the task reviewed in the previous paragraph seems impossible based on the activation of the posterior parietal or inferior temporal cortex alone. Accumulating studies support the possibility of the prefrontal and PPC/IT cortex simply being specialized for different aspects of WM, among other cognitive functions (Katsuki and Constantinidis 2012, Swaminathan and Freedman 2012, Crowe, Goodwin et al. 2013, Ibos, Duhamel et al. 2013, Jacob and Nieder 2014, Qi, Elworthy et al. 2015).

The primacy of the PFC in WM behavior is perhaps most vividly demonstrated in inactivation studies. Cooling experiments, which reversibly inactivate the underlying cortex by lowering its temperature, demonstrate much greater decreases in ODR performance after prefrontal cooling compared to

posterior parietal cooling (Chafee and Goldman-Rakic 2000), even when the inactivated areas have similar delay period activity (Chafee and Goldman-Rakic 1998). The results of these studies parallel the effects of reversible inactivation of the frontal eye fields via muscimol injections, which produce a significant impairment in memory-guided saccade performance (Sommer and Tehovnik 1997, Dias and Segraves 1999). In contrast, virtually no impairment was observed after the muscimol-induced inactivation of the posterior parietal cortex (Li, Mazzoni et al. 1999, Chafee and Goldman-Rakic 2000, Wilke, Kagan et al. 2012), even though posterior parietal inactivation produces consistent deficits in tasks that require attention or selection between multiple stimuli (Wardak, Olivier et al. 2002, Wardak, Olivier et al. 2004, Liu, Yttri et al. 2010, Wilke, Kagan et al. 2012). Small lesions to the dorsolateral PFC also produce impairment in WM performance for remembered stimuli in the contralateral space, an effect termed a “mnemonic scotoma” (Funahashi, Bruce et al. 1993, Funahashi 2015).

Visual Cortex

Recent human imaging studies have applied multi-variate pattern analysis (MVPA), examining the simultaneous pattern of activation of multiple voxels to different task conditions (Offen, Schluppeck et al. 2009), in order to successfully decode WM content from the primary (Harrison and Tong 2009, Albers, Kok et al. 2013, Xing, Ledgeway et al. 2013) and extrastriate visual cortex (Tong and Pratte 2012, Ester, Anderson et al. 2013, Sreenivasan, Vytlacil et al. 2014). MVPA analysis could not

decode information from the PFC, nor fully account for behavioral performance in the task (Harrison and Tong 2009, Sreenivasan, Vytlacil et al. 2014). As a result, some imaging studies have even claimed that the size of the primary visual cortex alone may be the best predictor of WM ability (Bergmann, Genc et al. 2014).

However, there are critiques that may call these results into question. A tacit assumption when comparing the results of MVPA analysis is that the structure of the voxel (typically in the order of 3 x 3 x 3mm) would be equivalent across different cortical areas. This may not be the case. Unlike the well-documented topography of visual space in the primary visual cortex, no retinotopic map (or other overarching organizational principle) has yet been revealed in the PFC, and the same retinal position is represented multiple times across the prefrontal surface. Precise stimulus location information, therefore, seems to be represented on an extremely fine spatial scale, with the entire visual hemifield possibly represented in prefrontal modules with a surface no larger than 0.5 x 0.5 mm (Constantinidis, Franowicz et al. 2001). Voxels averaging cortical volumes an order of magnitude larger are thus likely to eliminate stimulus information and fail to decode the information held in WM, even if this is robustly represented in activity of individual prefrontal neurons. Positive fMRI results, decoding features of a remembered stimulus directly from PFC have also been published: the orientation of a grating was successfully decoded from the PFC (Ester, Sprague et al. 2015). These results thus argue directly against models of WM that postulate a solely top-down control role for the PFC, and place feature storage networks in the visual cortex.

MVPA methods still yield undeniably positive findings of information decoding in visual cortex hence it is important to consider neurophysiological activity in the visual cortex during WM. By some accounts, most if not all areas of the visual cortex produce neurophysiological activity during working memory (Christophel, Klink et al. 2017). A more careful view of these results, however, leads to a more nuanced view. Some studies have reported stimulus-selective activity in a small percentage of V1 neurons during WM (Super, Spekreijse et al. 2001). However, activity that outlived the visual stimulus tended to be characterized by a suppression of firing rate below the baseline, rather than elevated persistent activity. Importantly, the paradigm used in these experiments relied on the presence of a background grating that remained visible after the foreground stimulus was no longer present, providing physical stimulation onto V1 neurons. Stimulus-selective activity in V1 during WM is likely due to top-down projections from higher associative cortices, since V1 activation appears first in superficial layers (Roelfsema 2015). The relatively "quiet" background activity in the V1 allows the observation of this subtle backwash from higher cortical areas, while the higher cortical areas themselves may be too noisy to detect these small signals. In addition to this backwash, fMRI activation may also be detecting presynaptic activation of V1 neurons from higher cortical areas (Logothetis and Wandell 2004). It is also unclear whether V1 activity can be predictive of behavior in WM tasks as this modulation was present for both correct and incorrect trials (Super, Spekreijse et al. 2001). On the other hand, studies comparing activity in three cortical areas in

the same animals required to remember the direction of motion of a random-dot display, found virtually no persistent discharges in visual area MT, but robust activation in parietal area MST, in addition to prefrontal persistent activation (Mendoza-Halliday, Torres et al. 2014). This suggests an abrupt generation of feature-selective persistent activity in areas beyond the visual cortex. A more recent review of cortical areas generating persistent activity, including both positive and negative findings, revealed activity concentrated in the PFC and areas directly connected to it (Leavitt, Mendoza-Halliday et al. 2017).

Subcortical Structures

Basal Ganglia

In primates, the basal ganglia are formed by the striatum, the globus pallidus (Gpe, Gpi), the substantia nigra (SNc, SNr), and the subthalamic nucleus. Cortico-striato-thalamo-cortical loops, which involve various subdivisions of the PFC, are then used to project thalamic and Gpi/SNr outputs back to cortical areas (Selemon and Goldman-Rakic 1985, Alexander, DeLong et al. 1986, Haber, Kunishio et al. 1995, Middleton and Strick 2002). Several nodes of these loops are targets of dopaminergic cells. The presence of these inputs is therefore implicated as a possible mechanism through which these loops may contribute to WM-related behaviour beyond purely motor control, binding sensory, motor, and motivational cortical information (Hadj-Bouziane, Meunier et al. 2003). Substantial persistent activity has been described in the basal ganglia (Hikosaka, Sakamoto et al. 1989,

Hikosaka, Sakamoto et al. 1989, Kermadi and Joseph 1995, Wise, Murray et al. 1996, Kawagoe, Takikawa et al. 1998). As in PFC, visual and persistent activity during memory delays in the caudate nucleus reflects combined visual and motivational (reward expectation) information needed to perform WM tasks (Kawagoe, Takikawa et al. 1998). The basal ganglia are thus implicated in WM as a possible support system for prefrontal-dependant maintenance.

Thalamus

The thalamus is a key structure in both cortico-striato-thalamo-cortical loops, as well as direct cortico-thalamic systems. These latter systems may combine with cortico-cortical connections to implement collective computation and large-scale fast communication, which may ultimately be the mechanism of cognitive binding (Houk and Wise 1995, Llinas, Ribary et al. 1998, Llinas, Leznik et al. 2002). Several thalamic nuclei are directly connection with association cortices, and most importantly, the mediodorsal nucleus (MD) and the medial nucleus of the pulvinar (Goldman-Rakic and Porrino 1985, Giguere and Goldman-Rakic 1988, Selemon and Goldman-Rakic 1988, Romanski, Giguere et al. 1997, Rouiller, Tanne et al. 1999). Neurophysiological studies in MD have shown typical persistent activity during memory tasks (Fuster and Alexander 1973, Tanibuchi and Goldman-Rakic 2003, Watanabe and Funahashi 2004), though more neurons are activated by the motor aspects of the task (Watanabe and Funahashi 2004, Watanabe and Funahashi 2004).

Hippocampus

The hippocampus is well accepted to play a critical role in the consolidation of WM into LTM. This role implies tight communication with the PFC. Indeed, hippocampal neurons are readily activated during WM tasks, such as the delayed-match-to-sample task, and exhibit stimulus-selective persistent discharges (Hampson, Pons et al. 2004).

Long-Range Loops

The review of areas where persistent activity has been observed hints at some general principles. The PFC and the brain areas to which it is directly connected to form a series of parallel, long-range loops comprising the association cortex of the parietal and temporal lobe, and cortico-striatall, and cortico-thalamic circuits, where persistent activity is readily detectible. Recent theoretical studies suggest that inter-areal reverberation is essential for the emergence of persistent activity, even when some of the component areas are incapable of generating persistent activity on their own (Mejias and Wang 2019). In this context, the PFC can be viewed as a core component but not as the exclusive site of WM maintenance.

Functional Organization of Working Memory Within the Prefrontal Cortex

Anatomical organization

The PFC can be divided into a lateral, a medial, and an orbital aspect, the former of which aspect can be further distinguished into a dorsolateral and a ventrolateral subdivision (Constantinidis and Procyk 2004). Multiple cytoarchitectonic areas make up the lateral aspect of the macaque PFC, including areas 8a (encompassing the frontal eye field) and 45, which lines the superior and inferior banks of the arcuate sulcus respectively, area 8b, which lies just medial to the arcuate, areas 9 and 12 in the superior and inferior convexities of the cortex respectively, area 46, which lines the banks of the principal sulcus and finally, area 10, which covers the frontal pole (Walker 1940). Area 46 has since been further subdivided along its mediolateral axis, into areas 46dr, 46d, 46v and 46vr, which line the medial rim, the medial and the lateral banks of the principal sulcus, and the lateral rim of the principal sulcus, respectively (Preuss and Goldman-Rakic 1991). It has also been divided along its anterior-posterior axis into a caudal aspect—sometimes referred to as area 9/46—and a rostral one (Petrides 2000). fMRI studies have implied the possibility of additional areas based on functional connectivity (Goulas, Stiers et al. 2017).

Anatomical studies have found that projections from the posterior parietal cortex terminate primarily in the dorsal PFC (areas 8 and 46, including both banks of the principal sulcus), while the inferior temporal cortex projects to areas 12 and 45

of the ventral PFC, thus suggesting a relative segregation between the areas. (Petrides and Pandya 1984, Selemon and Goldman-Rakic 1988, Cavada and Goldman-Rakic 1989). A relative segregation of inputs has also been observed across the anterior-posterior axis, with areas higher in the sensory and limbic hierarchies projecting to more anterior prefrontal subdivisions, while lower areas likewise project to the more posterior subdivisions instead (Gerbella, Borra et al. 2013, Barbas 2015, Borra, Ferroni et al. 2017).

Localization of Stimulus Selectivity

Guided by the pattern of anatomical connections, early neurophysiological studies suggested that spatial and object WM were subserved by the dorsolateral and ventrolateral PFC, respectively—a "domain-specific" organization that was thought to extend from the dorsal and ventral visual streams (Ungerleider and Mishkin 1982, Felleman and Van Essen 1991, Wilson, Scalaidhe et al. 1993, O Scalaidhe, Wilson et al. 1997, O Scalaidhe, Wilson et al. 1999). However, opposition to this proposal has noted that after a monkey is trained in tasks that require stimuli location and identity, neurons with selectivity to these factors can be found in all areas of the PFC. (Rao, Rainer et al. 1997, Rainer, Asaad et al. 1998). This would imply that the apparent functional specialization observed in the earlier studies was actually the result of task requirements, thus leading to the proposal of a new "integrative" model, where task requirements and cognitive process engaged are the primary

principle of organization rather than the type of stimulus representation (Rao, Rainer et al. 1997, Owen, Stern et al. 1998, Rainer, Asaad et al. 1998).

Spatial Selectivity

More recent studies have reexamined prefrontal selectivity for different stimulus properties by comparing dorsal and ventral prefrontal responses to the same stimuli. Spatial selectivity (Figure 6) proved a strong predictor of whether a neuron was recorded in the dorsal or the ventral PFC, with dorsolateral PFC neurons displaying significantly higher spatial selectivity, regardless of whether or not the monkeys were trained, passively viewing the task, or actively maintaining the stimuli in their WM (Meyer, Qi et al. 2007, Meyer, Qi et al. 2011, Riley, Qi et al. 2017). Ventrolateral prefrontal neurons also displayed substantial spatial selectively with well-localized receptive fields, but only to a comparatively lesser degree (Meyer, Qi et al. 2011, Riley, Qi et al. 2017). Importantly, a gradient of stimulus selectivity is present along the anterior-posterior axis, with the posterior areas containing the most highly selective neurons; anterior neurons exhibit little selectivity to stimuli per se but are more likely to represent task variables (Riley, Qi et al. 2017). Spatial selectivity also exhibits a strong temporal component. The proportion of spatially selective neurons, for example, peaks early after the appearance of the stimulus and diminishes with time; response latency also differentiates the highly selective posterior areas from the less selective anterior ones (Kadohisa, Kusunoki et al.

2015, Riley, Qi et al. 2017). A functional specialization between (posterior) dorsal and ventral areas is also most evident for peripheral stimuli. Experiments that positioned stimuli 10 to 14 degrees of visual angle away from the fovea were able to detect the dorsal and ventral gradient of specialization, while studies that tested stimuli within 4 to 6° degrees from the fovea found little or no differentiation between dorsal and ventral prefrontal activity (Rao, Rainer et al. 1997, Kadohisa, Kusunoki et al. 2015).

Studies manipulating neuronal activity provide a second line of evidence regarding the functional specialization of the PFC. For example, temporary inactivation experiments, for example, injections of the $GABA_A$ agonist muscimol, or lidocaine, or cooling of the underlying cortex in the dorsal prefrontal areas, including the frontal eye field, causes spatial WM deficits that localize in the contralateral visual field (Funahashi, Bruce et al. 1993, Sommer and Tehovnik 1997, Dias and Segraves 1999, Chafee and Goldman-Rakic 2000, Sawaguchi and Iba 2001, Suzuki and Gottlieb 2013, Noudoost, Clark et al. 2014). In contrast, inactivation of the ventral PFC does not impair spatial WM (Bichot, Heard et al. 2015).

How visual space is organized in the PFC remains unresolved. No obvious, retinotopic organization of visual space has been described across the surface of the PFC, or with dense electrode areas placed in posterior dorsal areas with strong overall spatial selectivity (Figure 7A). Some indications exist for at least some local organization, however (Figure 7B). Electrode penetrations into the principal sulcus

suggest an orderly progression of receptive field locations (Arnsten 2013). Some organization has been revealed across cortical layers, at least in terms of time course of activation (Figure 7C): neurons activated by visual stimuli are more likely to be encountered in intermediate layers; neurons with delay period activity in superficial; neurons with motor responses in deep, though this relative segregation is quantitative rather than qualitative (Markowitz, Curtis et al. 2015).

Feature Selectivity

In monkeys trained to perform behavioral tasks, dorsolateral prefrontal neurons (including the frontal eye fields) exhibit substantial selectivity for object shape, no less than that of ventrolateral neurons (Peng, Sereno et al. 2008, Meyer, Qi et al. 2011, Clark, Noudoost et al. 2012), and mirroring the selectivity observed in the posterior parietal cortex (Sereno and Maunsell 1998, Janssen, Srivastava et al. 2008). Interestingly, however, if the monkeys are not trained, shape selectivity is higher in the ventrolateral PFC instead (Meyer, Qi et al. 2011). Unlike spatial location, which can be studied systematically with parametric variation, the virtually infinite variety of shapes prevents the detection of a clear-cut dichotomy in shape selectivity between the dorsal and ventral PFC. However, neurons with high feature selectivity will only respond vigorously to a limited set of stimuli and are likely to produce uniformly weak responses to stimuli drawn from a small set, failing to differentiate between them. This precisely matches the response patterns of

inferior temporal neurons (Gross, Rocha-Miranda et al. 1972, Desimone, Albright et al. 1984, Tanaka, Saito et al. 1991, Fujita, Tanaka et al. 1992) and ventrolateral neurons. By contrast, when neurons were tested with stimulus sets requiring very narrow shape selectivity, such as faces and complex objects, neurons distinguishing between such stimuli were localized exclusively in the ventral PFC (O Scalaidhe, Wilson et al. 1997, O Scalaidhe, Wilson et al. 1999).

Selectivity to stimulus color follows a similar pattern. Only a small proportion of the total prefrontal populations exhibit color selectivity, in the order of 5 to 15%, similar to the proportional selectivity in posterior parietal neurons (Constantinidis and Steinmetz 2001, Lara and Wallis 2014). Among neurons that respond to colored squares, selectivity for different color is not significantly different in dorsal than ventral PFC (Riley, Qi et al. 2017). Combined fMRI studies and neurophysiological studies in the temporal lobe have suggested that neurons selective for faces, other shapes, and colors are clustered at distinct patches of cortex (Tsao, Freiwald et al. 2006, Popivanov, Jastorff et al. 2014, Chang, Bao et al. 2017). A handful of studies have suggested a similar clustering for color-selective neurons in the PFC as well, thus strongly implying the precise anatomical localization of color representations in WM (Lafer-Sousa and Conway 2013).

Plasticity of Stimulus Representations

Plasticity of neural responses also exhibits differential localization in the PFC (Figure 6). While training increases the number of responsive neurons in the ventral and anterior PFC, the posterior dorsal PFC exhibits the weakest effects of plasticity, with robust selectivity for spatial stimuli being present both before and after training (Meyer, Qi et al. 2011, Riley, Qi et al. 2017). The emergence of robust selectivity after training would thus seem to indicate that anterior and ventral responses are more sensitive to task variables and cognitive factors—particularly reward conditions—rather than stimulus properties, and this capacity for plasticity has been linked to the action of dopamine D1R receptors (Puig and Miller 2012). The great plasticity of the PFC cannot entirely erase the preexisting functional specialization gradient for spatial and object information between the posterior aspects of the dorsal and ventral PFC. Moreover, training does not exert an equal effect on all areas of the PFC, thus implying that the effects of training on any given neuron would depend upon that neuron's anatomical location.

Functional Implications of Dorsal and Ventral Prefrontal Inactivation

Although selectivity to different types of stimuli is revealing, the ultimate functional role of an area can be probed by examining the consequences of its activity on performance of WM tasks. Thus, experiments manipulating neuronal activity can be very informative on the functional roles of prefrontal subdivisions. Temporary inactivation experiments, for example, injections of the $GABA_A$ agonist muscimol, or

lidocaine, or by cooling of the underlying cortex, during WM are consistent with the specialized stimulus selectivity of dorsal and ventral PFC neurons. Inactivation of dorsal prefrontal areas, including the frontal eye field, decreases performance during spatial WM tasks (Sommer and Tehovnik 1997, Dias and Segraves 1999, Chafee and Goldman-Rakic 2000, Suzuki and Gottlieb 2013, Noudoost, Clark et al. 2014). Similarly, limited injections of muscimol in the dorsal PFC produce spatial WM deficits that localize in the contralateral visual field. This "mnemonic scotoma" effect is also produced by small, focal lesions (Funahashi, Bruce et al. 1993, Sawaguchi and Iba 2001).

Inactivation of a dorsal area, the frontal eye field, has negligible effects on object WM (Clark, Noudoost et al. 2014), even when monkeys are tested with objects for which frontal eye field neurons were shown to exhibit broad but significant selectivity in the same object-WM task (Clark, Noudoost et al. 2012). In contrast, inactivation of the ventral PFC impairs the ability to select objects based on their remembered features but not on their spatial location (Bichot, Heard et al. 2015). Location along the anterior-posterior axis is also critical for the effects of lesions. Lesions in the anterior aspect of the ventral PFC (area 47/12) that failed to produce deficits of feature WM (Rushworth, Nixon et al. 1997), whereas lesions in the posterior that succeeded (Bichot, Heard et al. 2015).

Although this discussion emphasizes spatial and object WM, we do not wish to suggest that these are the only functions of the dorsal and ventral PFC, respectively. The representation of task rules and associations reviewed in the

neurophysiological studies of the previous section fares prominently on the functional consequences of lesion studies. In this case too, the effects of lesions are dissociable between dorsal and ventral PFC. Thus, lesions in the posterior-dorsal PFC impair performing tasks relying on associations (e.g., green light means press left button, red light means press right), even after the basic rule of the task has been acquired (Petrides 2005). Lesions of the mid-dorsal PFC impair tasks such as the delayed nonmatch to sample, involving presentation of a stimulus and after a delay period, the same stimulus plus a new one, requiring selection of the newly added item. The deficit is more pronounced if the nonmatch stimulus is one the monkey is familiar with (Petrides 2005). Damage to the ventral PFC does not produce impairments of performing tasks requiring recognition or simple recall; its effects become apparent in tasks requiring active retrieval of information from memory, such as free-recall tasks (Petrides 1996). The ventral PFC also appears to be essential for learning a task by trial and error, and reversal learning, which requires learning new associations within a session (Rushworth, Nixon et al. 1997, Buckley, Mansouri et al. 2009, Rygula, Walker et al. 2010).

Neural Correlates of Working Memory of Other Modalities

Although visual WM is arguably the most popular modality in current research, the examination of additional sensory modalities is also highly instructive. In particular, the comparison between modalities reveals common themes on the neural basis of WM and informs the debates on the mechanisms of its generation and its localization in the brain. We will examine primarily somatosensory and auditory WM, since these are the two modalities most studied in nonhuman primates; reports of olfactory and gustatory WM will be briefly reviewed in the last part of this section.

Somatosensory Working Memory

Somatosensory information is processed by the primary somatosensory cortex (SI), which is situated in the postcentral gyrus. This information is then relayed to the secondary somatosensory cortex (SII) and higher-order somatosensory areas including areas 5 and 7b in the parietal lobe and the insular cortex (Ig), before ultimately being projected to the frontal lobe. The ventral rim of prefrontal area 46 is anatomically connected with parietal lobe regions, with the strongest connections stemming from area 7b (Cavada and Goldman-Rakic 1989). Two paradigms have been popular in the study of somatosensory WM. In one series of studies, subjects were required to remember the frequency of a vibratory stimulus and to determine if a second vibratory stimulus was of higher or lower frequency (as in Figure 8, bottom

panel). Neurons in the prefrontal cortex and other frontal areas (e.g., premotor and presupplementary motor area) have been shown to remain active in the delay period after the presentation of the first vibratory stimulus, and furthermore, their firing rate is graded, depending on the frequency of the stimulus (Romo, Brody et al. 1999). Similar to what has been observed in visual WM neurons, persistent somatosensory activity typically exhibits dynamics, with activity ramping up or down toward the end of the delay period, but vibratory stimuli are still represented in a stable subspace (Hernandez, Zainos et al. 2002, Brody, Hernandez et al. 2003, Wang, Li et al. 2015, Murray, Bernacchia et al. 2017). Frequency-selective persistent activity has also been described for cross-modal tasks (as in Figure 8), where the subject has to maintain in WM the frequency of a tactile stimulus and compare it with that of an auditory one (Constantinidis 2016, Vergara, Rivera et al. 2016).

In a second popular paradigm to study somatosensory WM, subjects are presented with a rod, out of view, whose surface is an embossed grating or other patterned texture, and the subjects have to maintain this tactile information in WM in order to determine whether a subsequent stimulus presented in the same fashion is the same or different.

Neurons in the PFC remain active after the presentation of such stimuli, and individual neurons exhibit selectivity for different patterns (Wang, Li et al. 2015). In this case too, subjects can perform cross-modal judgments, for example,

comparing the orientation of the tactile stimulus with that of a visual one (Wang, Li et al. 2015).

These experiments reinforce some of the ideas discussed in the context of the role of persistent activity in visual WM. Individual neurons readily generate persistent activity, which is specific for the properties of the stimuli maintained in memory and not related to motor preparation (which cannot be planned at the time of the first stimulus presentation) or their spatial location (which never varies in the somatosensory paradigms mentioned), so as to instantiate top-down control.

The localization of somatosensory WM has also been debated. Beyond the frontal lobe, haptic-selective persistent activity has also been demonstrated unequivocally in brain areas directly connected with the PFC, most importantly the posterior parietal cortex (Koch and Fuster 1989). In sensory cortical areas, results have been mixed. Tested with the vibratory WM paradigm, neurons in SII have been shown to discharge for a brief period after the offset of the stimulus, though these responses quickly decay (Romo, Hernandez et al. 2002). Frequency-selective vibratory information was not found in SI in these experiments (Romo and Salinas 2003). On the other hand, a small population of SI neurons exhibit sustained haptic-selective activity over the entire delay period for textures of remembered objects (Zhou and Fuster 1996). We may, therefore conclude that, in analogy to the visual cortex, primate somatosensory areas are not entirely uncoupled to working memory. Persistent activity driven by somatosensory stimuli is more prevalent and more robust in the prefrontal cortex, though traces of persistent activity can be

found under some conditions as early as in the primary somatosensory cortex. It appears therefore that a gradient of importance exists, with frontal areas simply being the most crucial for stimulus maintenance, being active across conditions, and enabling judgments such as cross-modal comparisons of remembered stimuli.

Auditory Working Memory

The auditory cortex is organized in a central primary auditory cortex core (A1) that lines the central bank of the Sylvian fissure and is surrounded by an area of higher-order auditory cortex, known as the belt cortex (A2). A tertiary auditory area, the parabelt cortex (A3), is located lateral to this belt. In analogy to the segregation of object and spatial information in the visual pathway that extends into the PFC, the rostral and caudal belt have been identified as processing pathways for the identity and spatial location of auditory stimuli, respectively, and their projections have been shown to terminate in discrete subdivisions of the frontal lobe (Jones, Dell'Anna et al. 1995, Rauschecker, Tian et al. 1995, Kosaki, Hashikawa et al. 1997, Kaas and Hackett 2000). Specifically, the rostral belt projects to the inferior convexity (areas 12 and 45), a region that also receives visual object information from the inferior temporal cortex, whereas the caudal belt cortex targets the caudal aspect of area 46 and area 8, a region that also receives visuospatial information from the posterior parietal cortex (Hackett, Stepniewska et al. 1999, Romanski, Bates et al. 1999, Romanski, Tian et al. 1999).

Neurons in the PFC generate persistent discharges in auditory WM tasks that represent remembered sound features, such as the identity of a monkey vocalization, or the spatial location of the sound source (Bodner, Kroger et al. 1996, Fuster, Bodner et al. 2000, Romanski and Goldman-Rakic 2002, Rama, Poremba et al. 2004, Plakke, Ng et al. 2013). Auditory information that is maintained in memory and generates persistent activity allows cross-modal judgments as well (Fuster, Bodner et al. 2000). Persistent activity representing the identity of multi-sensory, audio-visual stimuli has also been described (Sugihara, Diltz et al. 2006). In analogy to visual WM, persistent activity in auditory WM tasks also encodes more abstract information, for example, categories of sound stimuli (Cohen, Hauser et al. 2006). Activity-silent mechanisms of auditory WM have also been identified in the PFC, most importantly differential responses to the same stimulus when it appeared as a match or a nonmatch in WM task (Hwang and Romanski 2015).

The localization of auditory memory in the brain has also received substantial research attention. Beyond the PFC, persistent activity for auditory stimuli has been described in the superior temporal cortex (Scott, Mishkin et al. 2014). The primary auditory cortex also seems to possess a small population with stimulus-selective persistent activity (Gottlieb, Vaadia et al. 1989, Bigelow, Rossi et al. 2014) (Scott and Mishkin 2016), though the result has not been confirmed across all studies; in some auditory WM paradigms, persistent activity was not detected in the primary cortex (Lemus, Hernandez et al. 2009). Strong evidence exists for a causal role of the PFC in the maintenance of WM; reversible inactivation

experiments have revealed profound impairments in auditory and audio-visual WM tasks (Plakke, Hwang et al. 2015).

Gustatory and Olfactory Pathways

Gustatory cortical areas include the primary gustatory cortex in the postcentral gyrus and secondary areas such as the insula and precentral opercular areas. These areas largely project to areas 12 and 13 of the orbitofrontal cortex (Cavada, Company et al. 2000). The same orbital frontal areas are also targeted by the olfactory (pyriform) cortex, while the entorhinal cortex receives input directly from the olfactory bulb (Carmichael, Clugnet et al. 1994). Although few neurophysiological studies have employed gustatory and olfactory cues for tasks that engage WM in primates, there have been reports of persistent responses for gustatory stimuli, primarily in the secondary gustatory cortex and in the orbitofrontal cortex (Ifuku, Hirata et al. 2003, Lara, Kennerley et al. 2009). In one set of experiments, neurons in these areas sustained firing while the subject had to remember the identity of the gustatory stimulus (water or salt) in order to plan a response. Such sustained responses were absent from the primary gustatory cortex, the neurons of which encoded gustatory information during the stimulus presentation (Ifuku, Hirata et al. 2003). In another set of experiments, neurons in the orbitofrontal and gustatory cortex generated persistent discharges representing the identity of gustatory stimuli (different types of fruit juices) in the context of a match/nonmatch task. Some neurons with persistent gustatory discharges were

also encountered in the dorsolateral and ventrolateral PFC, though these were less frequent (Lara, Kennerley et al. 2009).

Working Memory Capacity

Limited capacity is a defining feature of WM and is widely viewed as a predictor of general cognitive ability (Luck and Vogel 1997, Colflesh and Conway 2007, Fukuda, Awh et al. 2010, Constantinidis and Klingberg 2016, Unsworth and Robison 2020). The limits of capacity are not exclusive to WM storage alone but also include the amount of perceptual information that we can process at any given time, traditionally exemplified in change blindness demonstrations (Simons and Chabris 1999). Further studies have suggested that this apparently blindness may be due to unattended objects being stored as ensemble statistics, where multiple, single item measurements are combined to form a singular group (Freeman and Simoncelli 2011). This would imply that only attended objects are directly stored in WM capacity, while all others would be stored as single or multiple ensemble statistic sets (Cohen, Dennett et al. 2016). We examine behavioral correlates of WM capacity first, and conceptual models that have been introduced to account for them, before we delve into the neural correlates that underlie the storage of information in WM and its limitations.

Behavioral Correlates of WM Capacity

Experimental tasks devised to quantify how many items a subject can maintain in memory often rely on “change detection”: subjects are typically presented with a set number of items and after a delay period they are asked to report if one of the items presented for a second time was identical to the first (single probe test), or if a

whole display presented for a second time is identical to the first. In a set of experiments that used a change detection task to present arrays of colored squares, oriented lines, variously sized lines, and other similar stimuli, participants were generally found to maintain up to three distinct items without difficulty or error (Cowan 2001). However, any arrays that demanded capacity for additional items beyond this point seemed to cause a systematic decrease in performance, demonstrating the extreme limitation of human WM capacity.

Analytical models of capacity have been developed that can provide estimates of the items a subject can maintain in memory correctly. Two popular examples include the Pashler and Cowan formulas for WM capacity, which both assume that WM is carried out through a limited number of discrete units in the change detection paradigm (Pashler 1988, Cowan 2001):

$$\text{Pashler: } k = N\left(\frac{h-f}{1-f}\right)$$

and

$$\text{Cowan: } k = N(h-f)$$

In both formulas, variable "h" represents the number of hits in the change detection paradigm, variable "f" represents the false alarm rate, variable "N" represents the set size of remembered material that is demanded by the paradigm, while variable "k" represents the estimated measure of capacity. Cowan's formula applies primarily to single-probe experiments of change detection; Pashler's formula is most applicable in whole-display paradigms (Rouder, Morey et al. 2011). A caveat is that these models predict perfect performance for any set size that is smaller than

capacity—an assumption that it is almost always incorrect due to the fact that occasional errors may occur at any set size due to lapses, for example, correctly recalling but misclicking on the task computer, or circumstantial distractions of external stimuli or internal thoughts.

The definition of the units of capacity is also tricky, as it is well known that WM may actually be carried out as the integration of multiple items into "chunks" (e.g., groups of digits). Alternatively, it is also possible that all features may be stored as a continuous ensemble, rather than being "chunked" into discrete units (further discussed below). A "unit" may also include multiple features of the same object: If the task requires storage of color and orientation simultaneously, in a conjunction task, performance does not decrease over maintenance of one feature at a time (Luck and Vogel 1997). On the other hand, spatial locations of objects seem to be stored more reliably and automatically in WM, even if they are not task relevant (Foster, Bsales et al. 2017). The estimated capacity may also dependent on the complexity of the objects stored in memory and the relative similarity of items in the display (Eng, Chen et al. 2005, Schurgin, Wixted et al. 2020).

Discrete Slots vs. Continuous Resource

Whether WM capacity is bound as a set number of discrete units—typically referred to as "slots"—or instead, if capacity is a flexible resource that allocates items continuously, and therefore, without a limit fixed to a single number is a matter of debate (Luck and Vogel 1997, Luria, Sessa et al. 2010). The discrete units model can

be traced to early models of WM capacity, introduced by Miller in 1956 as the "magic number seven" for the number of objects that could be held in WM at any given time, although more recent estimates suggest an even lower number of four or so objects (Cowan 2001). Evidence in favor of the former model comes from color recall tasks in which subjects indicate the closest shade on a spectrum wheel. The error distributions are consistent with what would be expected from discrete units, with each additional item in WM load decreasing the probability density of correct responses until capacity is reached, and performance completely collapses (Fukuda, Awh et al. 2010, Luck and Vogel 2013). Moreover, beyond the behavioral evidence, EEG recordings also revealed a steady increase in mean amplitude for each item that was added to capacity, which would eventually plateau when capacity was reached (Luck and Vogel 2013).

Flexible resource models point out the inverse relationship between WM capacity and object complexity. An extreme view of the discrete units model would predict that the number of items that can be stored in capacity would remain constant regardless of the complexity of the stored objects. However, this has not been experimentally supported. The addition of new features or objects increases the net complexity and decreases apparent WM capacity accordingly (Bays and Husain 2008, Luria, Sessa et al. 2010). Results of the error distributions in color recall tasks have also been revisited. The Target Confusability Competition model suggested that errors are primarily driven by the similarities between alternatives (Schurgin, Wixted et al. 2020). Behavioral evidence from a color detection task

indicated that stored objects with greater similarity to the target seem to receive a boost in their likelihood to be selected, thus providing a possible explanation for the abrupt collapse in performance that occurs when capacity is reached (Schurgin, Wixted et al. 2020).

For both discrete units and flexible resource models, there seems to be a link between attention and WM capacity, as prior studies have suggested that individuals who possess a higher WM capacity may be able to control and maintain the intensity of their attention on a specific object more effectively than individuals with low WM capacity (Colflesh and Conway 2007, Unsworth and Robison 2020). However, despite the general rule of WM capacity decreasing as remembered objects become more complex, not all types of WM seem to be affected equally in this manner. A critical example can be observed in a recent study that applied a recall task of increasing load to examine the effects of increasing task complexity on WM capacity, ultimately finding that verbal WM was not affected (Doherty and Logie 2016). As a result, verbal WM would seem to be more aligned with the expected results of the discrete unit models of WM capacity, despite these models being opposed by other studies, such as the color change detection task of Bays et al (Bays and Husain 2008).

Electrophysiological Correlates of WM Capacity

The number of items stored in WM manifests themselves in EEG activity that persists throughout the entire period in which the content is maintained in WM

(Vogel and Machizawa 2004, Sprague, Ester et al. 2014, Bae and Luck 2018, Bae and Luck 2019). The contralateral delay activity (CDA), a negative-voltage wave over posterior contralateral electrodes increases monotonically rising in amplitude as the more items are added (Figure 9), eventually flatting into an asymptote when the subject individuals reached their respective WM capacity limits (Vogel and Machizawa 2004). In fMRI studies, the posterior parietal cortex, particularly in the intraparietal sulcus, seems to be particularly important, as the levels of observed BOLD activation in this area best tracks WM capacity (Todd and Marois 2004). Interpretation of these macroscopic measures of brain activity hinges on the fundamental mechanisms of WM maintenance. Increased content may simply be more susceptible to drifts in activity, without reducing the activity of the neural ensembles that represent each given object (Schneegans and Bays 2018). Alternative models of WM capacity rely on the number of EEG cycles that can be simultaneously maintained, with each cycle representing a different object, as each cycle would representing the firing of a different neural ensemble (Lisman and Idiart 1995, Raffone and Wolters 2001).

Nonhuman primates are capable of mastering multiple-item WM tasks (Buschman, Siegel et al. 2011, Heyselaar, Johnston et al. 2011, Lara and Wallis 2012, Panichello, DePasquale et al. 2019) and it has been possible to investigate the representation of multiple-stimulus information in memory by single neurons (Warden and Miller 2007, Buschman, Siegel et al. 2011, Lara and Wallis 2012). In the context of the bump attractor model (Figure 10), the activity representing each

of multiple stored items held in memory can be thought of as a separate bump (Edin, Klingberg et al. 2009). As more items are added to memory, the total population activity would be expected to increase. However, when the capacity of the system is exceeded, the persistent activity representing some of the stimuli decays and the item cannot be recalled at the end of the delay period. In practice, experimental studies have not produced such a clear picture of the neural correlates of capacity. Prefrontal neurons have large receptive fields, and firing rate of individual neurons quickly saturates when more than a single item is maintained in memory, as does the information that can be decoded from their activity (Buschman, Siegel et al. 2011). Across the population of neurons, activity in the delay period does increase as additional items are stored in memory (Figure 10), by virtue of more neurons being recruited (Tang, Qi et al. 2019). However, responses to multiple item displays are not the linear summation of responses to individual items, which complicates the decoding of items held in WM (Lara and Wallis 2014).

Working Memory Duration

As with capacity, WM is defined, in part, by its limited duration, which is commonly believed to span along a range of seconds. Certainly, there is a significant separation between this, and the hours-to-decades long storage duration that has been observed to occur in LTM (Brooks 1975, Goelet, Castellucci et al. 1986, Sweller 2016, Pertzov, Manohar et al. 2017). However, there is a great deal of debate in regard to where one form of memory begins and the other ends. We will therefore attempt to reinitiate this line of inquiry by examining the neural correlates of WM duration across the visual, auditory, and tactile modalities, the effects of rehearsal, and the underlying nature of WM duration in regard to whether this property is limited by the decay or interference of stored information.

Distinguishing Maintained WM Duration from LTM

Due to the ambiguity in separating WM duration from LTM, many researchers have suggested that at least certain aspects of WM duration may actually be activated LTM (Ranganath, Johnson et al. 2003, Ruchkin, Grafman et al. 2003, Cowan 2005). This hypothesis is supported by the wide degree of overlap between the brain areas that are involved with LTM compared to the brain areas that are involved WM, particularly in the prefrontal cortex, and in the case of auditory WM, the hippocampus (Ranganath, Johnson et al. 2003, Kumar, Joseph et al. 2016). Aging also affects these areas simultaneously, resulting in similar deficiencies in both LTM and WM (Chen and Naveh-Benjamin 2012). Moreover, certain models of WM,

such as Baddeley's phonological loop, would require the representation of information in LTM, regardless of whether or not WM is a separate and distinct function (Hulme 1991, Baddeley 2012). However, the activated LTM explanation for WM duration also faces a degree of challenge, particularly in the rapid loss of information that is often observed in WM, compared to the extremely long lasting information in LTM (Pertzov, Manohar et al. 2017). If WM were truly activated LTM, we would expect that subjects would be able to retrieve information previously discarded from WM, even a significant period of time later. We would also expect that the impairment of LTM would result in the impairment of WM, though this was also refuted by prior examinations on the effects of head trauma (Brooks 1975).

Furthermore, despite being involved with similar brain areas, LTM and WM seem to rely on very different molecular mechanisms (Goelet, Castellucci et al. 1986, Dash, Moore et al. 2007). LTM, for example, requires the synthesis of new proteins, while WM can persist exclusively through the sustained neuronal activity of specific neuronal ensembles (Goelet, Castellucci et al. 1986, Qi, Meyer et al. 2012, Rigotti, Barak et al. 2013, Constantinidis and Klingberg 2016). Certain brain areas such as the hippocampus, also seem to play a much greater role in LTM compared to WM, ultimately implying that although WM likely shares a degree of overlap with LTM, it is unlikely that these functions can be considered the same (Lisman and Grace 2005). WM duration, along with capacity, thus represents key variables for demarcating the two memory mechanisms.

WM Duration for different modalities of stimuli

The limits of WM duration are exemplified by the progressive decrease in performance over time, regardless of modality. Performance is thought to be dependent on the activity of the brain areas maintaining each type of information. In auditory tasks, for example, the maintenance of single tones was correlated with sustained activity in the interconnected areas of the auditory cortex, the inferior frontal gyrus, and the hippocampus, among which the former two were also able to distinguish the maintained tone (Kumar, Joseph et al. 2016). Likewise, when tactile WM information was being maintained, the application of transcranial magnetic stimulation was able to demonstrate a causal role for the middle frontal gyrus, but only for areas that possessed anatomical connections with the primary somatosensory cortex (Hannula, Neuvonen et al. 2010, Zhao, Zhou et al. 2018). Importantly however, certain WM task contexts can result in overlap between otherwise modality-specific areas, even if the actual task stimuli are exclusively unimodal (Calvert, Bullmore et al. 1997, Gallace and Spence 2009). An excellent example can be observed with the maintenance of phonological information, such as silent lip movements—a purely visual stimulus that is also able to activate auditory areas when the lip movements resemble the formation of words (Calvert, Bullmore et al. 1997). Certain tasks involving spatial frequency, speed, direction and motion coherence, seemed to undergo only limited decreases in performance when the required duration was extended by a significant degree (Magnussen and Greenlee 1992, Blake, Cepeda et al. 1997, Magnussen, Idas et al. 1998). However,

these opposition points have been criticized for failing to account for the special status of spatial information in visual WM, as well as the possibility of non-articulatory rehearsal via attention, which has been experimentally shown to supplement recall (Cowan and AuBuchon 2008, Foster, Bsales et al. 2017).

The Role of Rehearsal in WM Duration

The maintenance of information in WM is greatly enhanced when conscious attention is devoted to the process, and various forms of rehearsal may be applied to this end. In particular, maintenance rehearsal is typically defined as the repetition of information, with a common example being the recitation of a phone number until the call is dialed, after which point, the number is rapidly forgotten (Gardiner, Gawlik et al. 1994, Oberauer 2019). Maintenance rehearsal results in similar neural activity as the original encoding of information when examined via EEG (Pinal, Zurron et al. 2014). Moreover, it seems possible that rehearsal may be "offloaded" to brain areas beyond the PFC, such as the modality specific areas that were previously noted. These critical insights thus reveal a possible key for explaining the neural mechanisms of rehearsal, and by extension, determining the degree to which this ability may ultimately be responsible for WM duration. On an extreme view, this has even led to the question of whether WM duration actually exists, or instead, if the phenomenon that we refer to as WM duration is actually a product of rehearsal, due to the fact that objects seem to persist indefinitely in WM as long as the rehearsal continues to be maintained (Wixted 1991).

The rehearsal explanation for WM duration is further supported by the fact that the loss of WM content over time does not seem to be an all or nothing event, but instead, can be mitigated by devoting additional cognitive resources to specific objects, as demonstrated by the classic Sperling et al. experiment. Specifically, individuals were briefly (between 15-500 ms) presented with a 3x4 grid of random letters, and then asked to report the letters at each location (Sperling 1960). The average results for this first experimental condition ranged between four or five accurate letters. However, in a second experimental condition, participants were only required to report a single row of letters, indicated by a tone that would sound 300 ms after the array disappeared. The average results for this second experimental condition ranged between three or four accurate letters, double the per-row average of the original experimental condition, which implied that neural representations of most letters were still available after 300 ms, and that they could survive longer once attention consciously directed to exclusively maintaining the prompted row, while allowing all others to rapidly fade (Sperling 1960).

Decay, Interference, and Drift Models of WM Duration

The cause of the WM duration limits has been debated. It has been theorized that either spontaneous decay of sustained neuronal activity, interference and eventual displacement by new items, or gradual drifting in the activity of the neural ensembles that represent each remembered object may be the mechanism of eventual loss of the information maintained (Portrat, Barrouillet et al. 2008,

Constantinidis and Klingberg 2016, Schneegans and Bays 2018). For example, several studies found that distractors cause greater decreases in WM performance when they match the modality or category of the task relevant information (Wickelgren 1965, Elliott and Strawhorn 1976, Proctor 1978). The implication is that the greater overlap between the task relevant information and the distractor leads to greater interference, and subsequently, greater reductions in performance. In turn, this idea would imply that WM is maintained as long as the relevant coding remains intact, however, the maintenance requires increasing cognitive resources over time. As a result, old content is gradually replaced by the new, unless attention is applied through abilities like rehearsal, to reserve cognitive resources on specific objects over time.

A challenge in neurophysiological studies that could chart neuronal activity as a function of visual working memory duration has been that it is difficult to avoid eye movements (and blinks) over prolonged delays, and neural activity driven by eye movements and retinal input may confound the interpretation of activity relating to the contents of working memory. However, some studies have succeeded (e.g. by allowing temporary timeouts in fixation) and they provided little evidence for decay of persistent activity during the duration of the trial (Balan and Ferrera 2003). Neurophysiological studies of auditory and tactile working memory which are not subject to this limitation have similarly provided little evidence of decay, at least for intervals of ~6 seconds (Romo, Brody et al. 1999, Plakke, Ng et al. 2013). It is only at delay intervals of up to 15 seconds that activity of some individual prefrontal

neurons begins to shut down, though other neurons are able to maintain persistent activity (Papadimitriou, Holmes et al. 2021). Even in these experiments, LFP recordings maintain their tuning for the preceding stimulus all the way up to the longest intervals tested, of approximately 15 seconds (Holmes, Papadimitriou et al. 2018).

In the context of the bump attractor model, strong evidence exists that inaccuracies in recall are the effects of drift rather than decay of activity in the network (Fig. 11), at least for short periods of delay, of 1-3 seconds. Although it is well established that the accuracy of recalling the location of a stimulus decreases monotonically as a function of time (White, Sparks et al. 1994), this does not necessarily imply decay of persistent activity. Instead, longer delays allow the bump of activity to drift and deviate more from the remembered location, on average. Critical predictions of the alternative, decay model, are not borne by the experimental data (Wimmer, Nykamp et al. 2014).

Working Memory across Species

Working memory is not exclusive to primates. Commonalities and differences between the WM mechanisms across and between different species reveals the roles of various neural mechanisms and their relative importance to various WM functions. For example, in virtually all animal models, the development of WM seems to be intertwined with additional cognitive systems, including the motor, sensory and attentional functions (D'Esposito, Ballard et al. 2000, Lowet, Gomes et al. 2018). Substantial challenges also exist. In the broader sense, the incompleteness and only vague definition of psychological terms used to explain the neural basis of cognitive function comes into sharper relief the further away we wander in the evolutionary scale (Buzsaki 2020). The homology or even existence of areas such as the prefrontal cortex across species is also debated (Laubach, Amarante et al. 2018). Here we will review the neurophysiology of WM across the rodent and avian models, focusing particularly on the varying capabilities of each taxonomic class, the role of persistent activity, and the cortical areas involved in it.

Rodent Models

The rodent model of WM, including mice and rats, is one of the most commonly applied models in current research. Rodents can readily be trained to perform working memory tasks, including spatial working memory tasks, based on the radial maze, delayed response and delayed alternation tasks. Moreover, working memory tasks based on other sensory modalities, including olfactory, auditory, and tactile

working memory have also been demonstrated. These paradigms have been used to address some of the same questions that have been raised in primate and human research, including the neural basis of working memory, where in the brain it is localized, the local and long-distance circuits that are essential for its maintenance and the limits of its capacity and duration. Some of the same debates are also present in the rodent literature.

Persistent activity has been reported across working memory tasks. Initial reports documented persistent activity in delayed response tasks (Erlich, Bialek et al. 2011), which might have been confounded with motor preparation, but newer studies have demonstrated persistent activity in a variety of tasks representing remembered features of stimuli across various modalities, e.g. auditory or tactile frequency (Kamigaki and Dan 2017, Esmaeili and Diamond 2019). In contrast to the primate persistent activity, however, individual rodent cortex neurons are not active through the entire delay period of WM tasks, but different groups of neurons become active for brief periods of time (Fig. 12), so that their activity "tiles" the entire delay period (Scott, Constantinople et al. 2017). Information may be encoded by neurons that do not exhibit persistent activity, and by different neurons that become active during the course of learning a behavioral task (Baeg, Kim et al. 2003).

As has been speculated in the primate model, different types of interneurons play a specialized role in the circuits that maintain working Memory. Optogenetic inhibition of different types of interneurons produce different results in WM

performance: inactivation of parvalbumin and somatostatin interneurons degrade working memory performance, whereas inactivation of VIP neurons may improve it (Kamigaki and Dan 2017). Similar to the primate brain, persistent activity generated in the frontal lobe is also dependent upon long-range circuits which includes the mediodorsal thalamic nucleus (Bolkan, Stujenske et al. 2017). MD inactivation degrades performance and activity in the frontal lobe.

As has also been discussed for the primate brain, working memory activity is not limited only to the rodent prefrontal cortex. Areas have also been implicated as major factors in WM, including the posterior parietal cortex, the frontal orienting fields, the hippocampus, and the motor cortex—particularly in the anterolateral area (Hampson, Simeral et al. 1999, Erlich, Bialek et al. 2011, Guo, Yamawaki et al. 2018, Lyamzin and Benucci 2019). These areas appear to be specialized in WM, and the frontal orienting fields, for example, have demonstrated selectivity to orientation in 37% of their population, with performance decreasing when these areas (Hampson, Simeral et al. 1999) are deactivated (Erlich, Bialek et al. 2011). Similarly, the hippocampus—typically associated with LTM—has demonstrated significant selectivity for spatial WM tasks.

Important differences in functions of prefrontal areas between rodent and primate species have also been noted. For example, lesions of the medial rodent PFC were reported to impair attention but not WM (Kahn, Ward et al. 2012) and to not affect task performance in a two-choice serial reaction time task, though other studies reported impairments following medial PFC lesions when the number of

possible targets in the serial reaction time task are increased (Muir, Everitt et al. 1996, Granon, Hardouin et al. 1998). The prospect of reconciling the conflicting results of these prior investigations is complicated by the lack of consensus regarding the anatomy of the rodent PFC and the question of whether the rodent and human PFC are truly homologous (Preuss 1995, Laubach, Amarante et al. 2018). In particular, a recent meta-analysis has suggested that although medial prefrontal and cingulate areas are present in the rodent prefrontal cortex, the lateral prefrontal cortex, which is the focal area of persistent activity generation, has no clear homolog in the rodent brain (Laubach, Amarante et al. 2018). It is therefore possible that unique WM functions evolved along with unique PFC areas in the primate.

Avian Models

Many avian species—particularly corvids—have demonstrated high levels of WM, including a capacity for storing up to four distinct items and the ability to perform highly complex tasks (Veit and Nieder 2013, Balakhonov and Rose 2017). However, they represent a unique model in WM, as their brains are highly dissimilar to the mammalian brain, particularly in terms of the structures typically associated with higher functions such as WW (Emery and Clayton 2005, Gunturkun 2005). Notably, the avian brain seems to lack a PFC, neocortex or any laminated component within its telencephalon, despite possessing a similar structure of modular networks (Reiner, Perkel et al. 2004, Emery and Clayton 2005, Kirsch, Gunturkun et al. 2008,

Shanahan, Bingman et al. 2013). This implies convergent evolution, with these two taxonomic classes taking separate developmental paths to independently pursue the same functions (Emery and Clayton 2004, Butler and Cotterill 2006, Balakhonov and Rose 2017).

The caudolateral nidopallium has been identified as an area important for working memory, demonstrating persistent activity when information is maintained in WM through a delay period in a wide variety of tasks. For example, during a spatial delayed-response task, at least 30% of the caudolateral corvid nidopallium demonstrated selectivity to both visual presentation and the maintenance of items during the delay period, thus resulting in decodable representations that remained stable throughout the course of the task (Rinnert, Kirschhock et al. 2019). Persistent discharges have thus been described that can maintain visual spatial location and object identity, auditory stimuli, numerical quantity, and abstract task rules (Kalt, Diekamp et al. 1999, Diekamp, Kalt et al. 2002, Veit, Hartmann et al. 2014). The parallel between the caudolateral nidopallium and the PFC is further supported by the fact that both areas are central hubs for motor and sensory activity, the development of which are known to be intertwined with WM (Kroner and Gunturkun 1999, D'Esposito, Ballard et al. 2000, Lowet, Gomes et al. 2018). The direct causality of the caudolateral nidopallium in WM is suggested by a study where the controlled lesioning of the caudolateral nidopallium was found to generate WM deficits that were proportional to the size of the lesion (Diekamp, Gagliardo et al. 2002). Recently however, these results have been questioned in certain avian

species, such as pigeons, due to the fact that persistent neural activity has not been observed in all studies of the caudolateral nidopallium, and may be modulated by factors, such as the presence of a task reward (Johnston, Anderson et al. 2017, Johnston, Porter et al. 2019). This may imply that the caudolateral nidopallium may only be the seat of certain aspects of WM. However, the wide range of cognitive abilities between different avian species may also result in some species simply lacking the ability to effectively conduct certain types of WM tasks.

Working Memory Development

Humans and nonhuman primates experience a protracted period of cognitive development that parallels the maturation of the prefrontal cortex (Yakovlev and Lecours 1967, Chugani, Phelps et al. 1987, Jernigan, Archibald et al. 1991, Pfefferbaum, Mathalon et al. 1994, Sowell, Delis et al. 2001). Working memory ability, in particular, undergoes significant improvements that are one of the hallmarks of human cognitive development (Fry and Hale 2000, Gathercole, Pickering et al. 2004, Luna, Garver et al. 2004, Crone, Wendelken et al. 2006, Spronk, Vogel et al. 2012, Isbell, Fukuda et al. 2015). A monotonic age-based increase in the performance of multiple types of working memory tasks has been reported, including improvement in the mean accuracy and reaction time of memory-guided saccade tasks (Fry and Hale 2000, Gathercole, Pickering et al. 2004, Luna, Garver et al. 2004). Tasks that assess verbal and visuo-spatial working memory and executive control show parallel developmental profiles (Gathercole, Pickering et al. 2004, Alloway, Gathercole et al. 2006). Developmental improvements in WM are not only characterized by an increase in correct trials, but also by a decrease in performance variability from trial to trial, which mirrors decreases in the variability of whole brain states of task-related activity (Montez, Calabro et al. 2017). On the other hand, neurodevelopmental disorders that manifest in childhood and late adolescence, such as ADHD and schizophrenia, result in corresponding deficits of working memory (Park, Holzman et al. 1995,

Pantelis, Barnes et al. 1997, Westerberg, Hirvikoski et al. 2004, Burgess, Depue et al. 2010, Lee, Cowan et al. 2010).

Age-related performance increases observed for working memory tasks are associated with structural changes due to white and gray-matter maturation in the PFC and its connections with other cortical regions in humans (Bunge, Dudukovic et al. 2002, Olesen, Nagy et al. 2003). Human fMRI studies have also shown that patterns of brain activation associated with the performance of working memory tasks undergo distinct changes between childhood and adulthood, typically involving increased activation of frontal and parietal areas, but also more effective deactivation of other areas, such as the default mode network (Klingberg, Forssberg et al. 2002, Kwon, Reiss et al. 2002, Burgund, Lugar et al. 2006, Olesen, Macoveanu et al. 2007, Simmonds, Hallquist et al. 2017).

The nonhuman primate model has been particularly fruitful for understanding working memory maturation. The male rhesus monkey (*Macaca mulatta)* enters puberty at approximately 3.5 years of age and reaches full sexual maturity at 5 (Plant, Ramaswamy et al. 2005, Herman, Zehr et al. 2006), roughly equivalent to human ages of 11 and 16 years, respectively thus making this a particularly well-suited model to inform human neurocognitive development. Similar to the human pattern of development, the monkey prefrontal cortex continues to mature during adolescence and into early adulthood (Lewis 1997, Fuster 2002, Malkova, Heuer et al. 2006), establishing connections with a network of brain areas, similar to that of the human PFC. Biochemical and anatomical

changes have also been characterized within the monkey prefrontal cortex from pre-puberty to adulthood, including maturation of interneuron morphology and connections (Lewis 1997, Hoftman and Lewis 2011).

Behavioral performance and neural activity in working memory tasks change markedly around the time of puberty (Fig. 13), a developmental event associated with the release of sex hormones and significant neurological change (Zhou, Zhu et al. 2013, Zhou, Zhu et al. 2016). Performance of working memory tasks is subtly but significantly higher in adult monkeys compared to adolescent monkeys that have entered puberty. The improvement involves both increases in mean performance and decreases in variability, similar to the improvements observed in humans between these two developmental stages (Montez, Calabro et al. 2019). Most importantly, persistent activity during the delay intervals of working memory tasks is higher in adult animals than adolescent ones. Even when comparing persistent activity from adolescent and adult monkeys obtained in sessions equated for performance, the adult prefrontal cortex is better able to generate persistent activity (Zhou, Zhu et al. 2016). Furthermore, the activity that follows presentation of distractors—stimuli that the monkeys have been instructed to ignore—is substantially lower in adulthood, suggesting that the adult prefrontal cortex is better able to filter distracting stimuli during working memory. These changes refer specifically to persistent activity. Responses to visual stimuli themselves are relatively unchanged, indicating that stimulus-driven neuronal responses are fully mature by puberty (Zhou, Zhu et al. 2016).

The observed changes in neural activity during childhood, adolescence, and adulthood suggest plasticity in the circuit that generates persistent discharges. An example of this may be observed in the local-circuit differences between adolescent and adult monkeys. Recordings from nearby pairs of neurons (within 0.5 – 1 mm from each other) reveal a higher percentage of spikes in adult than adolescent monkeys that are generated in near synchrony, within 2-3 ms of each other (Zhou, Zhu et al. 2014). This increase in "zero-lag synchronization" with age is the result of weakened inhibitory interactions in adulthood, either as a result of pruning of inhibitory connections, or decreases in the connectivity strength of pyramidal neurons onto interneurons, which lessens the net output of inhibitory connections as the prefrontal cortex matures (Gonzalez-Burgos, Miyamae et al. 2015). Interestingly, a decrease in zero-lag synchrony of prefrontal neurons has been recently implicated in an animal model of schizophrenia (Zick, Blackman et al. 2018), a condition that, among other pathological symptoms, compromises working memory.

Changes in neuronal activity and connectivity that are observed in aged monkeys somewhat reverse the changes that are seen between adolescence and adulthood. Most importantly, persistent activity during working memory tasks is substantially lower in old monkeys, though in this case too, firing rate during the stimulus presentation differed little between age groups (Wang, Gamo et al. 2011). Aged animals exhibit elevated cyclic-AMP (cAMP) signaling, which reduces persistent activity by opening Hyperpolarization-activated Cyclic Nucleotide–gated

channels (HCN - nonselective voltage-gated cation channels), and KCNQ, (Potassium voltage-gated channels). Persistent activity can be partially restored to more youthful levels by inhibiting cAMP signaling, or by blocking HCN or KCNQ channels (Wang, Gamo et al. 2011).

Working Memory Training

Working memory training may induce significant performance improvements at all stages of life, including adulthood (Constantinidis and Klingberg 2016). Computerized training has resulted in expanded WM capacity, as well as clinical improvements for patients of stroke, ADHD, schizophrenia (Klingberg, Forssberg et al. 2002, Westerberg, Jacobaeus et al. 2007, Jaeggi, Buschkuehl et al. 2008, Subramaniam, Luks et al. 2012) as well as aged adults (Anguera, Boccanfuso et al. 2013). However, there is a heated debate in regards to if and how these improvements may generalize ("transfer") to tasks that were not part of the training, including more complex day to day activities (Cortese, Ferrin et al. 2015, Schwaighofer, Fischer et al. 2015). The view against the benefits of training argues that improved performance levels are actually the result of task-specific skills instead of genuinely expanded abilities (Owen, Hampshire et al. 2010). Imaging studies have also produced conflicting results about the effects of training, with some studies suggesting increased activity in executive brain areas (Garavan, Kelley et al. 2000, Hempel, Giesel et al. 2004, Olesen, Westerberg et al. 2004, Dahlin, Neely et al. 2008, Subramaniam, Luks et al. 2012, Jolles, van Buchem et al. 2013) and others suggesting decreased activity instead (Schneiders, Opitz et al. 2011, Kuhn, Schmiedek et al. 2013, Schweizer, Grahn et al. 2013, Takeuchi, Taki et al. 2013). Thus increased activity has been suggested to reflect specialized selectivity, while decreases in activity are likewise suggested as an improvement in efficiency

(Constantinidis and Klingberg 2016). To resolve these debates, it is again necessary to understand changes at the level of neural activity as a result of training.

Initial Working Memory Training

The appearance of persistent activity seems to be ubiquitous regardless of the task type, the subject's prior training, or even whether they were required to perform a working memory task, at all. For example, persistent prefrontal activity has been observed in naïve monkeys that passively view stimuli without performing an active task, continuing even after the stimuli are no longer present (Meyer, Qi et al. 2007, Riley, Qi et al. 2017). This evidence—and the everyday ability to track our environment and recall information even when not explicitly prompted to do so ahead of time—would collectively argue against the view that working memory necessarily requires conscious effort. Thus, working memory seems to be an automatic function that is only subject to, but not dependent upon, conscious control. Importantly, the observed persistent activity in naïve monkeys is selective for specific properties of the stimuli, including spatial location, color, and shape (Meyer, Qi et al. 2007, Riley, Qi et al. 2017). There are also limits to the information that may be represented automatically. In particular, the decodable identity of a stimulus would not persist after the presentation of a new stimulus in the experiments discussed above, and the information of whether two successively presented stimuli were the same or not was largely absent in naïve animals (Meyer, Qi et al. 2007, Meyers, Qi et al. 2012, Riley, Qi et al. 2017).

Training of monkeys to perform a working memory task for the first time results in long lasting changes in neural activity (Fig. 14), including an overall increase in persistent activity, in addition to increases in the number of neurons that are activated after appearance of stimuli (Meyer, Qi et al. 2007, Mendoza-Halliday and Martinez-Trujillo 2017, Riley, Qi et al. 2017, Riley, Qi et al. 2018). These changes are evident even when the trained monkeys are tested with passive presentation of stimuli in the same fashion they did prior to training, and are predictably emphasized when they actively participate (Riley, Qi et al. 2017). Moreover, greater training seems to result in greater changes in activity that reflect the task performance levels at each point in time (Qi, Meyer et al. 2011, Tang, Qi et al. 2019).

Prefrontal subdivisions along the anterior-posterior and dorso-ventral axes of the PFC display varying degrees of plasticity after training, as already noted in the section of prefrontal areal specialization. Along the dorsal PFC, the highest levels plasticity occur in more anterior areas, including the mid-, and anterior-dorsal areas (area 46). On the other hand, the posterior dorsal prefrontal areas exhibit limited training effects beyond a small increase in the percentage of posterior-dorsal neurons that are activated by stimuli (Meyer, Qi et al. 2011, Constantinidis and Qi 2018, Riley, Qi et al. 2018). A similar gradient of plasticity runs along the dorso-ventral axis, with most plastic responses observed in the ventral PFC and least plastic in the more medial aspects of the dorsal PFC (Meyer, Qi et al. 2011). Areas along the principal sulcus exhibit intermediate levels of plasticity in terms of increases in persistent discharges after WM training.

Decoding information about remembered stimuli depends not only on their mean firing rate, but also the trial-to-trial response variability, and the positive correlation of firing rates between neurons, all of which are affected by changes in plasticity (Qi and Constantinidis 2012, Qi and Constantinidis 2012, Moreno-Bote, Beck et al. 2014). For example, training decreases the Fano factor of spike counts, a measure of variability, with the greatest decreases observed in neurons that exhibit persistent activity. This implies a refinement of neural selectivity and reliability of stimulus property representations. Similarly, the post training decrease in noise correlation—the spike-count correlation of persistent firing rates between pairs of neurons—also improves information decoding from simultaneously active neurons (Qi and Constantinidis 2012). Task training also alters the time course and dynamics of persistent activity (Kobak, Brendel et al. 2014, Tang, Qi et al. 2019). For example, the firing rates of some neurons are known to "ramp up" or decrease during the course of the trial, in order to encode information dynamically at different time points (Romo, Brody et al. 1999, Meyers, Qi et al. 2012, Stokes, Kusunoki et al. 2013). With training, these changes become more pronounced in trials compared to animals who were only assigned to passive fixation (Kobak, Brendel et al. 2014, Tang, Qi et al. 2019).

Plasticity when learning to perform additional tasks

Subjects that have already trained to perform a cognitive task demonstrate that persistent activity can undergo rapid modification when they are exposed to a new

task condition or type of stimulus (Sarma, Masse et al. 2016). Individual neurons may begin to be activated by stimuli that they were initially unresponsive to over the course of a single training session, when the animals are learning to associate these particular stimuli with motor responses that earn reward, e.g. “look to the left when stimulus A is presented” (Asaad, Rainer et al. 1998). The ability of prefrontal neurons to rapidly represent different aspects of stimuli in persistent activity is also illustrated after training in tasks that require animals to remember one relevant stimulus feature and ignore others (Mante, Sussillo et al. 2013). Dual-task paradigms reveal that persistent activity can rapidly decrease and increase again over the course of a single trial, as the subject alternates their focus between the two tasks (Watanabe and Funahashi 2014).

Training-induced increases in plasticity may also point to the specialization of underlying cellular and molecular mechanisms (Kuboshima-Amemori and Sawaguchi 2007) between specific prefrontal subdivisions. For example, recent studies have demonstrated the systematic variation of plasticity markers between limbic and eulaminate areas, including the relative scarcity of Calcium/calmodulin-dependent protein kinase II (CaMKII) in area 46d compared to more anterior limbic areas (Garcia-Cabezas, Joyce et al. 2017). Markers of cortical stability, including intracortical myelin, perineuronal nets, and parvalbumin have demonstrated the opposite pattern. Changes in neuronal morphology, molecular profiles of the synaptic apparatus, and the influence of neuromodulator systems have also been implicated in long-term prefrontal plasticity (Laroche, Davis et al. 2000, McEwen

and Morrison 2013), and may differ between areas. Further mechanisms for post training plasticity may be connected to observed changes in short-term prefrontal synaptic plasticity, depression, and facilitation (Hempel, Hartman et al. 2000).

The Effects of Training on WM Capacity

Although WM capacity was previously thought to be a fixed quality in individuals, recent research has shown that training can actually enhance WM capacity—particularly in children and young adults—with significant improvements that extend beyond the training-specific tasks (Klingberg, Forssberg et al. 2002, Klingberg, Fernell et al. 2005, Jaeggi, Buschkuehl et al. 2008). This is important for assisting the rehabilitation of individuals with memory related disorders, such as ADHD, and also for improving the cognitive abilities of neurotypical individuals. Moreover, the modification of WM capacity highlights possible neural mechanisms that may mediate capacity.

Firing rate

Training in tasks that require the simultaneous storage of multiple stimuli can induce plastic changes in prefrontal activity, activating more neurons, decreasing baseline firing, and increasing the rate of persistent activity in order to ultimately improve WM capacity—at least in regard to the trained task (Tang, Qi et al. 2019). Training may also occasionally decrease persistent activity, and the reason for this is currently debated as an improvement in efficiency or strategy (Schneiders, Opitz

et al. 2011, Kuhn, Schmiedek et al. 2013, Schweizer, Grahn et al. 2013, Takeuchi, Taki et al. 2013, Constantinidis and Klingberg 2016). It is also possible that change detected in fMRI may be due to decreased baseline activity, particularly when activity is averaged over long periods. Both humans and monkeys are known to develop personalized strategies that assist their performance in complex tasks, e.g. suggestive of grouping of multiple stimuli in memory based on their geometric arrangement (Tang, Riley et al. 2017). These strategies are reflected in the altered prefrontal response selectivity and dynamics when comparing single and multi-stimuli tasks, even if the stimuli sets are the same between both task types (Konecky, Smith et al. 2017, Tang, Qi et al. 2019). In particular, a greater percentage of neural activity from individuals who displayed post-training improvements in multiple stimuli task performance could be explained by "condition-independent" components, not related to any remembered stimuli (Tang, Qi et al. 2019).

Dopamine

A variety of PET studies have investigated the role of dopamine in WM capacity before and after training, finding that changes in D1R density can be mapped to changes in WM capacity in an inverted U graph (McNab, Varrone et al. 2009). This was reinforced by similar results that were observed when applying D1R antagonists (Sawaguchi and Goldman-Rakic 1991). An additional study was conducted to assess D2R density, finding that there was increased dopamine release in individuals who were currently performing a memory task that could be

linked to previously observed increases in BOLD activity after training (Backman, Nyberg et al. 2011). These insights are particularly compelling in light of the abnormal dopamine release that is observed in ADHD and schizophrenia, among other memory-related disorders, as a possible explanation for the associated WM deficits (Arce, Balice-Gordon et al. 2019, Yokokura, Takebasashi et al. 2020).

Functional connectivity

White matter volume and structure has long been associated with WM capacity (Takeuchi, Sekiguchi et al. 2010, Gibson, Purger et al. 2014, Yeung, Zdunek et al. 2014, Bathelt, Gathercole et al. 2018). Structural connectivity—particularly from the PFC to its afferent cortical areas—has previously been correlated with WM capacity during development (Edin, Klingberg et al. 2009, Thompson, Waskom et al. 2016, Bathelt, Gathercole et al. 2018). In fact, the role of functional connectivity does not seem to be limited to development alone. In a study where adults performed n-back WM training, improved performance was correlated with increases in functional connectivity, with the greatest increases occurring between the parietal cortex and a superior frontal region, as well as between the parietal cortex and the dorsolateral PFC (Thompson, Waskom et al. 2016). In the above study on functional connectivity—and many others that attempt to confront the effects of training on WM capacity—it is worth nothing that improved task performance does not necessarily equate to improved capacity. Instead, the observed improvement may be due to improved WM processing or duration, among other factors.

Arguably the most dramatic effects of WM capacity training can be observed in the case of professional mnemonists, who can rapidly store exceptional large quantities of items—usually letters or numerical digits—in their WM. For example, an individual known as SF was presented with random auditory digits at a rate of 1 second per digit, and, after 250 hours of practice, was able to raise his recall limit from 7 digits, to 80 (Chase and Erickson 1982). These results were later replicated by another subject, BB, though BB only received 86 hours of practice and required 5 seconds to store each digit in his WM to achieve maximum capacity (Kliegl 1987). Notably, nearly all professional mnemonists are known to apply various encoding strategies, such as the famous method of loci, where objects are associated with familiar locations, after which the objects may be recalled by imaging the locations (Verhaeghen and Marcoen 1996). These encoding strategies may therefore serve as an excellent method for demonstrating the effects of training on WM capacity in a clearly observable manner. However, the relative lack of research on the topic of encoding strategies may bring their use into question, and among followers of the previously discussed discrete units theory for WM capacity, there is the possibility that encoding strategies are not genuinely increasing WM capacity, but instead, simply serve to more effectively organize chunks of information into each discrete "unit" of an individual's existing WM capacity. Alternatively, it is also possible that the encoding strategies may serve to rapidly store information in long term memory, allowing the recall of LTM to supplant WM without requiring an increase in capacity. Ultimately, these questions can only be resolved by bringing the encoding strategies

of mnemonists back to the forefront of research and continuing the work that was started by the prior generation of neuroscientists.

Conclusions

Recent research has been able to provide a comprehensive link between the activity of neurons in the prefrontal cortex and other brain areas with a range of psychological phenomena related to working memory. The bump attractor model of visuospatial working memory provides a basic framework that links spiking activity with behavior. Synaptic mechanisms allow representation of latent information even in time intervals of no overt spiking activity, but they influence behavior inasmuch as they influence the generation of spiking activity. Limits of capacity and duration can be traced back to the properties of this model, as can improvement of working memory during childhood development and through adult training. Flexible behavior requires considerable neuronal plasticity, however some specialization of circuits mediating different types of information within the prefrontal cortex and between its connections is still evident.

A number of unanswered questions remain. The role of rhythmicity in working memory is still unresolved. Rhythmic discharges and local field potentials can be readily observed in working memory paradigms. The extent to which they represent information beyond what firing rate represents and whether they influence behavior is unclear. The circuits that maintain memory for objects have been more difficult to investigate. Whether fundamentally different mechanisms participate in object than spatial working memory maintenance is an open question. The intrinsic organization of the prefrontal cortex also remains unanswered. Hints of local organization suggest that the representation of information is not entirely random

across the cortical surface. Future studies addressing these questions will help tie cognitive phenomena to the activity of cortical neurons.

FIGURES

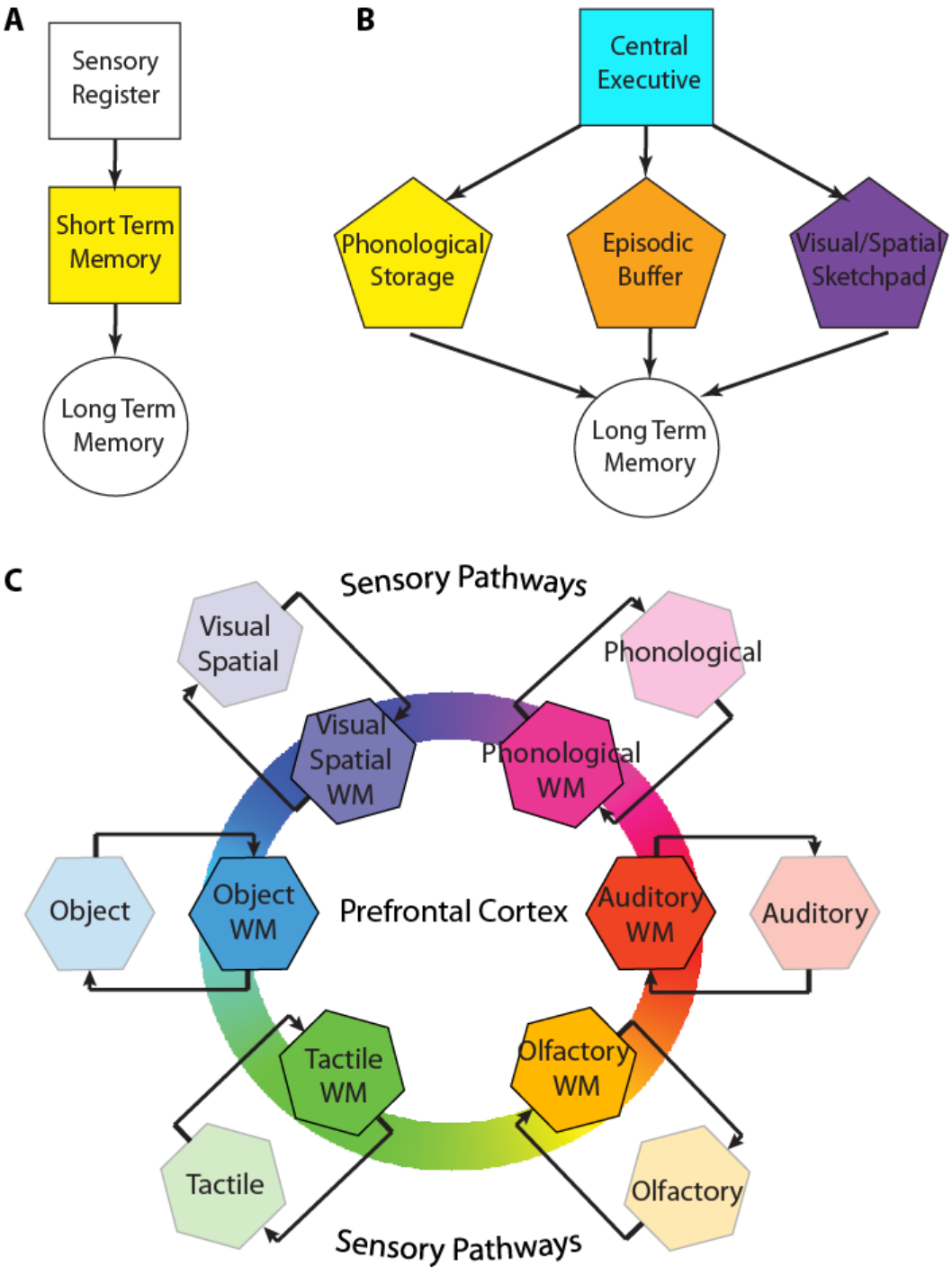


Figure 1. A. Schematic diagram of (A) the Atkinson and Shiffrin model, (B) Baddeley model, and (C) the model we are putting forth here (Jaffe and Constantinidis model). Colored squares in all schematic diagrams represent short/term working

memory stages. The prefrontal cortex plays a central role in the maintenance of working memory through activity that reverberates in short- and long-distance loops including the later stages of the sensory pathways that originally transmit sensory information to working memory.

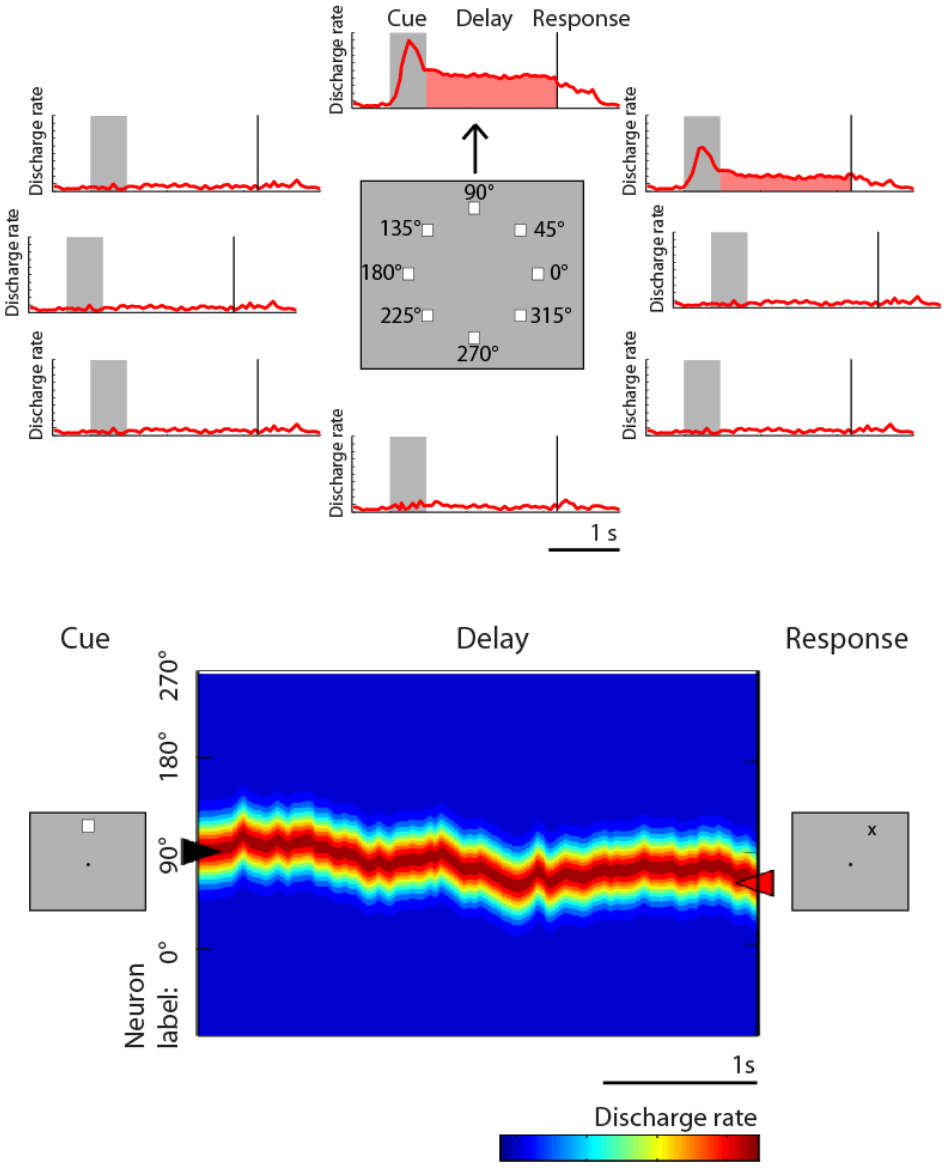


Figure 2. From single neuron responses to the bump attractor. Top, schematic illustration of responses of a single neuron to the ODR task for spatial working memory of a stimulus that appears at 8 different locations. Bottom, the population of neurons represent the stimulus location by the bump of activity in the network. Drifts of this activity result in errors. Adapted from Klingberg and Constantinidis, 2016.

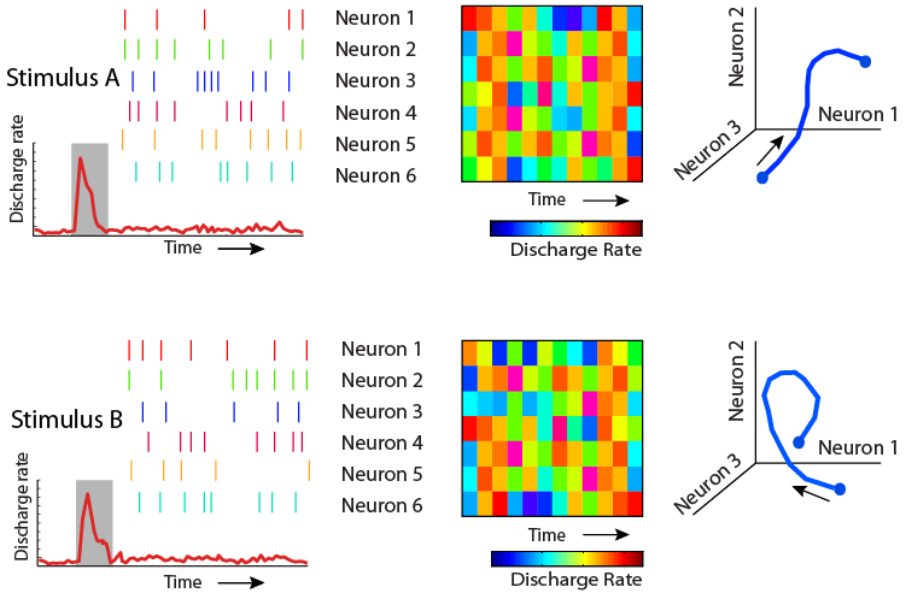


Figure 3. Schematic illustration of types of dynamic coding. Two different stimuli may not elicit an overall increase in firing rate during the delay period of working memory tasks (left panels). Their identity may still be decoded based on the pattern of spikes of individual neurons, which may be separable for different stimuli (middle panels). Furthermore, the pattern of activity of individual neurons may evolve dynamically in time, i.e. different neurons reach maximum activity at different time points during the trial, resulting in different trajectories in state space (right panels).

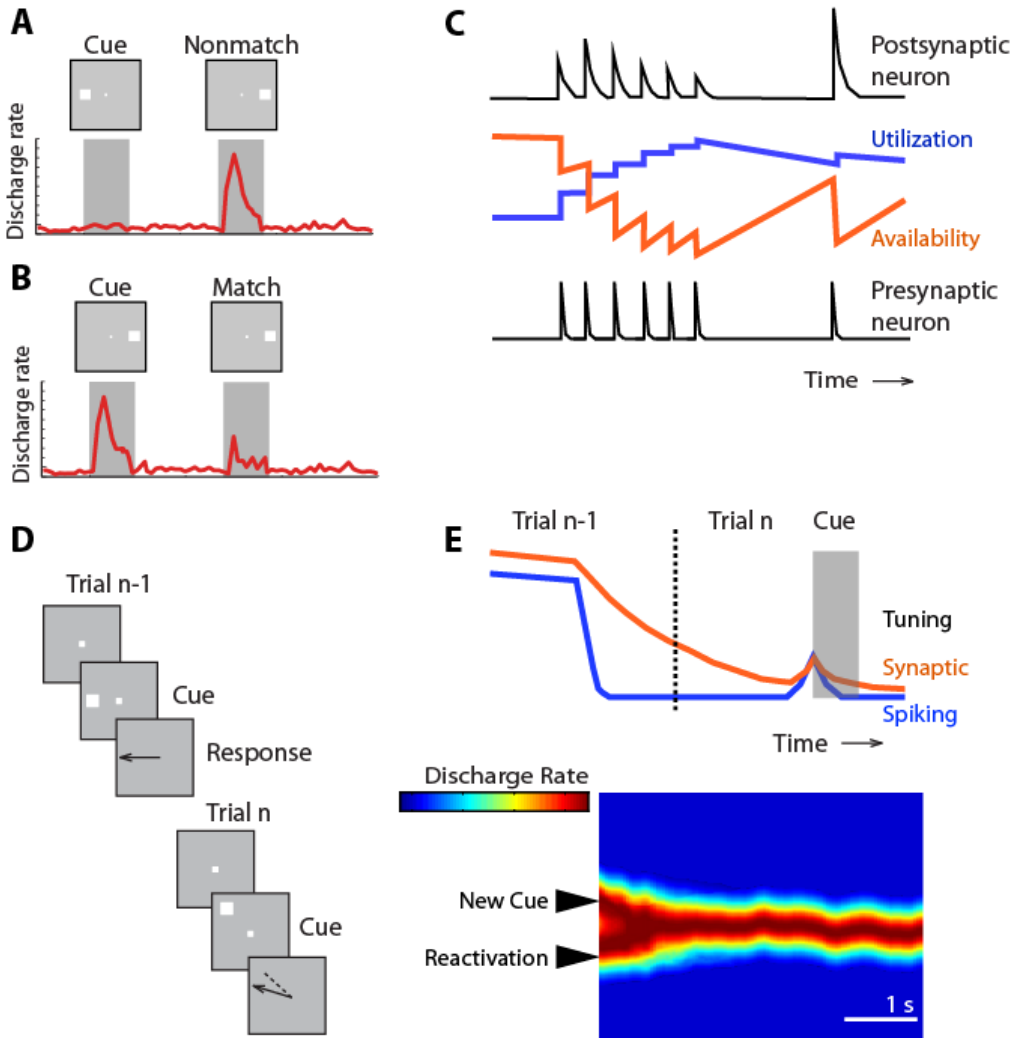


Figure 4. A-B. Schematic demonstration of the phenomenon of match suppression: a modulation of a stimulus firing rate depending on a previous stimulus, which does not depend on persistent spiking over the time interval between the two stimuli. **C.** The synaptic model of working memory. Utilization and availability of synaptic resources (e.g. Calcium concentration) modulate the magnitude of the postsynaptic potential generated by spikes elicited when a stimulus is presented for the first time, and when the same stimulus is presented for the second time. **D.** Schematic illustration of serial bias in working memory. The location of the previous

stimulus influences the memory of the current stimulus location. **E.** During the inter-trial interval (dotted line), firing rate no longer represents the previous stimulus and the neuron exhibits no selectivity for the location of the preceding stimulus. However, this information continues to be maintained by synaptic mechanisms and when the next trial begins, selectivity for the preceding stimulus re-emerges. The interaction of this reactivated bump and the bump of activity caused by the cue appearance causes a slight deviation (bias) which is evident in the pattern of behavioral performance. Panel C, adapted from Mongillo et al., 2008; panel E from Barbosa et al., 2020.

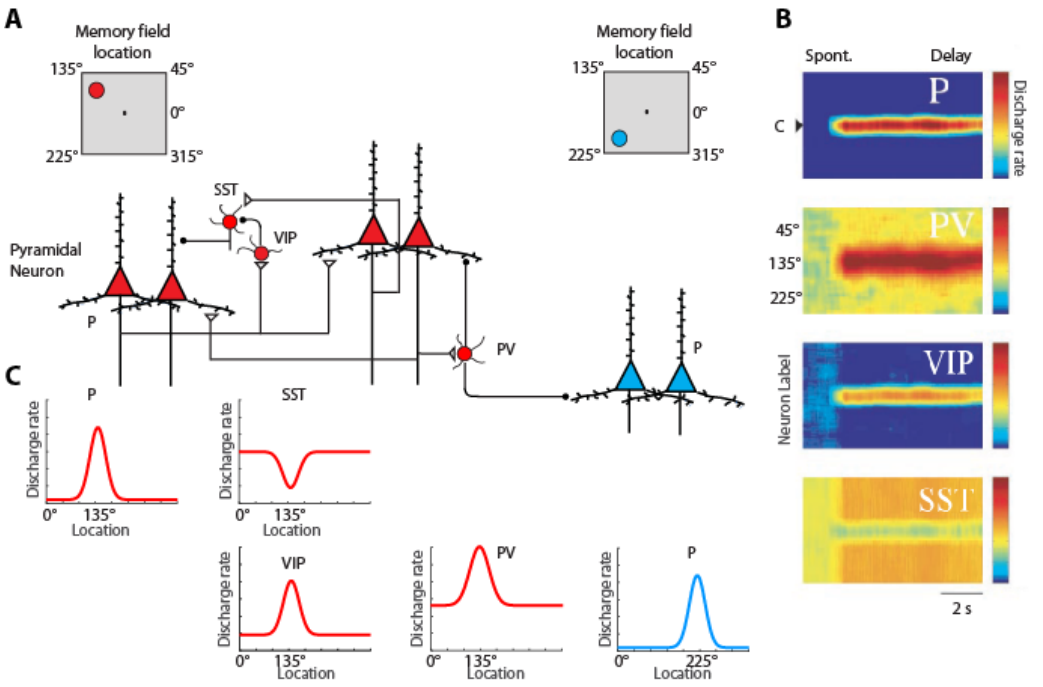


Figure 5. Division of labor model. Selectivity of Pyramidal Neurons (P) and three types of interneurons: Parvalbumin (PV), Calretinin (CR/VIP) and Calbindin (CB/SST) are shown. **A.** Insets on top are meant to illustrate that red-colored neurons on the left side of the figure are driven by a stimulus at the upper left of the screen, the 135° location, whereas blue-colored neurons on the right side of the figure are maximally activated by a stimulus in the lower left, 225° location. Excitatory synapses connect pyramidal neurons with similar preferences in the delay period that follows a stimulus in the upper left. **B.** Heat maps representing the activity of different neurons are plotted by preference for stimulus location (y-axis), as a function of time (x-axis). Adapted from Li et al., 2020.

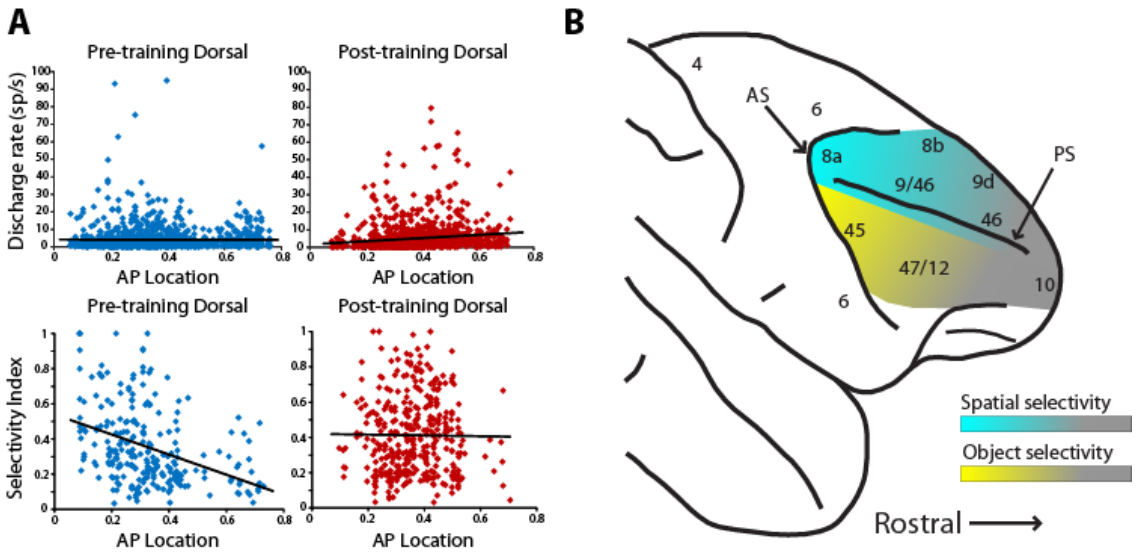


Figure 6. Stimulus selectivity along the dorso-ventral and anterior-posterior axes of the PFC before and after training. **A.** Firing rate and selectivity of dorsal prefrontal neurons at different positions along the anterior-posterior axis to spatial stimuli, before and after training. **B.** Schematic illustration of spatial and object selectivity across the PFC. Panel A from Riley et al., 2018; panel B from Constantinidis and Qi al. 2018.

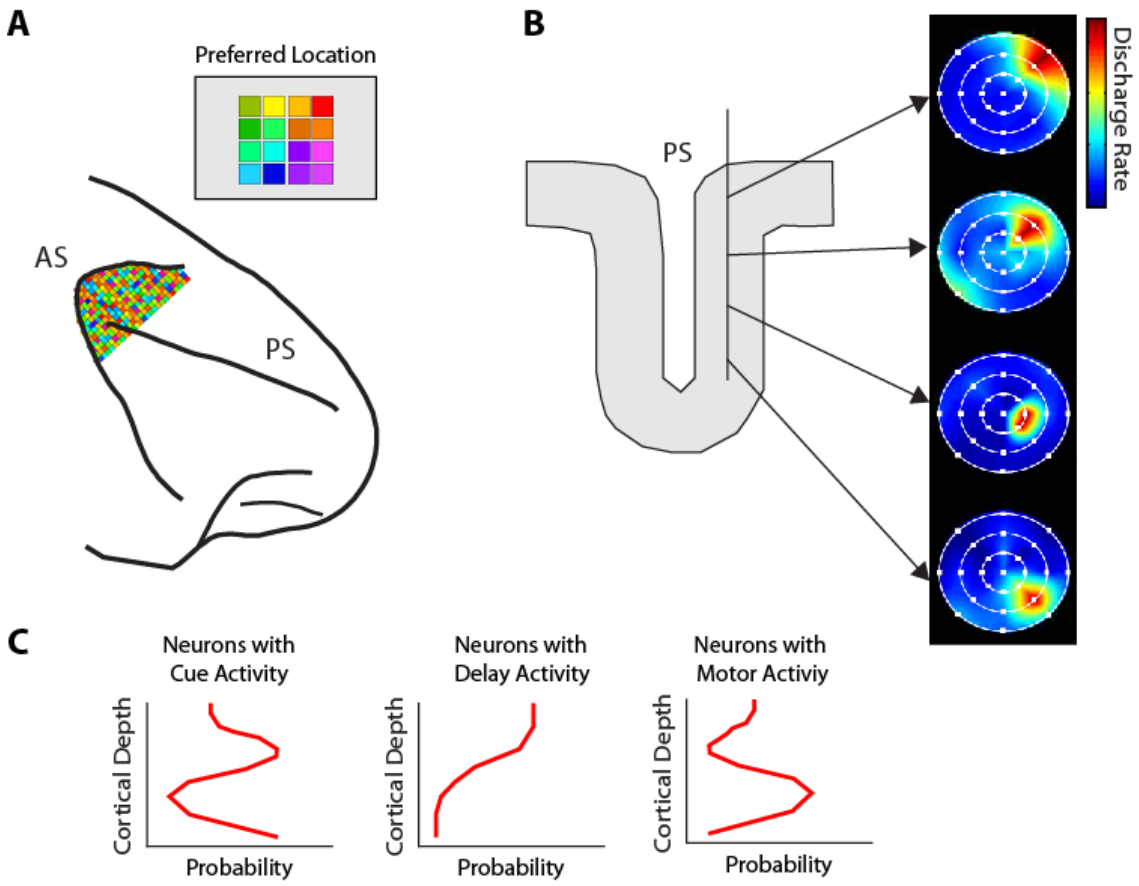


Figure 7. Local organization of the prefrontal cortex. **A.** No retinotopic map of visual space is present in dorsal posterior PFC, whose neurons do display strong selectivity for space. **B.** Some local organization and orderly progression of receptive fields is revealed in tangential electrode penetrations in the principal sulcus. **C.** Organization of activity across the depth of the PFC. Cue responses predominate in middle layers; delay period activity in upper layers; response activity in deeper layers of the PFC. Panel A based on Leavitt et al., 2018; panel B based on Arnsten et al., 2013; panel C based on Markowitz et al., 2015.

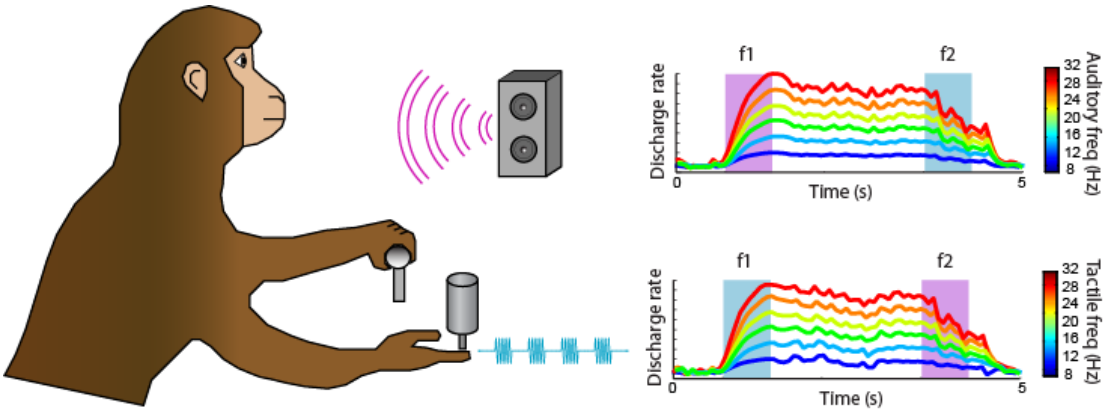


Figure 8. Schematic illustration of a cross-modal WM task. The monkey needs to remember the frequency of either a tactile or auditory stimulus and compare it with a stimulus of either modality. Right panels represent schematically the responses of a single neuron in the pre-Supplementary Motor Area for an auditory stimulus (purple bar) followed by a tactile stimulus (blue bar), or vice versa (bottom). Responses of the neuron during the delay period were modulated monotonically depending on the frequency of the stimulus, adhering to the same monotonic code of firing rate as a function of frequency for both the remembered auditory and tactile stimuli. From Constantinidis, 2016, referencing the data of Vergara et al., 2016.

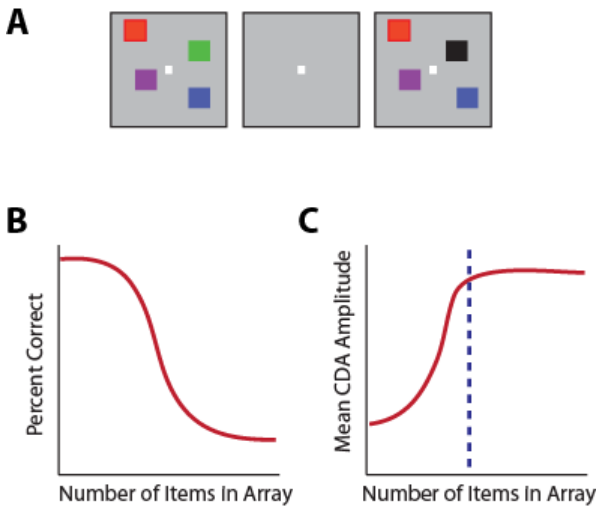


Figure 9. Properties of working memory capacity in human subjects. **A**. Change detection paradigm. Subjects need to maintain in memory a set of colored squares and report if a second display was identical to the first or not. **B**. Performance on the task declines as a function of set size. **C**. Mean Contralateral delay activity measured from posterior EEG electrodes as a function of elements in the array. Peak CDA amplitude predicts capacity (vertical) line. Based on Fukuda et al., 2010.

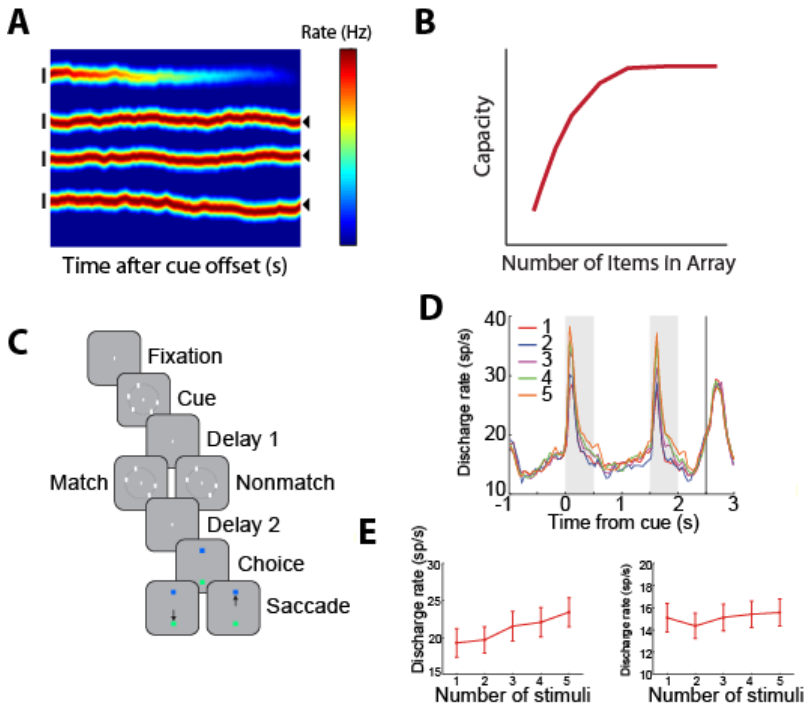


Figure 10. **A.** Predicted model of working memory capacity in the context of the bump attractor. Multiple stimuli are encoded at the beginning of the trial. An item that decays, fails to be recalled at the end of the trial. **B.** Capacity curves of monkeys. **C.** WM capacity task. **D.** Population firing rate for displays with different numbers of stimuli. **E.** Average firing rate during the cue and delay period. Adapted from Tang et al., 2019.

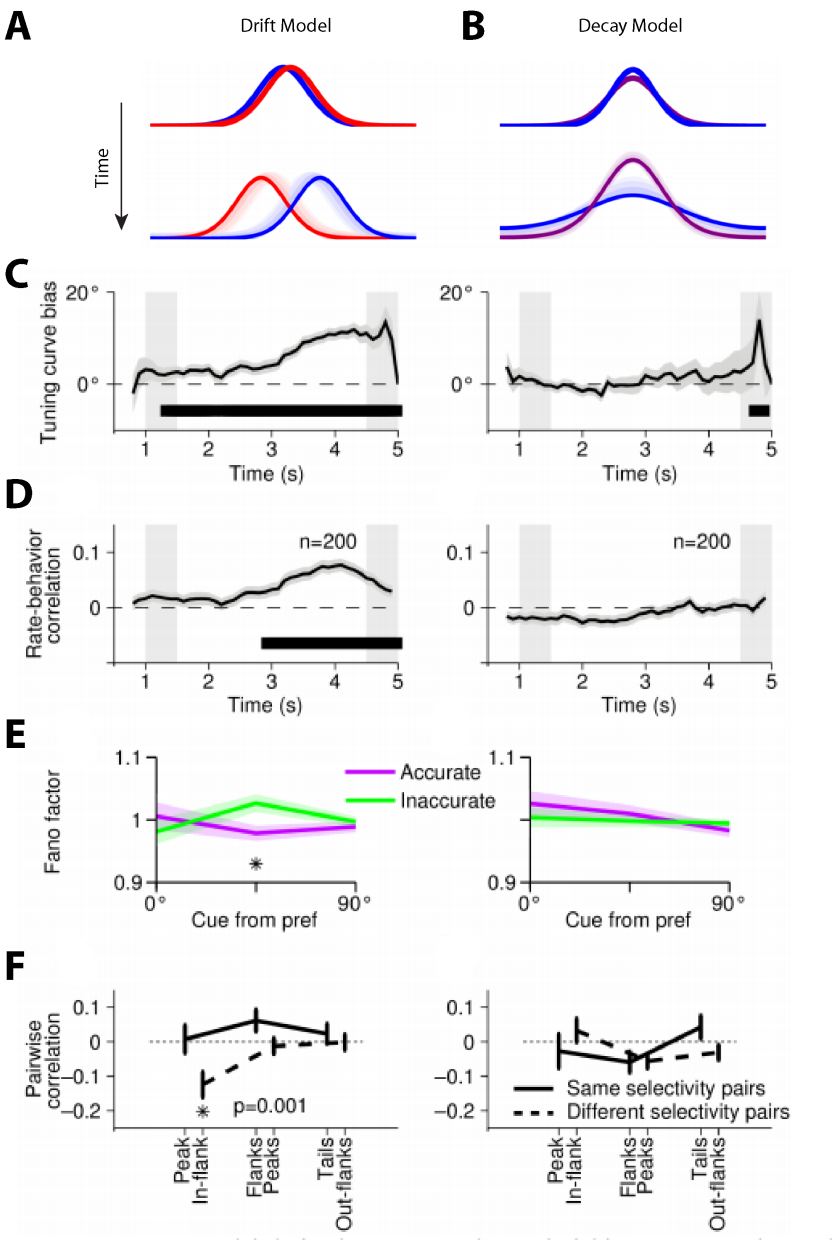


Figure 11. Predictions of **A,** the bump attractor model suggesting that drift is the main source of imprecision in working memory vs. **B,** a model where the bump of activity decays. **C.** For the bump attractor, but not the decaying bump model, the tuning curve bias computed from trials with clockwise versus counterclockwise deviations becomes increasingly positive through the delay. **D.** Only for the bump

attractor model, spike counts in response to flank stimuli correlated with behavioral deviations in the direction of the neuron's preferred cue as delay progressed. **E.** For the bump attractor model, but not the decaying bump model, Fano Factors computed separately for trials when a flank stimulus was presented are larger when considering inaccurate trials with large behavioral deviations (solid line) as compared to accurate trials (dashed line). **F.** Only for the bump attractor model, neuron pair noise correlation is negative for those pairs with dissimilar tuning, when responding to a middle flank stimulus. From Wimmer et al., 2014.

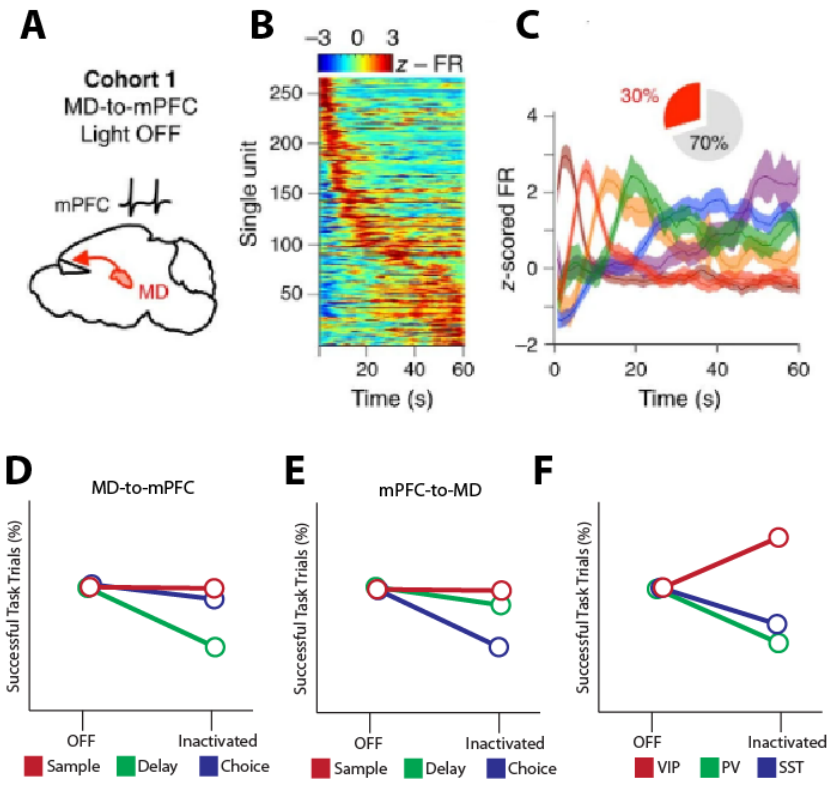


Figure 12. **A.** Patterns of activation of mouse working memory, recorded from mPFC. **B.** Individual neurons are typically active over a short period. **C.** Mean z-scored firing rate of delay-elevated units identified after clustering into six groups based on temporal correlation in firing rates from dark red (earliest) to purple (latest). Approximately 30% of all mPFC neurons in the dataset exhibited significant delay-elevated activity (inset). **D.** Behavioral effects of inactivation of MD-to-mPFC terminals. **E.** Behavioral effects of inactivation of mPFC-to-MD terminals. **F.** Behavioral effects of inactivation of different mPFC interneuron populations during an auditory working memory task. Panels A-E from Bolkan et al., 2017. Panel F is based on Kamigaki and Dan, 2017.

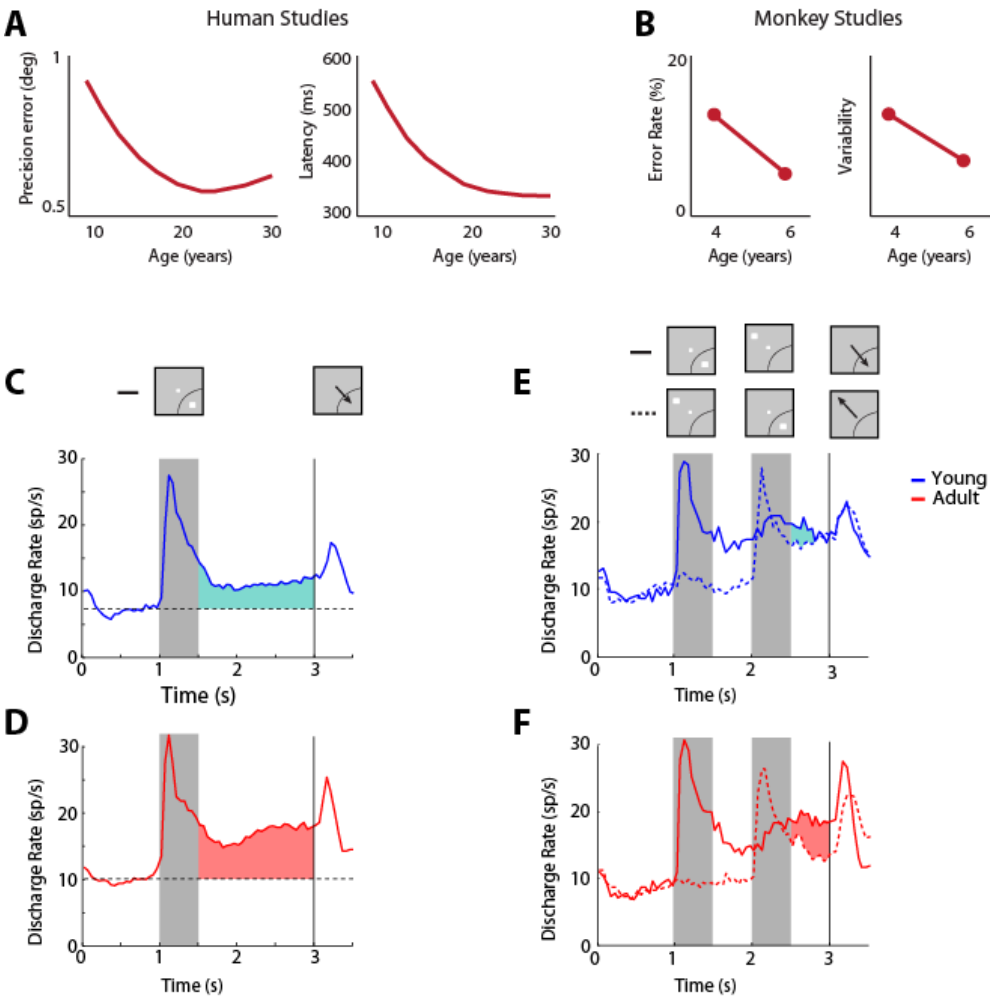


Figure 13. **A.** Behavioral improvement of human working memory in adolescence. **B.** Behavioral improvement of monkey WM in adolescence. **C.** Persistent discharges in the adolescent and **D,** adult PFC in the Oculomotor Delayed Response task. **E-F.** Activity in the adolescent and adult PFC in the distractor task. Panel A based on Simmons et al., 2017; panels B-F based on Zhou et al., 2016.

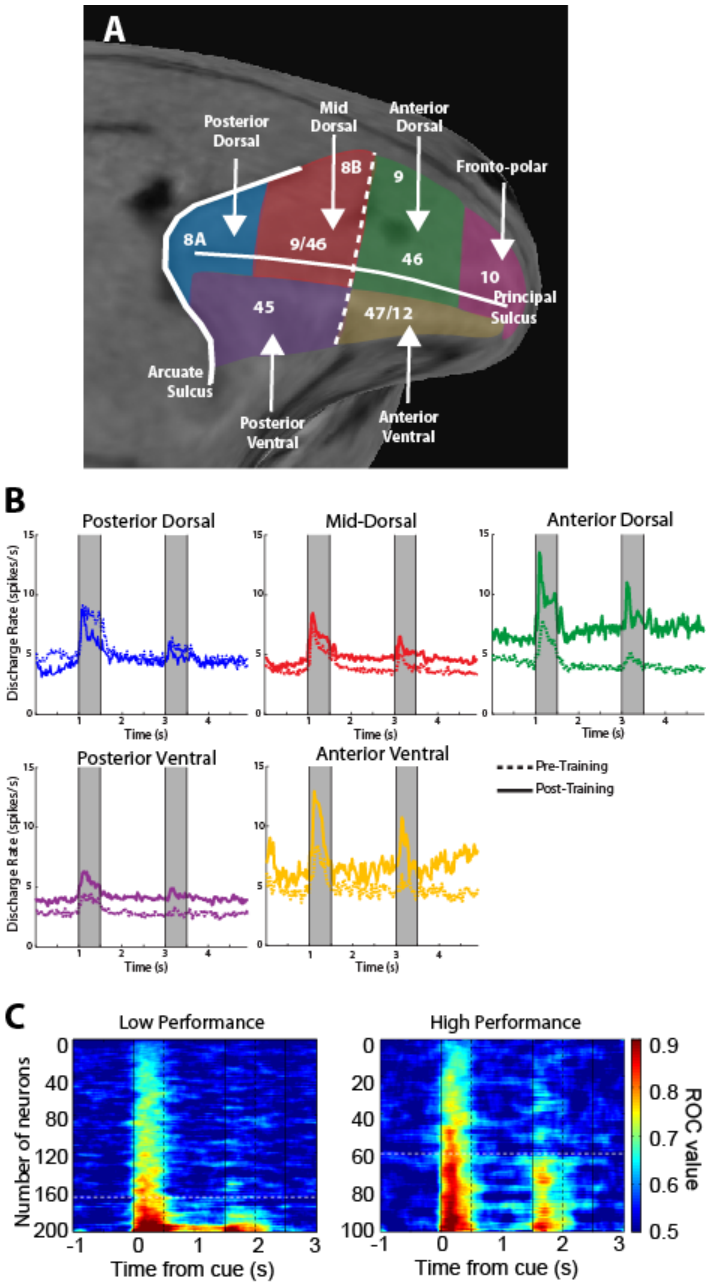


Figure 14. **A.** Anatomical MRI of the monkey lateral prefrontal cortex with anterior/posterior and dorsal/ventral subdivisions indicated relative to the Principle and Arcuate Sulci. **B.** Mean firing rate of neurons recorded in these subdivisions in monkeys both before and after they were trained to perform spatial WM tasks. Gray bars represent stimulus presentations. Data are shown separately for each

prefrontal region. **C.** ROC analysis for recordings from low-performance and high-performance sessions, as monkeys were trained in a multi-stimulus WM task. Panels A and B, based on Riley et al., 2018. Panel C from Tang et al., 2019.

Chapter 2: Emergence of Nonlinear Mixed Selectivity in Prefrontal Cortex after Training

The following chapter was adapted from a published article: Dang, W.*, R. J. Jaffe*, X. L. Qi and C. Constantinidis (2021). "Emergence of Nonlinear Mixed Selectivity in Prefrontal Cortex after Training." J Neurosci 41(35): 7420–7434. Any difference in stylistic variations is due to the requirement of the journal.

* W.D. and R.J.J. contributed equally to this work.

Abstract

Neurons in the PFC are typically activated by different cognitive tasks, and also by different stimuli and abstract variables within these tasks. A single neuron's selectivity for a given stimulus dimension often changes depending on its context, a phenomenon known as nonlinear mixed selectivity (NMS). It has previously been hypothesized that NMS emerges as a result of training to perform tasks in different contexts. We tested this hypothesis directly by examining the neuronal responses of different PFC areas before and after male monkeys were trained to perform different working memory tasks involving visual stimulus locations and/or shapes. We found that training induces a modest increase in the proportion of PFC neurons with NMS exclusively for spatial working memory, but not for shape working memory tasks, with area 9/46 undergoing the most significant increase in NMS cell proportion. We also found that increased working memory task complexity, in the form of simultaneously storing location and shape combinations, does not increase the degree of NMS for stimulus shape with other task variables. Lastly, in contrast to the previous studies, we did not find evidence that NMS is predictive of task performance. Our results thus provide critical insights on the representation of stimuli and task information in neuronal populations, in working memory.

SIGNIFICANCE STATEMENT

How multiple types of information are represented in working memory remains a complex computational problem. It has been hypothesized that nonlinear mixed selectivity allows neurons to efficiently encode multiple stimuli in different contexts, after subjects have been trained in complex tasks. Our analysis of prefrontal recordings obtained before and after training monkeys to perform working memory tasks only partially agreed with this prediction, in that nonlinear mixed selectivity emerged for spatial but not shape information, and mostly in mid-dorsal PFC. Nonlinear mixed selectivity also displayed little modulation across either task complexity or correct performance. These results point to other mechanisms, in addition to nonlinear mixed selectivity, representing complex information about stimulus and task context in neuronal activity.

INTRODUCTION

Working memory (WM) is broadly defined as the ability to encode, maintain, and manipulate information in the conscious mind over a period of seconds without the presence of any sensory inputs. As a core component of complex cognitive abilities such as planning and reasoning, the true importance of WM ultimately depends on whether it can maintain and manipulate task relevant information in a task relevant manner (Baddeley 2012). Multiple variables, including external sensory inputs and internal task requirements, must be encoded in order to achieve the level of adaptability in WM that is necessary for complex tasks. The mechanisms that underlie this encoding process across time and neuronal population is one of the most important questions in current WM research.

When individuals are required to maintain objects in their WM, neurons from a network of brain regions may exhibit selective and sustained increases or decreases in their activity in order to represent the remembered objects through these unique patterns of activity (Constantinidis and Procyk 2004). The prefrontal cortex (PFC) plays a leading role in this network, and by extension, in the use of WM (Riley and Constantinidis 2016). For example, when the PFC is damaged or degraded, whether through trauma, illness, or experimental lesions, performance in WM tasks seems to decrease dramatically (Morris and Baddeley 1988, Curtis and D'Esposito 2004, Rossi, Bichot et al. 2007).

Individual PFC neurons typically encode more than one variables, and the exact variables encoded are task dependent (Asaad, Rainer et al. 2000, Mansouri, Matsumoto et al. 2006, Machens, Romo et al. 2010, Warden and Miller 2010, Qi, Elworthy et al. 2015). More interestingly, a portion of neurons exhibit nonlinear mixed selectivity (NMS) for different variables, which means that their response to the combination of variables cannot be predicted by the linear summation of their responses to single variables (Rigotti, Barak et al. 2013, Parthasarathy, Herikstad et al. 2017, Johnston, Palmer et al. 2020). Theoretical studies have shown that NMS is useful for linear readouts of flexible, arbitrary combinations of variables (Buonomano and Maass 2009, Rigotti, Ben Dayan Rubin et al. 2010, Fusi, Miller et al. 2016), and may also control the trade-off between discrimination and generalization (Barak, Rigotti et al. 2013, Johnston, Palmer et al. 2020).

Despite the proposed importance of NMS on theoretical grounds, some experimental studies have failed to detect neurons with NMS (Cavanagh, Towers et al. 2018). It is therefore possible that NMS may manifest exclusively in a limited set of PFC subdivisions or alternatively, that NMS emerges exclusively after training to perform specific types of cognitive tasks. Moreover, the implications of NMS on other aspects of neural encoding, such as code stability, have not yet been investigated. We were therefore motivated to analyze and compare neural data from rhesus macaque monkeys before and after training. Here we report results of NMS as a function of task training, performance of different types of working memory tasks, and correct and error trials, across different prefrontal areas.

METHODS

Animals: Data obtained from six male rhesus monkeys (*Macaca mulatta*), ages 5–9 years old, weighing 5–12 kg, as previously documented (Riley, Qi et al. 2018), were analyzed in this study. None of these animals had any prior experimentation experience at the onset of our study. Monkeys were either single-housed or pair-housed in communal rooms with sensory interactions with other monkeys. All experimental procedures followed guidelines set by the U.S. Public Health Service Policy on Humane Care and Use of Laboratory Animals and the National Research Council's Guide for the Care and Use of Laboratory Animals and were reviewed and approved by the Wake Forest University Institutional Animal Care and Use Committee.

Experimental setup: Monkeys sat with their heads fixed in a primate chair while viewing a monitor positioned 68 cm away from their eyes with dim ambient illumination and were required to fixate on a 0.2° white square appearing in the center of the screen. During each trial and in order to receive a liquid reward (typically fruit juice), the animals maintained fixation on the square while visual stimuli were presented either at a peripheral location or over the fovea. Any break of fixation immediately terminated the trial and no reward was given. Eye position was monitored throughout the trial using a non-invasive, infrared eye position scanning system (model RK-716; ISCAN, Burlington, MA). The system achieved a < 0.3°

resolution around the center of vision. Eye position was sampled at 240 Hz, digitized and recorded. The visual stimulus display, monitoring of eye position, and synchronization of stimuli with neurophysiological data were performed with in-house software implemented on the MATLAB environment (Mathworks, Natick, MA), utilizing the Psychophysics Toolbox (Meyer and Constantinidis 2005).

Pre-training task: Following a brief period of fixation training and acclimation to the stimuli, monkeys were required to fixate on a center position while stimuli were displayed on the screen. The stimuli shown in the pre-training passive spatial task were white 2° squares, presented in one of nine possible locations arranged in a 3 × 3 grid with 10° distance between adjacent stimuli. The stimuli shown in the pre-training passive feature task were white 2° geometric shapes drawn from a set comprising a circle, diamond, the letter H, the hashtag symbol, the plus sign, a square, a triangle, and an inverted Y-letter. These stimuli could also be presented in one of nine possible locations arranged in a 3 × 3 grid with 10° distance between adjacent stimuli.

Presentation began with a fixation interval of 1 s where only the fixation point was displayed, followed by 500 ms of stimulus presentation (referred to hereafter as cue), followed by a 1.5 s “delay” interval (referred to hereafter as delay1) where, again, only the fixation point was displayed. A second stimulus (referred to hereafter as sample) was subsequently shown for 500 ms. In the spatial task, this second stimulus would be either identical in location to the initial stimulus, or

diametrically opposite the first stimulus. In the feature task, this second stimulus would appear in the same location to the initial stimulus and would either be an identical shape or the corresponding non-match shape (each shape was paired with one non-match shape). Only one nonmatch stimulus was paired with each cue, so that the number of match and nonmatch trials were balanced in each set. In both the spatial and feature task, this second stimulus display was followed by another "delay" period (referred to hereafter as delay2) of 1.5 s where only the fixation point was displayed. The location and identity of stimuli was of no behavioral relevance to the monkeys during the "pre-training" phase, as fixation was the only necessary action for obtaining reward.

Post-training task: Four of the six monkeys were trained to complete an active spatial, feature and conjunction WM tasks. These tasks involved presentation of identical stimuli as the spatial and feature tasks during the "pre-training" phase, but now monkeys were required to remember the spatial location and/or shape of the first presented stimulus, and report whether the second stimulus was identical to the first or not, via saccading to one of two target stimuli (green for matching stimuli, blue for non-matching). Each target stimulus could appear at one of two locations orthogonal to the cue/sample stimuli, pseudo-randomized in each trial.

The conjunction task combined the active spatial and feature tasks, using the same shape stimuli and presented at the same possible locations, with the same timing. In a single recording session, only four shape-location combinations

involving two shapes and two locations were used. The conjunction task was the most complex task in the current study, as the monkeys were required to simultaneously store two different types of information, location and shape, in their working memory.

Surgery and neurophysiology: A 20 mm diameter craniotomy was performed over the PFC and a recording cylinder was implanted over the site. The location of the cylinder was visualized through anatomical magnetic resonance imaging (MRI) and stereotaxic coordinates post-surgery. For two of the four monkeys that were trained to complete active spatial, feature and conjunction WM tasks, the recording cylinder was moved after an initial round of recordings in the post-training phase to sample an additional surface of the PFC.

Anatomical localization. Each monkey underwent an MRI scan prior to neurophysiological recordings. Electrode penetrations were mapped onto the cortical surface. We identified 6 lateral PFC regions: a posterior-dorsal region that included area 8A, a mid-dorsal region that included area 8B and area 9/ 46, an anterior-dorsal region that included area 9 and area 46, a posterior-ventral region that included area 45, an anterior-ventral region that included area 47/12, and a frontopolar region that included area 10. However, the frontopolar region was not sampled sufficiently to be included in the present analyses.

Neuronal recordings: Neural recordings were carried out in the aforementioned areas of the PFC both before and after training in each WM task. Subsets of the data presented here were previously used to determine the collective properties of neurons in the dorsal and ventral PFC, as well as the properties of neurons before and after training in the posterior-dorsal, mid-dorsal, anterior-dorsal, posterior-ventral, and anterior-ventral PFC subdivisions. Extracellular recordings were performed with multiple microelectrodes that were either glass- or epoxylite-coated tungsten, with a 250 μm diameter and 1–4 MΩ impedance at 1 kHz (Alpha-Omega Engineering, Nazareth, Israel). A Microdrive system (EPS drive, Alpha- Omega Engineering) advanced arrays of up to 8-microelectrodes, spaced 0.2–1.5 mm apart, through the dura and into the PFC. The signal from each electrode was amplified and band-pass filtered between 500 Hz and 8 kHz while being recorded with a modular data acquisition system (APM system, FHC, Bowdoin, ME). Waveforms that exceeded a user-defined threshold were sampled at 25 μs resolution, digitized, and stored for off-line analysis. Neurons were sampled in an unbiased fashion, collecting data from all units isolated from our electrodes, with no regard to the response properties of the isolated neurons. A semi-automated cluster analysis relied on the KlustaKwik algorithm, which applied principal component analysis of the waveforms to sort recorded spike waveforms into separate units. To ensure a stable firing rate in the analyzed recordings, we identified recordings in which a significant effect of trial sequence was evident at the baseline firing rate (ANOVA, $p < 0.05$), e.g., due to a neuron disappearing or

appearing during a run, as we were collecting data from multiple electrodes. Data from these sessions were truncated so that analysis was only performed on a range of trials with stable firing rate. Less than 10% of neurons were corrected in this way. Identical data collection procedures, recording equipment, and spike sorting algorithms were used before and after training in order to prevent any analytical confounds.

Data analysis – Neural Selection: Data analysis was implemented with the MATLAB computational environment (Mathworks, Natick, MA), with additional statistic tests implemented through Originlab (OriginLab Corporation, Northampton, MA) and StatsDirect (StatsDirect Ltd. England). Peristimulus time histograms (PSTHs) for illustrations were calculated through the moving window average method with a Gaussian window that had a 200 ms standard deviation, with the shaded area indicating two times standard error cross trials. For all tasks, only cells with at least 12 correct trials for each cue-sample location/shape pairs were included in the analysis. To classify neurons of the spatial task into different categories of selectivity, we performed two-way ANOVAs determining the influence of task factors on the neuron's spatial tuning. We set up this analysis in two different ways. In the initial 2-way ANOVA, we analyzed firing only for the sample (second) stimulus in the task. The two factors were the location of the sample stimulus and its match or nonmatch status, thus quantifying whether tuning for the sample location depends on whether it is match or nonmatch. In a subsequent 2-way ANOVA, we

analyzed firing rate of both the cue (first) and sample (second) stimulus presentations, from match trials only. The two factors were stimulus location and the task epoch (first or second stimulus presentation), thus quantifying whether tuning is dependent on the order of stimulus presentation. Neurons with classic selectivity (CS) exhibited a main effect of only one factor without significant interaction term. Neurons with linear mixed selectivity (LMS) exhibited main effects of both factors without a significant interaction term. Neurons with NMS exhibited a significant interactions term. Finally, non-selective (NS) neurons exhibited no significant main effects nor interaction term. Similarly, the two factors for feature task ANOVA analysis were stimuli shape x matching status, and stimuli shape x task epoch for the trial. The $P<0.05$ level was chosen as the threshold for statistical significance for all ANOVA analysis. The animals were unable to predict the matching status of the trials during the fixation, cue and delay1 periods, and as a result, the proportion of cells with significance for matching factor and the interaction term would not be expected to exceed 5% during these first three periods. Time bins used to calculate spike rate for different task stages are displayed in Figure 1. Choice period was defined as the 1 second interval after the choice array appears onscreen.

Dimensionality Reduction. Our method for measuring dimensionality can be explained as follows: Let M be a dataset of size $N \times T \times C$, where N represents number of neurons, T indicates the number of trials, and C indicates the number of different

conditions. In the case of our task, C would refer to all possible combinations of stimulus identity (location or feature) and match or nonmatch status. Each entry in this matrix represents the firing rate in a specified time bin, namely the stimulus presentation period, or the delay period that follows it. Population response can thus be plotted across conditions and trials in an n-dimensional space. The plotting space is very high dimensional (dimensionality equals the number of cells), but most variance can be accounted for by a smaller number of axes if neural data are linearly correlated across conditions. As a result, the neural dimensionality is measured by the number of neural response subspace axes that explain this consistent variance, within the original n-dimensional space. In a fictitious noiseless case, where the response to each condition is identical in each trial, the dimensionality of the matrix could be reduced to that of the trial-averaged version of size N x C. The dimensionality of the matrix is equal to the number of nonzero singular values, which can be obtained by singular value decomposition (SVD), defined as:

$$M = U\Sigma V^T$$

In our context, M is the real number neural response data matrix. U is an N x N matrix, the columns of which provide a basis set of eigenvectors capturing the correlation among neurons, the weighted linear combination of which accounts for the variance in the original data. Σ is an N x C diagonal matrix with non-negative values equal to the square root of the eigenvalues, assigning weights to the columns of U. The dimensionality of the matrix is equal to the number of nonzero

singular values. Finally V is a C x C matrix mapping neuronal responses to stimulus conditions. However, the real data contain random noise that pushes data points away from a low dimensional manifold and thus dimensionality results are almost always overestimated. For this reason, we used the cross-validation method developed by Ahlhaim and Love to obtain a more accurate estimate of dimensionality for the real-word data (Ahlheim and Love, 2018). The basic idea is that the SVD operation could also be written in a linear equation form as follows:

$$\mathrm{M} = u_1\sigma_1 v_1^T + u_2\sigma_2 v_2^T + \ldots + u_c\sigma_c v_c^T$$

in which $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_c \geq 0$ are singular values, u and v are the corresponding vectors from U and V respectively. One can create a low dimensional approximation of M by gradually adding more terms that explain more variance. In the dimensionality estimation process, data from a total of j trials from each condition were first partitioned into three sets: a training set containing j-2 trials, a validation set containing 1 trial and a testing set containing 1 trial. SVD was then applied to the averaged training dataset and used to reconstruct the training data from the U, Σ, and V-transpose matrices using increasing numbers of dimensions from 1 to the dimensionality of the training data. For each reconstructed matrix of different dimensionality based on the SVD, the Pearson correlation coefficient was computed between this result and the validation trial. This process was repeated over all possible partitions of the data into training and validation/testing sets. The dimensionality k that produced the highest mean correlation between the SVD reconstructed training data and the validation trial determined the dimensionality of

the neural representation. The correlation of this representation with the held-out test data, provided the final reconstruction correlation, as an estimate of reconstruction precision. A similar method has also been recently used to estimate the dimensionality over time for neural data (Cueva, Saez et al. 2020). The dimensionality of the sample and delay2 period in the spatial and feature task was calculated on 50 resamples of a 200-cell pseudo-population in the corresponding datasets.

Comparisons between areas and tasks. Only PFC areas with more than 50 cells in both pre- and post-training time points were included in the subdivision mixed selectivity comparison analysis. Thus, for the feature task, only data from the mid-dorsal, posterior-dorsal and posterior-ventral PFC were analyzed. For the spatial task, only data from the mid-dorsal, posterior-dorsal, posterior-ventral, anterior-dorsal and anterior-ventral PFC were analyzed.

Neural data from tasks that applied the exact same visual stimuli were used to compare mixed selectivity between feature/spatial and conjunction tasks. For example, to compare the feature and the conjunction tasks, we started by selecting a subset of conjunction trials, in which both visual stimuli appeared at the same location as the corresponding feature task trials. Then, since all eight shapes were used in a single recording session for the feature task, a subset of trials, in which same shape pairs were used as the corresponding conjunction task, could be

chosen as the feature dataset. Our prior methods of ANOVA analyses could thus be applied for comparison across these datasets.

For comparing mixed selectivity in success and error trials from the spatial task, we first examined F scores from the ANOVA of two task variables—the stimuli location and matching status. In this analysis, we utilized neurons that had at least 3 match and non-match trials for both the correct and error dataset, in at least 3 stimuli locations. The number of minimum trials and stimuli locations were chosen to maximize the average trial number for each cell included into the analysis, while still retaining a sufficiently large sample (i.e. >150 cells). The same number of trials from each stimuli location were randomly chosen in the correct and error dataset. This randomized trial selection process was repeated 50 times in order to make the best use of the uneven number of available trials in two datasets. We also analyzed factors of stimuli location and task epoch. For this analysis, we used neural data from match trials only, and from neurons that had at least 4 correct and error trials, in at least 4 stimuli locations. To calculate dimensionality in the spatial task in correct vs erroneous conditions, we first identified 56 cells with at least 4 trials in the same four conditions in both the correct and error datasets. We then randomly selected 50 cells to construct a pseudo-population for measuring dimensionality using the previously described SVD based method. This process was repeated 50 times to obtain a confidence interval.

For decoding analysis of stimuli identity, matching status or saccade directions, spiking responses from 1 second before cue onset to 5 seconds after

cue onset were first binned using a 400 ms wide window and 100 ms steps to create a spike count vector with a length of 57 elements. A pseudo-population was then constructed using the spike count vectors from all the available neurons of all the available animals, thus resulting in a dataset with 96 trials, as if they were recorded simultaneously. The population response matrix was z-score normalized before being used to train the decoder. A support Vector Machine (SVM) decoding algorithm with a linear kernel was implemented using the MATLAB fitcecoc function to decode stimuli location, stimuli shape, the match/non-match status of trials, or the saccadic direction. A 10-fold cross validation method was used to estimate the decoder performance and 20 random samplings were implemented to calculate a 95% confidence interval. For the spatial and feature task, the decoding baseline for sensory information or saccadic direction was 12.5%, since there were 8 different options, and 50% for the matching status, since there were only 2 different options.

In the pre-training vs post-training decoding analysis (Fig. 8 A), linear (CS and LMS) and nonlinear (NMS) neurons are first defined by their pre and post training responses in the sample or delay2 period. Each classified population was then applied to decode sensory information (location and shape) and matching status. A randomization test was used to determine the time points at which decoding performance was significantly different between different selectivity categories. In short, we constructed the null distribution by randomly reassigning the cell selectivity labels under comparison, and re-computing the maximum absolute difference across all time points of the data in each iteration. This procedure was

repeated for 5000 times. A difference was deemed to be significant if the true response difference occurred at the extremes of this null distribution ($p < 0.05$, two-tailed). Since every point in the null distribution is the maximum of all time points, this method already corrected for the multiple comparisons.

For saccadic decoding in the spatial task (Fig. 11), we first constructed pseudo-populations by identifying neurons with correct and error trials in the same conditions (defined by saccadic direction and matching status). We then used correct trials to train a saccadic direction decoder. Only correct trials were utilized in the training dataset, as these would contain accurate representations of the saccadic direction information. Our decoder was subsequently employed to predict saccade direction in a test dataset of both correct and error trials.

A Random "coloring" process was utilized as previously described (Rigotti et al. 2010) for the purpose of calculating the number of implementable binary classifications. Specifically, trials with 16 experimental (8 stimuli identities × 2 matching status) conditions were equally and randomly assigned one of two labels as the correct answer for binary classification training, making a total of 12870 combinations with an equal number of conditions on each side of targeted decision boundary. An SVM decoder was then trained for every binary classification. A threshold of 80% correct for cross validated decoding performance was used to define an "implementable" binary classification.

To more directly compare our results with previous research, a decoder was constructed to decode the matching or nonmatching status of a trial after pure

selectivity for matching information was removed from each cell as reported previously (Rigotti et al. 2010). An SVM with a polynomial kernel was used to decode experimental conditions defined by cue location x matching status (16 conditions in total). We used the MATLAB fitcecoc function for this classification as well. In this case, the method produced a16-bit binary output based on the firing rates of the neurons, which was compared to the expected binary output of each experimental condition; data from each trial were then classified into the category that was most similar (needed the smallest error correction). Four conditions were compared using this non-linear decoder: 200 randomly selected informative cells (CS, LMS and NMS defined in sample period) in the post training condition, with unaltered selectivity profiles; 200 randomly selected informative cells in the post training condition ,with classical selectivity for matching status removed; 200 randomly selected informative cells in the pre training condition, with classical selectivity for matching status removed; 200 randomly selected classical selective cells in the post training condition, with classical selectivity for matching status removed. The inclusion of the last condition is necessary to confirm the effectiveness of the pure selectivity removal process. We removed the classical selectivity for certain variable (e.g. matching status) of certain neuron by contaminating the firing rate recorded during one condition (sample match cue) with spikes recorded in the other condition (sample nonmatch to cue). More specifically, every time we sampled a spike count from a trial in condition $c \in T_1$, an additional noise source was artificially superimposed by adding a spike count sampled from a randomly

selected trial in the same time bin, but belonging to a different condition $c' \in T_2$, in such a way that the classical selectivity for task variable T was equalized. The classical selectivity removal procedure can thus be formally denoted as the following equation:

$$v_i^c(t) = r_i^c(t) + \sum_c w\, r_i^{c'}(t)$$

Where $r_i^c(t)$ indicates spike count for neuron i at time t from a trial belonging to condition c, where $c \in T_1$, and $v_i^c(t)$ refer to the response of the same neuron at the same time point, after removing the classical selectivity for task variable T, $\{T_1, T_2, \ldots\} \in T$. Finally, $wr_i^{c'}(t)$ represents a random variable that selects a condition $c' \notin T_1$ at random.

Only informative neurons (CS, LMS, and NMS neurons) in the delay2 period were used for the cross temporal decoding analysis (Fig. 9), since we wanted to explore the decoding dynamics during the delay period. The linear SVM decoder was trained on individual time points and thus had 57 linear decision boundaries. The same dataset was then classified by every decision boundary in the vector to produce a 57x57 matrix—a process that was repeated 20 times in order to eliminate the noise. The decoding performance matrix for each condition was normalized individually to highlight the coding dynamics rather than absolute performance. A permutation-based test was used to determine the difference in the stability of different cell population recordings. For every permutation cycle, we randomly shuffled the labels of selectivity category for all cells, then we train separate decoders for different populations base on the new shuffled labels, two sample t

tests were ran on all corresponding pairs between decoding performance matrix of linear and nonlinear populations, and all p-values was recorded. We ran 1000 such permutations to construct a null distribution of the chosen statistic value (p-value). P values obtained from the original dataset that exceeded 5% threshold were considered statistically significant.

Data availability: All relevant data and code will be available from the corresponding author on reasonable request. Matlab decoder code for figure 8 and 9 is available at https://github.com/dwhzlh87/mixed-selectivity

RESULTS

Extracellular neurophysiological recordings were collected from the lateral PFC of six monkeys before and after they were trained to perform the match/nonmatch WM tasks (Meyer, Qi et al. 2011, Riley, Qi et al. 2018). The task required them to view two stimuli appearing in sequence with an intervening delay period between them, and after a second delay period to report whether or not the second stimulus was identical to the first. The two stimuli could differ in terms of their location (spatial task, Fig. 1A), shape (feature task, Fig. 1B), or both (conjunction task, Fig. 1C). If the second stimulus matched with the first, monkeys would saccade towards a green target during a subsequent interval. Otherwise, they would saccade to a blue target at a diametrical location. Monkeys were able to perform all three tasks with accuracy much higher than chance level (mean performance: spatial task 86.2%, feature task 82.1%, conjunction task 81.0%).

A total of 1617 cells from six monkeys (table 1-4) were recorded while the animals viewed the spatial stimuli passively, and 1493 cells from five monkeys were recorded while the animals viewed the feature stimuli passively, prior to any training. A total of 1104 cells from three monkeys and 1091 cells from two monkeys were collected while the animals were performing the active spatial and feature tasks, respectively, which were mutually called “post-training”. We also collected neural data from 247 neurons for the passive spatial task from two monkeys after they were trained in the active spatial task. An additional 975 cells from two

monkeys were collected while they were performing the active, post-training conjunction task.

Types of selectivity in individual neuronal responses

In our tasks, the context of a given stimulus depends upon the task interval and sequence in which it is presented. We first considered how selectivity for stimulus location and shape in the spatial, feature and conjunction WM tasks may vary when the same sample stimulus appears as a match (i.e. it is preceded by a cue at the same location/shape) or a nonmatch (i.e. is preceded by a cue stimulus of a different location/shape). The neuronal firing rate in a dataset is therefore a function of the stimulus location or shape, and whether the sample stimulus matched the cue stimulus. We used a 2-way ANOVA with factors of stimulus location/shape and match/nonmatch status to classify neurons into four categories of selectivity. CS neurons exhibited a significant main effect to only one of the factors (stimulus identity or matching status) and had no significant interaction term. In Fig. 2, the first exemplar set of plots depicts such a CS cell, selective exclusively for the location of the stimuli, which does not respond differently regardless of whether the stimulus appeared as a match or nonmatch. The second exemplar of Fig. 2 displays another CS cell not selective for the location of the stimuli but demonstrating higher mean response when the stimulus appeared as a non-match. LMS neurons exhibited a significant main effect for both factors but had no significant interaction term. The third exemplar of Fig. 2 displays such an LMS neuron demonstrating a

higher mean response when stimuli appear as non-match, while simultaneously displaying the same rank order preference for location. NMS neurons exhibited a significant interaction effect, as shown in the last exemplar in Fig. 2, a neuron demonstrating different selectivity pattern for locations under match vs. non-match conditions. Finally, NS indicated the neurons with no selectivity to any factors or their interaction under consideration. These analyses were performed using the firing rate recorded during the stimulus presentation period, and again, using the firing rate recorded during the delay period that followed it.

A second type of NMS was identified in terms of selectivity for stimulus sequence, that is, whether the same stimulus appeared first (cue) or second (sample). To avoid the confound of the match or nonmatch status of the second stimulus, we relied exclusively on match stimuli. This form of NMS was also evaluated through a 2-way ANOVA model, identifying CS, LMS, NMS, and NS neurons in terms of how the neurons represented the exact same stimulus when it appeared as a cue and as a match stimulus.

Effects of training on NMS

When we used the factors of stimulus location/shape and match/nonmatch status for our two-way ANOVA, we found that training in the spatial WM task increased the proportion of NMS cells in both the sample period and the delay period that followed the sample (sample period: pre-training proportion=6.2%, post-training proportion=12.3%, two-sample proportion test, $z=5.31$, $p=1.13\times10^{-7}$; delay2 period:

pre-training proportion= 2.8%, post-training proportion=6.2%, two-sample proportion test, z=4.62, $p=4.86\times10^{-5}$). However, this increase in selectivity was not exclusive to NMS cells. The proportion of CS cells also increased in the delay period following the sample (pre-training proportion= 10.6%, post-training proportion=14.8%, two-sample proportion test, z=3.19, p=0.0014).

The increase in NMS cells was not evident for all types of training. When we looked at the proportion of change across the pre-training and post-training feature task, we only found an increase of proportion for CS cells (sample period: pre-training proportion= 12.0%, post-training proportion=15.7%, two-sample proportion test, z=2.65, p=0.0081; delay2 period: pre-training proportion= 9.0%, post-training proportion=22.6%, two-sample proportion test, z=9.37, $p=3.4\times10^{-20}$). No significant increase in the proportion of NMS cells was observed (sample period: pre-training proportion=5.8%, post-training proportion=6.7%, two-sample proportion test, z=1.01, p=0.314; delay2 period: pre-training proportion= 4.2%, post-training proportion=4.6%, two-sample proportion test, z=0.522, p=0.602) (Fig. 3 A,B).

Similar results were observed when we used the factors of stimulus location/shape and task epoch (cue vs. match) for the two-way ANOVA instead (Fig. 3 C,D). For the spatial task, training increased the proportion of NMS cells, at least in the delay period (stimulus period: pre-training proportion= 5.7%, post-training proportion=7.4%, two-sample proportion test, z=1.78, p=0.075; delay period: pre-training proportion= 2.1%, post-training proportion=6.2%, two-sample proportion test, z=5.03, $p=4.94\times10^{-7}$). A similar increase was observed for the CS cells

(stimulus period: pre-training proportion= 21.3%, post-training proportion=25.3%, two-sample proportion test, z=2.37, p=0.018; delay period: pre-training proportion= 24.2%, post-training proportion=31.1%, two-sample proportion test, z=3.93, $p=8.53\times10^{-5}$). In the feature task, only the proportion of CS cells changed (stimulus period: pre-training proportion= 21.9%, post-training proportion=32.6%, two-sample proportion test, z=6.05, $p=1.45\times10^{-9}$; delay period: pre-training proportion= 27.6%, post-training proportion=41%, two-sample proportion test, z=7.16, $p=8.32\times10^{-13}$). The proportion of NMS cells with an effect in the stimulus period remained relatively unchanged for the cue/match period (pre-training proportion= 3.4%, post-training proportion=4.7%, two-sample proportion test, z=1.59, p=0.112), as well as the delay period (pre-training proportion= 2.4%, post-training proportion=3.8%, two-sample proportion test, z=1.95, p=0.051). We did not have sufficient power to perform this analysis individually for all animals, but for two subjects with sufficient data in both the feature and spatial tasks, lack of substantial NMS in the feature task after training was evident in both. For the sample period in subject A, the NMS proportion was 3.6%, and 6.5% for the pre-training and post-training periods, respectively (two-sample proportion test, z=0.93, p=0.352) For subject A, it was 6.4%, and 6.9%, respectively (two-sample proportion test, z=0.37, p=0.709). In the delay2 period, the proportion of NMS neurons in monkey A was 1.8%, and 3.8% for the pre- and post-training periods, respectively (two-sample proportion test, z=0.93, p=0.407). For monkey E it was 5.5%, and 4.8%, respectively (two-sample proportion test, z=0.42, p=0.672).

To further validate our proportional measure for NMS and compare our results to previous research on NMS in the PFC, we plotted the F scores for the interaction term (i.e. stimulus identity × matching status) in both the spatial and the feature task (Fig. 4 A). We found that this measure of NMS for individual cells increased specifically for the spatial task, indicated by a slightly higher mean F score values after training for the spatial task. Importantly, this slight increase in mean is not trivial considering that the proportion of NMS is relatively low.

To quantify the change in NMS in an alternative way, we also measured the dimensionality of population responses in the sample and delay2 period for the spatial and feature task. Again, this analysis confirmed the results of our cell proportion measure (Fig. 4 B). For the spatial task, there was a significant increase of dimensionality after training for both types of nonlinear mixed selectivity. For NMS defined as the location x matching effect, i.e. different selectivity for the same stimulus when it appeared as a match or a nonmatch, dimensionality increased after training in the sample period (pre-training dimensionality= 5.72, post-training dimensionality=10.33, two-sample t test, $t(98)=12.21$, $p=2.18\times10^{-21}$) as well as in the delay2 period (pre-training = 3.25, post-training dimensionality=6.29, two-sample t test, $t(98)=9.39$, $p=2.51\times10^{-15}$). For NMS defined as the location x epoch effect, i.e. different selectivity for a stimulus when it appears first in the sequence vs. second in the sequence dimensionality increased after training for the stimulus presentation period (pre-training dimensionality= 2.58, post-training dimensionality=3.61, two-sample t test, $t(98)=6.42$, $p=5.01\times10^{-9}$) as well as in the

delay period (pre-training dimensionality= 2.67, post-training dimensionality=2.91, two-sample t test, t(98)=5.76, $p=2.51\times10^{-8}$). For the feature task however, no significant increase was observed in the mean F score (Fig. 4A) or dimensionality (Fig 4B). This was true for both the shape x matching comparisons (sample period: pre-training dimensionality= 2.72, post-training dimensionality=2.73, two-sample t test, t(98)=0.027, p=0.978; delay2 period: pre-training dimensionality= 2.43, post-training dimensionality=2.11, two-sample t test, t(98)=3.49, $p=7.29\times10^{-4}$) and the shape x epoch comparisons (stim period: pre-training dimensionality= 3.08, post-training dimensionality=2.31, two-sample t test, t(98)=10.27, $p=1.36\times10^{-17}$; delay period: pre-training dimensionality= 2.18, post-training dimensionality=2.11, two-sample t test, t(98)=0.84, p=0.40). In accordance with previous research (Rigotti et al., 2013), we found that the increased dimensionality in the spatial task was accompanied by an increased number of implementable binary classifications (Fig. 4D), as well as enhanced decoding after pure selectivity was removed (Fig. 4C).

Previous studies of nonlinear mixed selectivity have suggested that increased linear classification boundary choices would be beneficial for downstream readout cells, which are mostly likely action related. We thus investigated the saccade related signals in recorded cells to determine whether action information exists preferentially in different selective categories (CS vs NMS), by defining different types of selectivity (CS, LMS, NMS) through the same method we used for the sample and delay2 period. Although there was strong selectivity for saccade locations, we ultimately found very little NMS for saccade direction and

matching status (Fig. 11A). Moreover, linear cells were found to contain more decodable information for saccade direction, even after controlling for the large proportion difference (Fig. 11B).

Regional localization of NMS

To assess whether specific sub-regions of the PFC may be specialized for NMS, we divided the lateral PFC into regions (Fig. 5 A) and analyzed the respective neurophysiological data from five of these regions in order to determine the different areas' proportional contributions to the observed changes in NMS. We examined NMS defined by location/shape and match/nonmatch status in the sample period and ultimately found that the mid-dorsal subdivision underwent the greatest proportional change in NMS cells for the spatial task after training (Fig. 5 C), without a comparable increase in the proportion of CS neurons (Mid-dorsal: CS 21.7% pre-training to 19.0% post- training, NMS 8.2% pre-training to 21.8% post-training). For the feature task however, the proportional change in NMS cells was relatively small, with moderate increases in CS and LMS observed in all three analyzed areas (Fig. 5 B).

NMS in task context

Previous theoretical studies linked NMS with more flexible readouts of multiple task variables, thus leading to the hypothesis that task complexity may modulate NMS.

To test this hypothesis, we compared the neural responses to different shapes at the same location when the stimuli appeared as match or nonmatch in the conjunction task, to the same neurons' responses to the same stimuli when they appeared in the feature task. In the conjunction task, animals needed to simultaneously remember both location and shape of visual stimuli, while in the feature task, they were only required to remember shape. Although the hypothesis predicted that the conjunction task would result in greater NMS than the feature task when the sensory stimuli was the same, this was not what we observed. No significant differences were observed for either CS cells or NMS cells in either the sample period or the delay2 period that followed. For CS cells in the sample period, the feature task proportion was 11.9%, whereas the conjunction task proportion was 9.9% (exact matched pair sample proportion test, F=1.229, p=0.197); in the delay2 period the respective proportions for the feature task was 10.8%, and for the conjunction task, 11.6% (exact matched pair delay2 proportion test, F=1.069, p=0.681). For NMS cells in the sample period the feature task proportion was 4.1%, and the conjunction task proportion was 4.4% (exact matched pair sample proportion test, F=1.031, p=0.901). In the delay2 period, the feature task proportion= 4.6%, conjunction task proportion=5.9% (exact matched pair delay2 proportion test, F=1.278, p=0.266) (Fig. 6 A).

We also examined changes in individual cells' selectivity across the feature and conjunction tasks. A change in selectivity type was frequent between tasks; most CS and NMS cells defined as such as in the feature task, changed their

selectivity type in the conjunction task (CS cells: 85.2% in sample period, 77.2% in delay2; NMS cells 88.6% in sample period, 89.7% in delay2). The finding implies that CS and NMS selectivity is task specific, and an unstable mapping exists between different tasks (Fig. 6 B). Although the percentage of NMS neurons was close to that expected by chance, NMS cells with a larger degree of interaction in one task tend to also fall into the NMS category in the other task (Fig. 6 C). No significant difference was observed in the proportion of NMS cells when we performed a similar comparison between the spatial and conjunction task; the proportion of NMS cells in the spatial task was 18.3%, whereas in the conjunction task it was=14.1% (two-sample proportion test, $z=0.68$, $p=0.494$). In the delay2 period, the same proportions were 7.0% and 11.3%, respectively (two-sample proportion test, $z=0.88$, $p=0.381$).

The comparison of the naïve and trained conditions allowed us to test the overall incidence of NMS in different populations of PFC neurons, sampled randomly before and after training, which was carried out over the course of several months. If NMS were critical for the representation of task-relevant information, we would expect more neurons to exhibit NMS, when animals are passively viewing stimuli vs. when they are actively performing the task and storing representations of the stimuli in their WM. We therefore applied a two-way ANOVA to compare the neural responses of neurons between the active and passive spatial tasks after the monkeys had been trained to perform the active spatial task. We ultimately observed a small but not significant increase in the proportion of cells that coded

matching status during the sample period, as well as an increase in the proportion of cells coding sensory information in the delay1 period when the animal was prompted to report the matching decision, (passive proportion= 9.3%, active proportion=11.7%, exact matched pair proportion test, F=1.385, p=0.362) (Fig. 7 A). Interestingly, a large proportion (CS: 81.8% , NMS:52.2%) of cells changed their selectivity category across tasks , especially for CS cells (Fig. 7 B), and the degree of NMS does not seem to be predictive of whether a given neuron would fall in the same selectivity category in both tasks (Fig. 7 C).

Information encoding by NMS neurons

Prior research has demonstrated that training leads to increases in both persistent activity and the incorporation of task relevant information in neural populations and these effects were more pronounced during the delay2 period in our task (Meyers et al., 2012). We therefore wished to test whether the increase of neurons exhibiting NMS would be greater in the second delay interval, as well. The current study revealed that this was not the case. The result suggests that the plasticity involving the increase in delay period activity and the increase in the magnitude of NMS are independent of each other.

We also investigated the relative contribution of NMS to encoding new task information. We used a linear SVM decoder to decode sensory information (location and shape) and match/or nonmatch status information to quantify the amount of task relevant information contained in linear (CS and LMS) and NMS cells. Since the

cell selectivity category could be defined by their response in either sample or delay2 period, we randomly selected equal numbers of linear and nonlinear cells in both task epochs for each comparison. The random selecting process was repeated multiple times to obtain a confidence interval. We ultimately found that linear and nonlinear cells contain comparable amounts of linearly decodable information in regard to both sensory information and task relevant information. The only observed difference between the decodable information in the linear and nonlinear cells occurred in the post training feature task, where linear cells were observed to contain more stimulus information (mean performance 46.6% for CS, 29.6% for NMS during sample period) in the sample period.

We applied cross temporal decoding to compare classic and linear mixed cells with regard to population coding dynamics during the delay period. If information were represented by a stable pattern of activity, the classifier trained at one timepoint would be expected to work equally effectively at other time points where the information is present. Conversely, if information were represented by dynamic patterns of activity, then the decision boundary at one time point would not contribute to decoding information at other time points. The most prominent result from this analysis is that NMS cells produced significantly more stable code for matching information in the spatial task, compared to CS and LMS cells, as indicated by higher performance off the diagonal during the delay2 period (Fig. 9). The contrast between the CS+LMS vs NMS cells supported the idea that NMS is particularly important for downstream readout. A stable code is generally easier to

read-out, since it does not require the downstream circuit to track the time elapsed in order to decode information. In the context of our task, the sensory information is no longer necessary during the delay2 period for the behavioral response, posing a likely explanation for why only the code for match/nonmatch information is stable in the delay2 period for NMS cells. Importantly, this stability was not observed in CS+LMS cells, despite the matching status also being represented in these cells. Another intriguing and related phenomenon revealed by our decoding analysis (Fig. 8B) is that stimulus information could be decoded above chance level for the spatial, but not the feature task in the delay2 period, although, as noted, maintenance in memory of either the first or second stimulus is no longer necessary at this stage. This supports the idea that PFC processes the spatial and shape information differently in the regions we recorded from.

NMS in correct and error trials

The presence of decodable information in the PFC does not necessarily imply that the information is utilized by the subjects to guide behavior. In order to decipher the role of NMS in guiding behaviors, we examined the F score of the main effects and their interaction in the ANOVA test in correct vs. error trials for the spatial task (Fig. 10 A,C), which displayed higher NMS levels than the feature or conjunction tasks. Similar to the pre- vs. post training comparisons, we examined two types of mixed selectivity: stimulus location vs. matching status and stimulus identity vs. task epoch. The number of trials and task variables were matched for each cell to avoid

confounds in the comparison. The mean F score in correct trials for the location variable in the stimulus epochs for the location × epoch comparison was equal to 1.86, while for the error trials the mean was 2.59 (paired t test, $t(147)=3.38$, $p=9.42 \times 10^{-4}$). The effect extended into the delay epochs, where the average F score for location in correct trials was 1.43, and that of error trials was 1.79 (paired t test, $t(150)=2.61$, $p=0.010$). However, we did not find any differences in the F score for the interaction terms in any comparison. The increased F score in error trials for stimuli location may reflect higher variability in error trials in coding or maintaining corresponding information. Alternatively, the increased F score may reflect erroneous association of a combination of conditions. In either case however, dimensionality collapse does not seem to be the primary cause of errors. A direct measurement of neural response dimensionality revealed that error trials with the most observed NMS did not undergo a decrease of dimensionality (active spatial task, Fig. 10 B,D).

Our analyses above reveal no obvious decrease of NMS representing stimulus identity and match status in error trials. However, to determine whether NMS may be involved with converting match status to action in our task, we decoded saccade direction from correct and error trials. If neurons exhibiting NMS (now defined by the interaction of saccade location and match status) could predict erroneous saccade direction more effectively, we might be able to conclude that NMS plays an important role in behavior, despite the lack of dimensional collapse for stimuli presentation and delay error trial representations of stimuli identity and

matching status. However, our results ultimately failed to indicate a significant decrease of such NMS in error trials, either (Fig. 11D).

DISCUSSION

Selectivity for different types of information is critical in representing the plethora of stimuli and task contexts that can be maintained in WM. NMS is thought to be critical in that respect, as it allows efficient representation of flexible, arbitrary combinations of variables (Buonomano and Maass 2009, Rigotti, Ben Dayan Rubin et al. 2010, Barak, Rigotti et al. 2013, Fusi, Miller et al. 2016, Johnston, Palmer et al. 2020). Consistent with this idea, increased dimensionality in NMS has been highlighted as a potential means of increasing the efficiency of WM task performance (Rigotti, Barak et al. 2013, Johnston, Palmer et al. 2020) and dimensional collapse characterizes task errors (Rigotti, Barak et al. 2013). Moreover, all task relevant information could be decoded from NMS neurons alone, despite their relative scarcity, with decoder accuracy actually increasing as the task became more complex (Rigotti, Barak et al. 2013). NMS is assumed to emerge with training in complex tasks that combine multiple types of information, or in multiple tasks, even without an explicit requirement to combine such information (Lindsay, Rigotti et al. 2017, Johnston, Palmer et al. 2020). However, this idea has not been tested experimentally until now. Our study, by virtue of analyzing neural recordings before and after training in a series of cognitive tasks, directly tested these postulates. We found that NMS resulted in a modest increase with training, but only for some tasks, and furthermore, task complexity was not a predictor of NMS emergence. A causal relationship between success and dimensionality—and by

extension, NMS—was also not supported by our results, as we did not observe any significant changes in NMS between error and success trials. These insights refine and qualify the role NMS plays in WM, and identify a number of open questions.

Effects of training on neural responses

WM is considerably plastic and at least some aspects of it, such as mental processing speed and the ability to multitask, can be improved with training (Klingberg, Forssberg et al. 2002, Klingberg, Fernell et al. 2005, Bherer, Kramer et al. 2008, Jaeggi, Buschkuehl et al. 2008, Dux, Tombu et al. 2009). WM training has been proven particularly beneficial for clinical populations, e.g. in the case of traumatic brain injury, attention deficit hyperactivity disorder (ADHD), and schizophrenia (Klingberg, Forssberg et al. 2002, Westerberg, Jacobaeus et al. 2007, Subramaniam, Luks et al. 2012). However, the verdict of whether WM training confers tangible benefits on normal adults and whether these benefits transfer to untrained domains, remains a matter of heated debate. (Fukuda, Awh et al. 2010, Owen, Hampshire et al. 2010, Cortese, Ferrin et al. 2015, Peijnenborgh, Hurks et al. 2015, Schwaighofer, Fischer et al. 2015, Constantinidis and Klingberg 2016).

This malleability of cognitive performance is thought to be mediated by the underlying plasticity in neural responses, most importantly within the PFC (Constantinidis and Klingberg 2016). In a series of prior studies, we have investigated changes in PFC responsiveness and selectivity (Meyer, Qi et al. 2011, Qi, Meyer et al. 2011, Meyers, Qi et al. 2012, Riley, Qi et al. 2018), as well as other aspects of neuronal discharges such as trial-to-trial variability and correlation

between neurons (Qi and Constantinidis 2012, Qi and Constantinidis 2012). This led to our present analysis where—guided by experimental and theoretical predictions (Rigotti, Barak et al. 2013)—we examined NMS as another potential source of enhanced ability to represent WM information after training.

In agreement with our hypothesis, we found that training increased the proportion of neurons that exhibit NMS. However, training does not seem to be a prerequisite, as NMS was also observed in animals that were naïve to any cognitive training. Prior research has established that the human and primate PFC represent stimuli in memory even when not prompted to do so (Foster, Bsales et al. 2017), or without training in WM tasks (Meyer, Qi et al. 2007). Our finding of NMS neurons in naïve monkeys provides another exemplar of that principle. However, NMS only increased for certain types of task information and not for others. Task variables mixed significantly more with spatial over shape information, thus suggesting that the role of NMS is not universal for all types of working memory. Considering the fact that the spatial information in current study is largely nonoverlapping and low dimensional (stimulus identities could be represented by one or two numbers), while the shape information is relatively high-dimensional, one possible explanation of the observed difference is that the distinction in the degree of NMS may reflect the qualitatively different ways of representing information with different dimensionality in working memory. In any case, our results suggest that NMS may not be as ubiquitous as previously believed, across all tasks, all prefrontal areas, and individual subjects. These results are consistent with some prior studies

that have also failed to uncover substantial NMS in the tasks they employed (Cavanagh, Towers et al. 2018). Examining where NMS failed to appear—and where WM representations fail to spontaneously appear— will be an important area of future investigation for NMS.

The greatest increase in the proportion of neurons that exhibited NMS for the spatial working memory task was observed at the mid-dorsal region during the sample presentation period (Fig. 5). This disproportionate increase in NMS neurons was associated with a modest decrease in neurons that exhibit CS, as predicted by theoretical studies (Lindsay, Rigotti et al. 2017). However, this finding, too did not generalize across conditions. During the delay periods of the spatial task, we saw an across-the-board increase in neurons with CS, which were much more abundant in the trained than the naïve PFC (Fig. 3). In fact, the increase of encoding of matching information in the feature task after training was driven almost exclusively by CS cells, suggesting a potential division of labor between NMS and CS in the PFC for different types of information.

Task Complexity and Difficulty

Another potential factor that determines the emergence of NMS is task complexity. NMS may arise exclusively in highly complex tasks that require subjects to maintain and combine multiple types of information in their WM, simplifying the involved neural circuits to achieve greater efficiency (Rigotti, Barak et al. 2013). We thus tested this concept by applying a dataset that relied on three tasks which differed in

complexity (and overall difficulty). The spatial and feature tasks each required maintenance of a single stimulus property in memory (location or shape). The conjunction task required both. Surprisingly however, we did not observe a higher incidence of NMS in the conjunction task when compared to the feature task. Moreover, we observed a much lower incidence of NMS in the feature task compared to the spatial task despite the fact that the latter was no more complex or difficult for the monkeys to perform (Meyer, Qi et al. 2011, Riley, Qi et al. 2018). This implies that NMS in the PFC may not be necessary for certain types of information, like object shape, even when the task complexity is high. Our tasks required modest flexibility for stimulus representations: presentation of two stimuli in sequence requires choice of the green or blue target depending on their relative location, shape, or location-shape combination. However, the flexibility of the task was bounded by the fact that the same basic match/nonmatch rule applied across all tasks. It is possible that a task that required more flexible representation (e.g. if the monkeys were additionally cued in each trial to select the green choice target for either the match or the nonmatch) would result into emergence of even more neurons with mixed selectivity. Future research is therefore necessary to assess this possibility.

Regional Specialization

Different types of information are represented across the dorso-ventral and anterior-posterior axes of the PFC (Constantinidis and Qi 2018), and examining the

regional distribution of NMS neurons within the PFC therefore bears a clear importance. We found that NMS was most strongly demonstrated in the mid-dorsal area for the spatial task and the posterior dorsal area for the feature task. The specialization of different PFC subregions in processing location versus feature information is still a topic of debate. Current evidence suggests that the degree of specialization may depend on what stimuli one is using. For highly specific stimuli that require within-category discrimination like faces, the selective cells are clustered in anatomically defined patches in the ventral PFC (O Scalaidhe, Wilson et al. 1997, O Scalaidhe, Wilson et al. 1999), while for more general stimuli varying in shape and color, the coding seems to be distributed in both the dorsal and ventral portions of the PFC (Rao, Rainer et al. 1997, Meyer, Qi et al. 2011, Kadohisa, Kusunoki et al. 2015). In the current study we used simple geometrical shapes to probe the selectivity for feature, and we sampled from both dorsal and ventral portion of PFC. It is noticeable that in the current study, the encoding of spatial information is stronger compared to feature information in the sampled neurons. This is in agreement with previous reports that the degree of selectivity is stronger on a single neuron level for location compared to simple geometrical shapes in the dorsal portion of the PFC (Constantinidis and Qi 2018).

Information Content and Task Performance

A critical issue regarding the role of Mixed Selectivity is whether nonlinear match selectivity, by virtue of representing information more efficiently, is also more

necessary for effective task performance (Rigotti, Barak et al. 2013). We relied on a linear SVM decoder to decipher the specific information that may be represented by NMS cells, compared to CS cells. In the current study, we found similar quantities of information could be decoded from (equal-sized) populations of CS and NMS neurons, though the coding dynamics for some types of information were significantly different between CS and NMS cells. Similarly, when we compared the NMS levels of successful and failed task trials, we were surprised to find that there was no appreciable difference. This suggests that loss of information encoded in a nonlinear manner may not be the primary factor of successful WM guided behavior. An important caveat for this conclusion is that the combination of small and unbalanced number of error trials in match vs. nonmatch conditions make the detection power for the interaction fairly small in our analysis. Moreover, with very little NMS presented even in success trials after matching the trial number for the error condition, a floor effect may have prevented a further decline from becoming apparent. Nonetheless, our result reinforces the idea that NMS is not necessary in all tasks, without which performance fails. An interesting observation in this analysis was that the proportion of cells tuned to spatial location was elevated in error trials. The result may imply that task success also depends on the task relevance of the represented information, with error trials incorporating greater quantities of task irrelevant spatial information and therefore unnecessarily drawing away WM resources without benefit. Ultimately, by comparing and evaluating the

conditions in which NMS emerges, we may decipher its true role in WM and other cognitive functions.

FIGURES

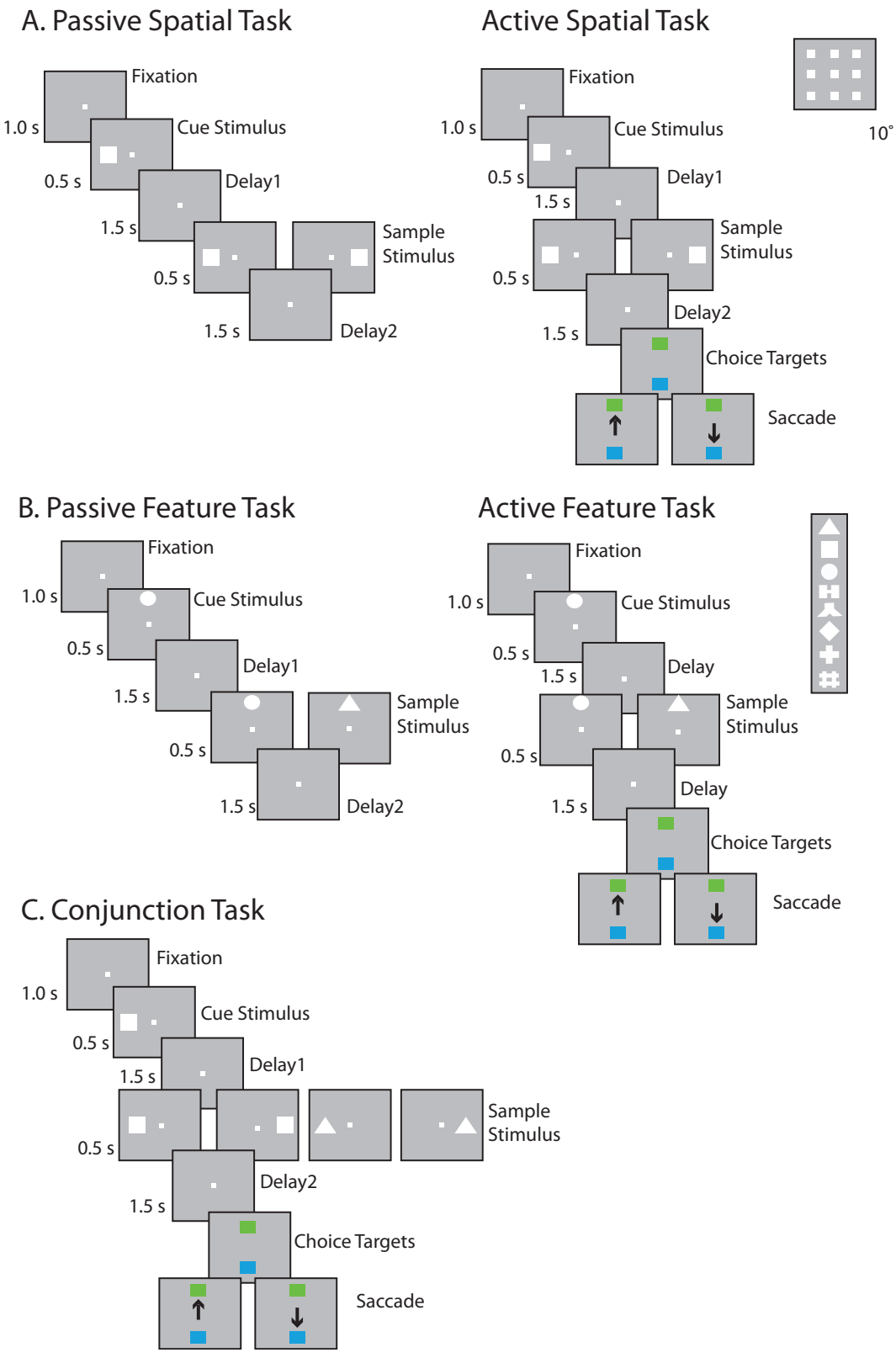

A. Passive Spatial Task
Fixation
1.0 s
Cue Stimulus
0.5 s
Delay1
1.5 s
Sample Stimulus
0.5 s
Delay2
1.5 s
Active Spatial Task
Fixation
1.0 s
Cue Stimulus
0.5 s
Delay1
1.5 s
Sample Stimulus
0.5 s
Delay2
1.5 s
Choice Targets
Saccade
10°
B. Passive Feature Task
Fixation
1.0 s
Cue Stimulus
0.5 s
Delay1
1.5 s
Sample Stimulus
0.5 s
Delay2
1.5 s
Active Feature Task
Fixation
1.0 s
Cue Stimulus
0.5 s
Delay
1.5 s
Sample Stimulus
0.5 s
Delay
1.5 s
Choice Targets
Saccade
C. Conjunction Task
Fixation
1.0 s
Cue Stimulus
0.5 s
Delay1
1.5 s
Sample Stimulus
0.5 s
Delay2
1.5 s
Choice Targets
Saccade

Figure 1. Task structure and stimuli used. The animals were required to maintain center fixation throughout both active and passive task trials. At the end of active tasks trials however, monkeys were required to make a saccade to a green target if the stimuli matched or to a blue target if the stimuli did not match. (A) Spatial location match-to-sample task, nine possible cue locations in a session shown in the inset. (B) Shape feature match-to-sample task, 8 possible shapes in a session shown in the inset. (C) Spatial-shape conjunction task, up to two locations and two stimuli shapes were used for any single particular session. Stimuli in all tasks extended 2 degrees of visual angle.

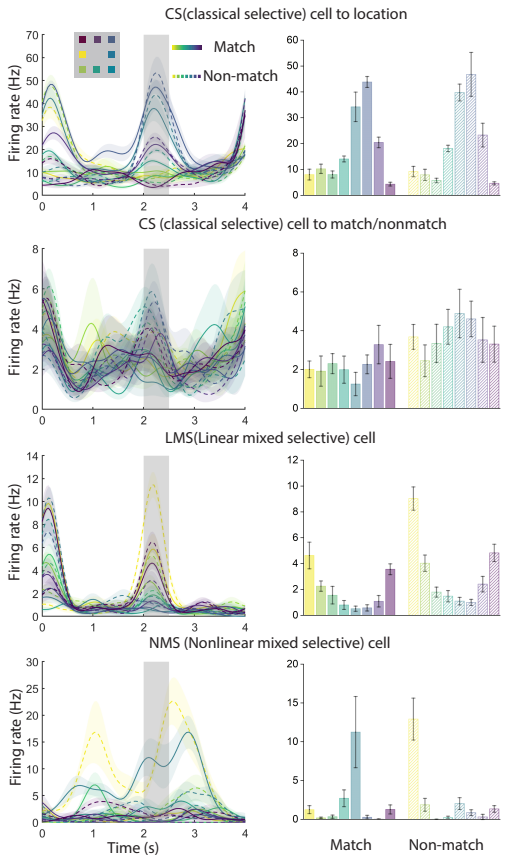


Figure 2. Exemplar neural responses from the spatial task for CS (classical selective), LMS (linear mixed selective) and NMS (nonlinear mixed selective) cells, defined by the task variables of stimulus location and match status. Selectivity classification were based on the spike responses of the 500 ms sample period (indicated by the gray shaded area from 2s to 2.5s). The locations the stimuli were color coded, with a solid line/bar representing when the stimulus was a match with the cue, and a dash line/bar representing when the stimulus was a nonmatch with the cue. Shaded regions and error bar indicate 2 times SE of firing rate.

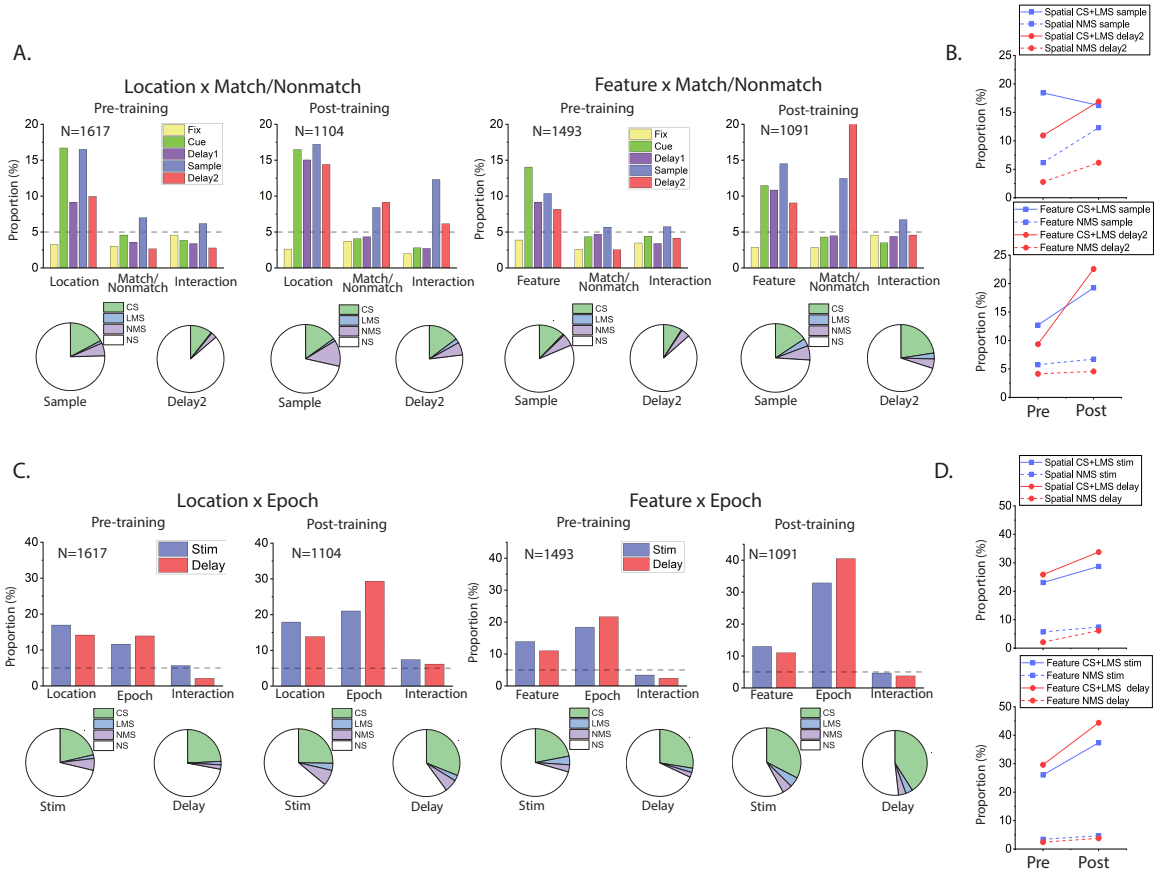


Figure 3. Training increased mixed selectivity preferentially in the spatial task. (A) Bar graphs show the proportions of cells tuned to stimuli identities (Location/Shape), matching status and their interaction (i.e. NMS) in different stages of the task trials, both before and after the animals were trained for the active tasks. Pie charts show the proportion of different selectivity categories (NS, CS, LMS and NMS) in the sample and delay2 periods of both tasks, both before and after the animals were trained for the active tasks. The green areas in the pie chart are combined proportion of classically selective cells for both factors used in the two-way ANOVA(B) Plots of corresponding proportion changes. (C) Same as (A) but examining the interaction between stimuli identities (Location/Shape), and task

epoch (cue/delay1 vs sample/delay2 period), instead of trials matching status. (D) Plots of corresponding proportion changes.

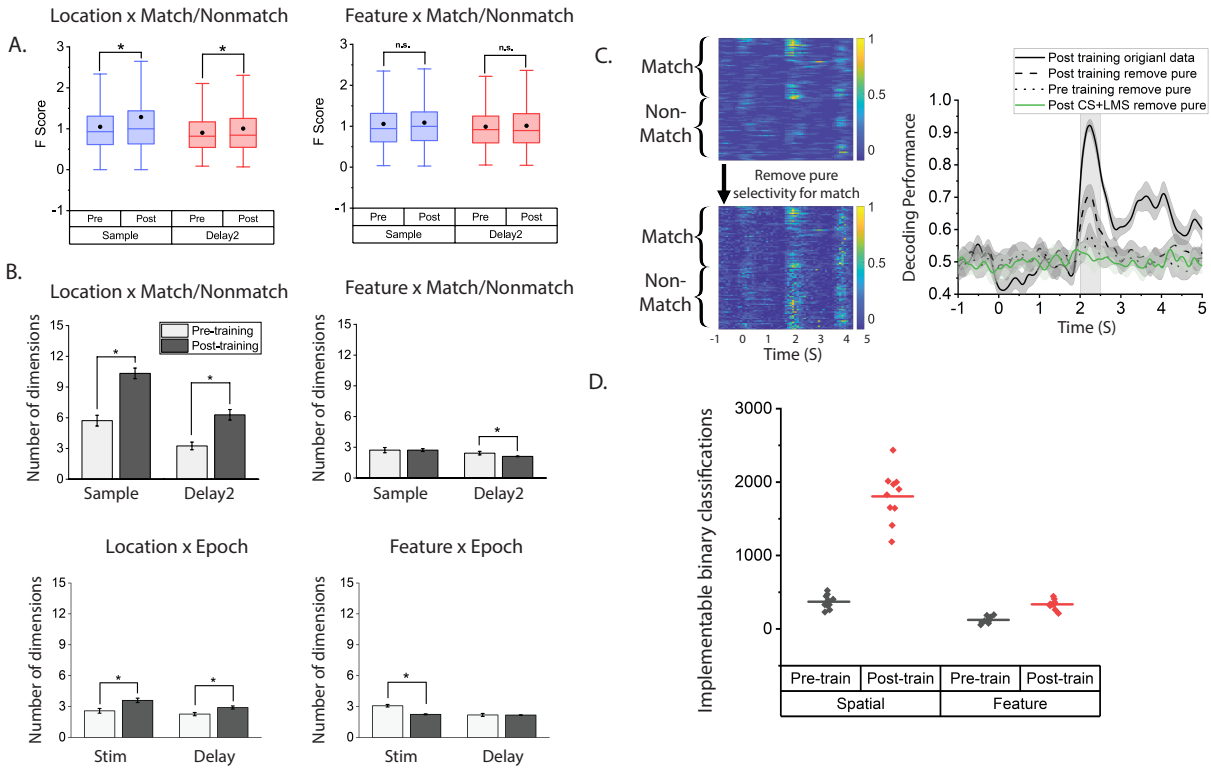


Figure 4. (A) Analysis of the F scores from all recorded neurons for the interaction term (stimuli identity × match/nonmatch) shows that the degree of mixed selectivity increased after training for the spatial task only. Black dots in the box represent mean, box boundaries indicate 25%-75% range, and whiskers represents 1.5 IQR. (B) Dimensionality measure of neural responses in the spatial (left) and feature (right) task, before and after training in the active tasks. To measure dimensionality, SVD was applied to a matrix containing data across all task conditions, based on the factors used for each analysis. In the case of Location x Match/Nonmatch NMS analysis, data included sample firing rates for all location x Match/Nonmatch combinations. For Location x Epoch NMS analysis, data included firing rates for different locations in either the cue or the sample period of match trials. (C)

Decoding with pure-selectivity removed. Left: An example of an NMS cell in the sample period with pure selectivity to match/nonmatch status removed. Color in the heat maps coded for the normalized neuronal response ranging from 0 (blue) to 1(yellow). The y axis in the heat map organizes trials by matching status, and then by stimuli location. Right: Decoding for matching status for the spatial task before and after removing pure selectivity of match/nonmatch status across pre and post-training. 200 cells in each dataset were randomly chosen to construct ten pseudo-populations in order to calculate the confidence interval. Nonlinear information increased after training. (D) Number of implementable binary classifications for different tasks and in different training stages. Data from sample period for both tasks.

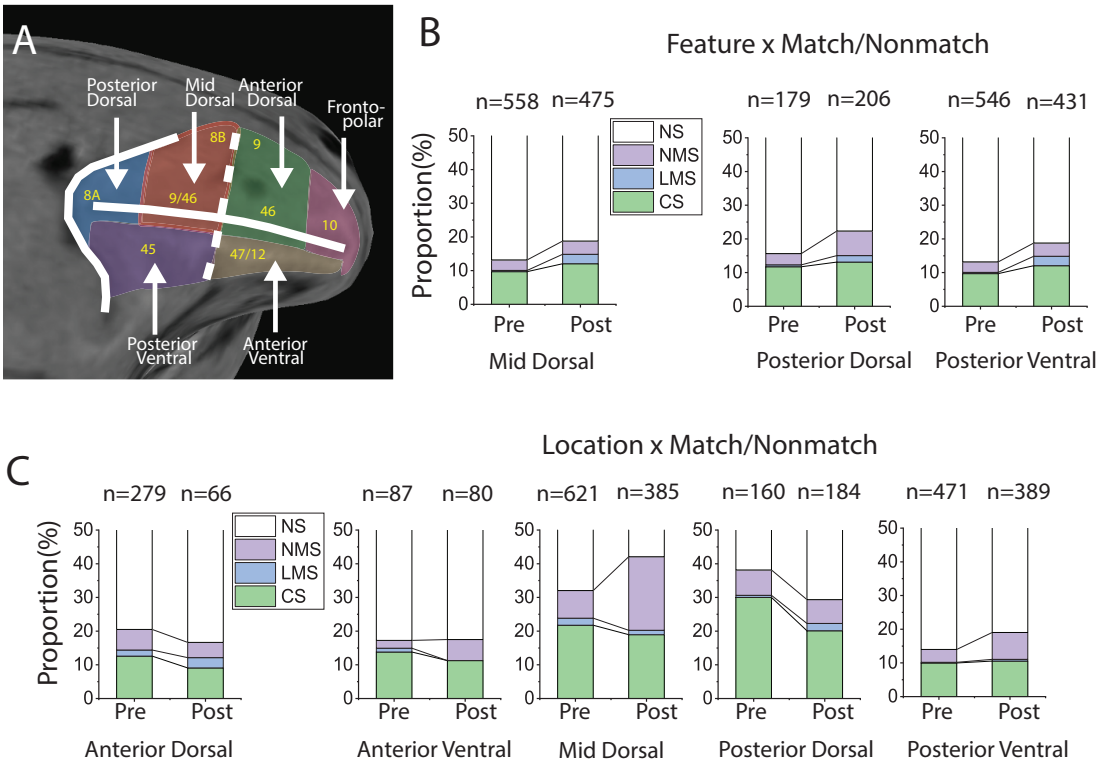


Figure 5. Cell selectivity changes by brain regions. (A) PFC subdivisions that were utilized for recording in the current study. (B) The effects of training in for the active feature task on the proportion of different selectivity categories (NS, CS, LMS and NMS) in the sample period. There were significant increases in the proportion of LMS cells in all three PFC regions included for analysis, but relatively low increases in the proportion of NMS cells. (C) The effects of training in the active spatial task on the proportion of different selectivity categories (NS, CS, LMS and NMS) in the sample period. The greatest increase in NMS occurred at the mid-dorsal region.

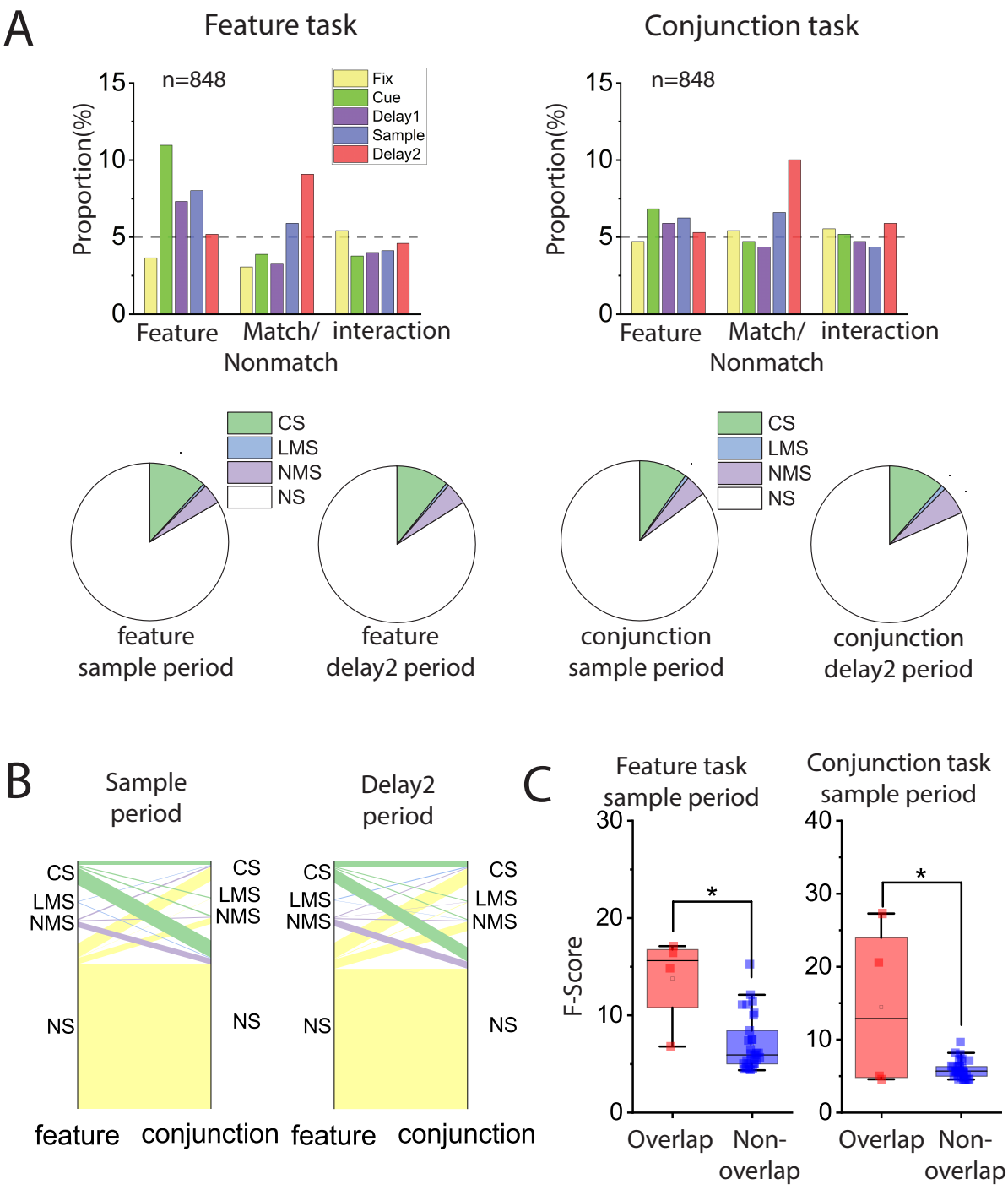


Figure 6. Cell selectivity in tasks with different task complexity. Analyses were performed on neural data from the same population of cells, with matching numbers of trials in the feature and the conjunction tasks. Only trials with the same

stimuli were included in this analysis. (A) Examining interaction (NMS) across stimulus preference and matching status. Bar graphs show the proportions of cells tuned to stimuli shape, trials matching status and their interaction in different stages of both the feature and conjunction tasks. Pie charts display the proportion of different selectivity categories (NS, CS, LMS and NMS) in corresponding sample and delay2 periods. (B) Flow diagram shows cell selectivity category mapping across tasks. On the left side of each plot, cell selectivity category (NS, CS, LMS and NMS) in the feature task was color coded. The thickness of lines represents number of cells. The right side of each plot represents selectivity categories of the same cells in the conjunction task. The composition of cell selectivity category with reference to the other task is shown by proportion of different colored lines (C) F scores of the interaction term in the ANOVA were compared between cells that were classified as NMS cell in both tasks (overlapping cells), and those only classified as NMS in one of the tasks (non-overlapping cells).

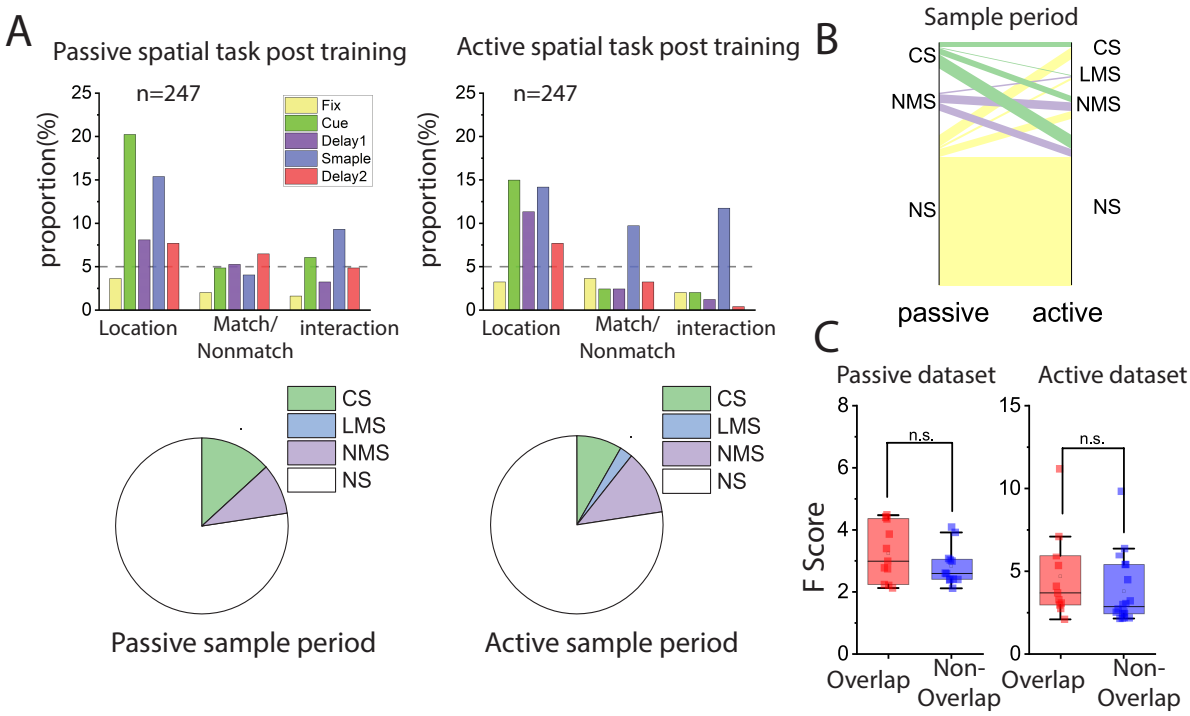


Figure 7. Cell selectivity in tasks with the same sensory input but different behavioral requirements. Analyses were performed on neural data from the same population of cells, with matching number of trials in the passive and active spatial tasks. Only trials that had the exact same stimuli pairs in both tasks were included in this analysis. (A) Examining interaction (NMS) between stimulus preference and matching status. Bar graphs show the proportions of cells tuned to stimuli location, trials matching status and their interaction in different stages of the tasks. Pie charts display the proportion of different selectivity categories (NS, CS, LMS and NMS) in the sample period. (B) Flow diagram shows cell selectivity category mapping cross tasks. On the left side of each plot, cell selectivity category (NS, CS, LMS and NMS) in the passive task were color coded. The thickness of lines

represents the number of cells. On the right side of each plot, cells belonging to different selectivity categories in the active spatial task were clustered together, so the composition of cell selectivity category with reference to the other task is shown by proportion of different colored lines. (C) F scores of the interaction term in the ANOVA were compared between cells that were classified as NMS cell in both tasks (overlapping cells), and those only classified as NMS in one of the tasks (non-overlapping cells).

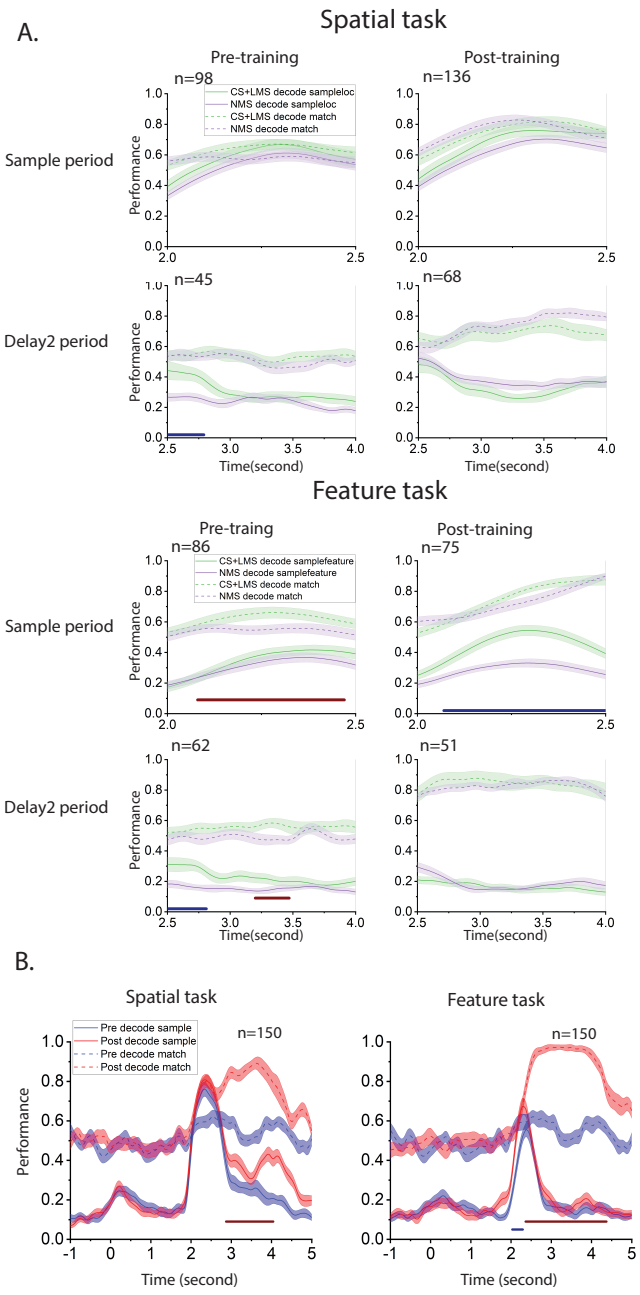


Figure 8. (A) Decoding task variables with an equal number of linear (CS+LMS) and nonlinear cells (NMS), at different task epochs and different training stages. The selectivity categories (CS, LMS and NMS) were defined by spiking count in corresponding epochs (sample and delay2) with reference to selectivity to stimuli identity (stimuli location or shape) and matching status. Decoding performance for

different task epoch and training stage combinations are displayed in separate subplots. In each subplot, the y axis represents decoding performance, with results for stimuli identity and matching status represented by solid and dash line respectively. Cell selectivity categories are color coded (green for CS+LMS, purple for NMS). Numbers on the x axis represents time relative to onset, with the sample period occurring from 2-2.5 seconds, and delay2 period occurring from 2.5-4 seconds. Red bars under each subplot indicate the time points when the decoding performance for the matching status differs significantly between linear and nonlinear cells. Blue bars under each subplot indicate the time points when the decoding performance for the sample stimulus identity differs significantly between linear and nonlinear cells. For the spatial task, we found that despite the significant change in proportion of NMS cells after training, NMS does not seem to represent increased quantities of linearly decodable information per cell compared linear cells. For the feature task, linear cells are generally more selective in many conditions, especially when coding stimuli identity after training in the active task. (B) Incorporation of new information after training for the spatial and the feature tasks. Linear SVM decoders were trained to classify either stimulus identity or matching status with z-score normalized pseudo-population response. Color bars indicate time points when the performance for NMS and linear cells differs significantly; red bar for decoding matching status, blue bar for decoding stimuli identity.

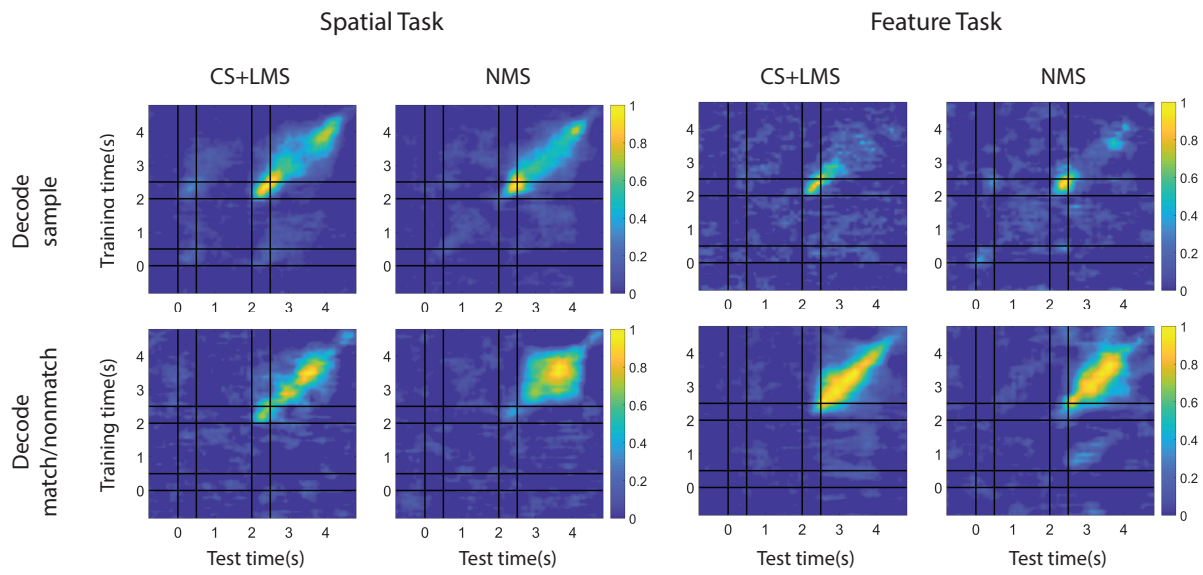


Figure 9. Coding dynamics of pure and linearly selective cells (CS and LMS) vs. NMS cells. Linear kernel SVM decoders were trained to perform cross-temporal decoding with different selectivity populations in the delay2 period, for both spatial and feature tasks, as indicated by the Y-axis. The decoder was then required to predict whether a match or non-match occurred at each time point based on a different test set of data, as indicated by the X-axis. Normalized decoding accuracy is indicated in the color bar, demonstrating how spatial and feature WM representations can be decoded from specific patterns of neural activity. Coding of matching information for NMS cells is more stable across time for the spatial task.

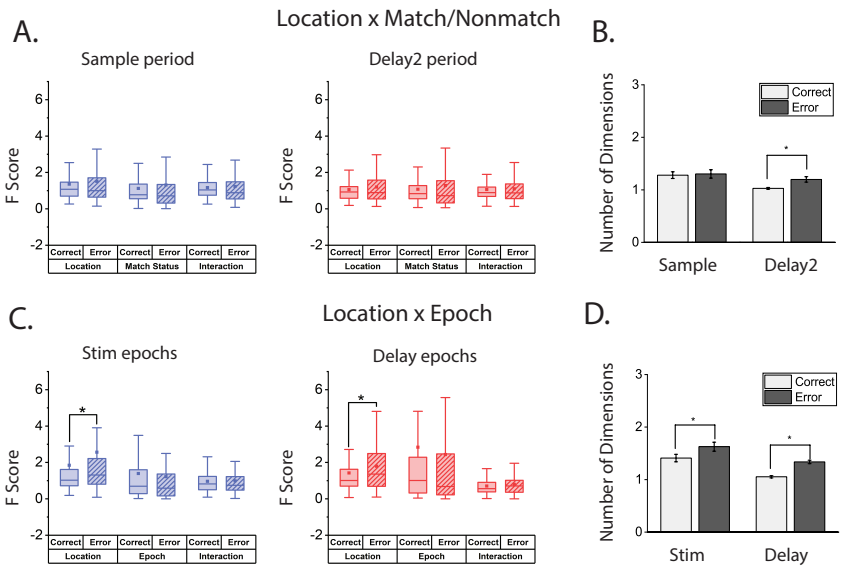


Figure 10. (A,C) Comparison of cell selectivity in correct and error trials in the same population for the spatial task, after controlling for trial number and location pairs used. Two forms of mixed selectivity were examined (location x matching, location x task epoch). No change (location x matching delay2 period) or increase (location x task stim and delay epochs, location x matching sample period) in the mean F score of the interaction term of ANOVA results from all recorded cells were observed. Higher F score for the variable of stimuli location was observed in error trials in the location x Epoch comparison. Box boundaries represent 25%-75% data range, whiskers indicate 1.5 IQR and squares indicates means across cells. (B,D) Dimensionality measure for the correct and error dataset in the spatial task, with different definitions of mixed selectivity(Left: location x match/nonmatch; Right: location x Epoch). No decrease of dimensionality was found in error trials.

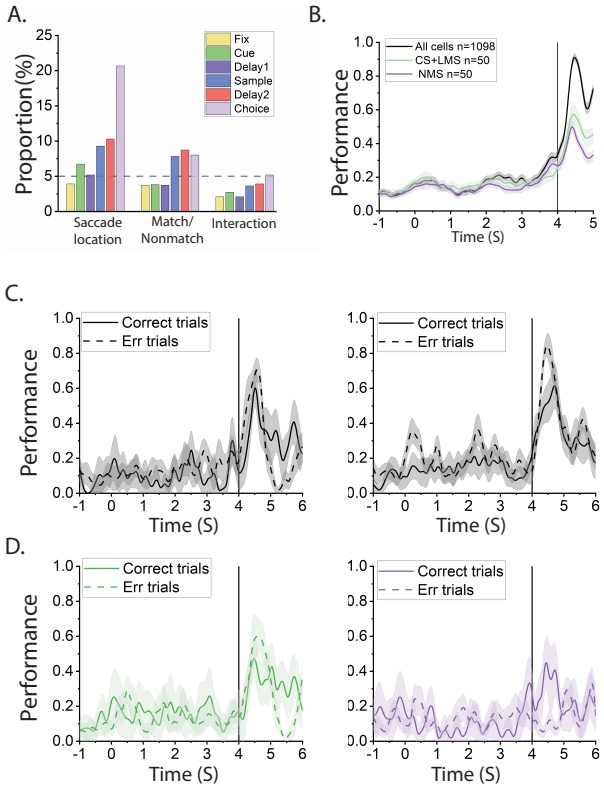


Figure 11. (A) Bar graphs show the proportions of cells tuned to saccade location, matching status and their interaction (i.e. NMS) in different stages of the task trials, for the active spatial task. (B) Decoding of saccadic direction using all recorded cells, or the same number of CS+LMS/NMS cells in the active spatial tasks. Time 0 at the x-axis indicted the onset of the cue. Black line indicates onset of choice array. (C) Saccadic direction decoding was equally effective in correct and error trials, regardless of pseudo-population size. Left: Decoding saccadic direction using 239 cells with 4 randomly selected correct and error trials (2 trials in 2 the same 2 conditions are randomly selected, thus resulting in 4 trial for each cell. Condition was defined by saccade location x matching status).Right: Decoding saccadic direction using 155 cells with 2 correct and error trials in the same conditions. (D)

Decoding saccadic direction using different populations from the same cells in (C). Left: Decoding performance using 54 CS+LMS cells with 4 correct and error trials. Right: Decoding performance using 10 NMS cells with 4 correct and error trials. In all panels, the shaded range represents 95% confident interval for 10 randomly constructed pseudo-population. Time point 0 is the onset of cue, and the black line at 4 seconds represents the onset of the choice array.

Chapter 3: Prefrontal Mechanisms of Rule Learning

ABSTRACT

Rule learning is associated with lasting changes in prefrontal activity. However, experiments typically focus on learning a single set of rules or task and there remains a significant gap regarding how related mechanisms may reflect behavioral improvements in different contexts, especially when examining across different tasks and modalities. We therefore recorded single units from chronic electrode arrays implanted in the prefrontal cortex of four monkeys as they were trained to perform spatial and object working memory tasks with the goal of assessing the resulting activity changes that would be induced by rule learning. Progression of training allowed behavioral improvements to be correlated with a variety of neural effects that could be observed across different task contexts, including increases or decreases of both firing rate and decoding, increases in the proportion of firing rate variance that was unexplained by sensory stimuli and motor actions, and the increased separation of population response trajectories in state space. Our results thus reveal how rule learning induces plasticity of prefrontal cortical activity, and the aspects of neural activity changes that were unique to individual tasks and modalities or common across them. Our results ultimately reveal new patterns of training effects, identifying the generalized prefrontal mechanisms that are responsible for rule learning.

INTRODUCTION

Rule learning requires knowledge of the values of available alternatives and their dependence on context (Costa, Tran et al. 2015, Hamid, Frank et al. 2021, Averbeck and O'Doherty 2022). Possible actions may then be assigned a continuously updated place within a hierarchy of values (Tomov, Yagati et al. 2020, O'Doherty, Rutishauser et al. 2021, Xu, Hosokawa et al. 2024). Such hierarchies allow relevant actions to be selected based on whichever is deemed to have the highest values in the current state, with recent research linking these values to a variety of neural correlates in higher cortices (Averbeck and O'Doherty 2022, Moneta, Garvert et al. 2023, Grossman, Man et al. 2026). Rules may be acquired through trial-and-error training, meaning that the prior trial's outcome can be used to infer the best choice in the current trial (Lloyd, Becker et al. 2012). Neural correlates of rule learning have been described in the prefrontal cortex (PFC); the activity of individual neurons changes as associations and contingencies are learned (Asaad, Rainer et al. 1998, Suchow, Fougnie et al. 2014, Jun, Lee et al. 2024). These changes may progress at variable time scales, reflecting rapid events or progressive understanding of a rule through training (Miller and Constantinidis 2024).

Experimental studies, particularly in humans, have produced conflicting results making it difficult to identify common mechanisms of task learning. Training in paradigms that require working memory (WM) represents an illustrative case study. WM training may lead to higher levels of activity across the PFC (Garavan,

Kelley et al. 2000, Hempel, Giesel et al. 2004), suggesting that repeated engagement with training strengthens the neural representations that underlie short-term information maintenance and manipulation. By contrast, other studies have demonstrated decreases in activity which have been interpreted as improvements in efficiency, instead (Schneiders, Opitz et al. 2011, Takeuchi, Taki et al. 2013). There are also studies that have demonstrated increases or decreases in different prefrontal areas (Miller, Tambini et al. 2022), and more subtle changes in neural dynamics rather than overall levels of activity, such as functional connectivity and network modularity, rather than overall changes in activity (Gordon, Breeden et al. 2014, Finc, Bonna et al. 2020). Neurophysiological studies in non-human primate models have not been able to resolve this debate, either, as they typically focus on one, or limited variations of a task, typically in a pair of monkeys, making it difficult to determine whether general principles exist (Sarma, Masse et al. 2016, Mansouri, Freedman et al. 2020, Tang, Riley et al. 2022).

We were therefore motivated to determine the mechanisms of rule learning by examining prefrontal activity in monkeys across learning of a variety of tasks requiring spatial and object working memory. We sought to identify changes in neural activity that parallel the related behavioral improvements and determine if they were unique to the task trained or generalized across separate contexts and modalities.

METHODS

Test Subjects

Data obtained from four rhesus monkeys (*Macaca mulatta*, ages 5-9 years, weighing 5-12kg, three male and one female), were analyzed in this study. None of these animals had any prior experimentation experience at the onset of our study. Monkeys were single housed in communal rooms with sensory interactions with other monkeys. Access to water was restricted during training to motivate task performance. All experimental procedures followed guidelines set by the US Public Health Service Policy on Humane Care and Use of Laboratory Animals and the National Research Council's Guide for the Care and Use of Laboratory Animals as reviewed and approved by the IACUC (Institutional animal care and use committee) of Wake Forest University and/or Vanderbilt University.

Experimental Setup and Task Overview

In all three of our experimental tasks, the monkeys sat with their heads fixed in a primate chair while viewing a monitor positioned 68 cm away from their eyes with dim ambient illumination and were required to fixate on a white square of 0.2 degrees appearing in the center of the screen. All tasks required remembering the location or shape of a cue stimulus. Two tasks presented successive stimulus pairs, then prompted a saccade to one of two targets to indicate if stimuli matched. Two monkeys (Subjects NI and MA) were assigned to the spatial version of this task,

and one monkey (Subject FI) was assigned to the shape version of this task. In a second shape task, the cue was followed by a stimulus pair, and a saccade to whichever matched the cue. Two monkeys (Subject FI and IT) were assigned to this second shape task. During each trial of every task, the monkeys were required to maintain fixation on the square while visual stimuli were presented either at a peripheral location or over the fovea, to receive a juice reward. Any break of fixation immediately terminated the trial, and no reward was given.

Eye position as monitored in the spatial match-nonmatch task via a non-invasive, infra-red eye position scanning system (model RK-716; ISCAN, Burlington, MA). Eye position was sampled at 240 Hz, digitized, and recorded. Similarly, eye position was also monitored in both object tasks via another non-invasive, infra-red eye position scanning system (model ETL-200; ISCAN), sampled at 500 Hz, digitized, and recorded. The system achieved a <0.3° resolution around the center of vision. The visual stimulus display, monitoring of eye position, and synchronization of stimuli with neurophysiological data were performed with in-house software implemented on the MATLAB environment (The MathWorks), using the Psychophysics Toolbox (Meyer and Constantinidis 2005).

Match-Nonmatch Tasks and Pre-Training Paradigm

A total of three monkeys (Subjects NI, MA, and FI) were trained in either a spatial (Subjects NI and MA) or object (Subject FI) version of a match-nonmatch task based on the multi-phase training paradigm of a prior study (Tang, Riley et al. 2022). This

task required the monkeys to maintain fixation, observe two stimuli appearing in sequence separated by delay periods, and to indicate if the two stimuli appeared at the same location or not by making an eye movement to one of two choice targets. Training was broken down into a four-phase discretized training paradigm that were designed to ensure that specific task elements would be acquired in sequence (Fig. S1).

The first phase (Fig. S1b) of training involved training the monkeys to make an eye movement to one of two choice targets to determine that only one of these targets would be rewarded. This phase began with the monkeys being exposed to match trials, requiring an eye movement to the “Diamond” choice target to produce reward. The first stimulus appeared always at the same location (to the right of fixation), followed by a very brief delay period (0.25 s) and a second presentation of the stimulus at the same location. After the second delay period, the two choice targets appeared with the fixation point turning off, either above or below the fixation point, but randomly switching between trials. In the absence of the fixation target, the monkeys quickly foveated one of the choice targets, and they learned through trial and error that the “Diamond” choice target was rewarded. On a subsequent training day, nonmatch trials were introduced. Now the first stimulus appeared again at the right location, but it was followed by a nonmatch stimulus. When the choice targets appeared at the end of the trial, it was the “H” shape that was rewarded. The monkeys quickly reversed and saccaded to the “H” choice target. Phase I of training involved delivering match and nonmatch trials in blocks

with decreasing numbers of trials before alternating. Importantly, at Phase 1 of this training paradigm, the monkey could simply saccade to the choice targets, determine which one was rewarded during the block, and then repeatedly select it in following trials to continue receiving reward for the rest of the session.

In the second phase, the monkey was presented with alternating blocks of match and nonmatch trials, of decreasing block length, until they were randomly mixed, requiring the monkey to gradually learn to associate the second stimulus location with the corresponding choice target (Fig. S1c). This meant associating the concept of "match" with the "Diamond" and "nonmatch" with the "H" shape. Examining the rule learning that occurred over the course of this second phase formed the foundation of our current experiment (Fig. 1a). Notably, at this point in training, the cue stimulus always appeared at the same location. However, the previous multi-phase training paradigm also included a third phase wherein the monkeys had to generalize the task to new stimulus locations, appearing at a 3 × 3 grid (Fig. S1d), as well as a fourth phase, wherein the delay period was increased to place a greater demand on WM (Fig. S1e).

Our current training paradigm thus utilized the same spatial match-nonmatch task from the second phase of this previously described four-phase training paradigm, as well as a parallel object version of this match-nonmatch task that was created specifically for the needs of our current experiment. In either case, our monkeys were given a window of 2 seconds to begin fixation, or the trial would be automatically aborted. Following this initial acquisition of the fixation point, the

trial would begin with a fixation epoch in which no stimuli would be shown (1.0 s). This was followed by the presentation of a single stimulus that was always the same and was referred to as the cue (0.5 s). Only a single match/nonmatch stimulus combination was applied in either version of the task, and so the initial cue stimulus was the same in every trial, always a square that appeared to the right of fixation in the spatial version of the task, and always a circle that appeared over the fixation in the object version of the task. This was followed by a very brief delay epoch (0.25 s) and then a second stimulus presentation, known as the sample, which could vary in either location (appearing to the left or right of fixation in the spatial task), or shape (appearing as a triangle or circle in the object task). After the second delay epoch (0.25 s), the two choice targets (Diamond and H-shape for the spatial task, Green and Blue for the object task) appeared with the fixation point turning off, either above or below the fixation point, but randomly switching between trials (0.5 s) to prevent confounds. Only one of these two choice targets would produce reward in each trial, and this was determined by the cue and sample stimuli that were previously presented. If these stimuli matched each other, a saccade to the "Diamond" and "Green" choice targets would be rewarded. Likewise, if these stimuli did not match, a saccade to the "H-shape" and "Blue" targets would be rewarded instead (Fig. 1a).

At the start of our experiment, all monkeys were already able to maintain fixation and had been exposed to the visual stimuli that would eventually be incorporated into their respective tasks. Moreover, monkeys had been taught a pre-

training version of the spatial and object match-nonmatch task, wherein sessions would present trials that used only one of the two possible stimulus combinations, without any trials that used the other stimulus combination. The presented stimulus combination that was selected for any single training session would alternate between the two options over the course of multiple sessions but would not mix. This meant that the single rewarded choice target remained constant throughout the span of the entire training session and thus taught the monkeys to make an eye movement to one of two choice targets to determine that only one of them would be rewarded. Then, once the monkey learned which choice target would be rewarded for the session, they could simply repeatedly select that target for the rest of the trials in that session to continue receiving reward. This discretized pre-training paradigm ensured that different task elements would be acquired in sequence. Specifically, monkeys had no need to pay attention to which of the two possible cue and sample stimulus combinations were being used in each trial and were thus not yet aware of the difference between trials with matching or nonmatching stimuli.

Our current training paradigm was designed to introduce a new task element by progressively mixed the two possible stimulus combinations into the same training sessions. This was carried out by requiring for the monkeys to periodically switch between trials with matching and nonmatching stimuli after a set block of correct trials, the size of which was constant and manually predetermined at the start of each session. As a result, the rewarded choice target would change with

each trial block, and the monkeys had to learn to identify the difference between trials with matching or nonmatching stimuli, since the presented stimulus combination would be responsible for determining which choice target would be rewarded at the end of each trial. The constant size of these blocks that was predetermined at the start of each session ensured that trials with matching or nonmatching stimuli would be balanced. During the first training sessions when the monkey had not yet begun to learn the new rules of this experiment, they were expected to perseverate in their behavior, and as a result, they would almost always fail the trial in which there was a reversal between matching and nonmatching stimuli. This led to a consistent Drop In Performance (henceforth abbreviated as DIP) at the reversal trial, which was calculated by subtracting the mean performance level of the trial that directly preceded the reversal trial from the mean performance level of the reversal trial itself, examined separately for each session. As a result, the shorter these trial blocks became before a reversal between the two possible stimulus combinations, the more the monkeys would be forced to adapt their behavior to avoid failure.

Object Choose-Match Task

Two monkeys (Subjects IT and FI) were trained in an additional object working memory task, the object choose-match task as well. As with the spatial and object match-nonmatch tasks, each trial began with a fixation epoch in which no stimuli were shown (1.0 s). This was followed by a single cue stimulus (either a circle or a

triangle) that was displayed at the center of the screen (0.5 s), and then a delay epoch (0.25 s), after which a pair of target stimuli (both a circle and a triangle) were displayed on diametrically opposite sides of the screen (0.5 s), prompting the monkeys to indicate which of these matched the first presented stimulus via an eye movement. The configuration of these target stimuli would randomly switch between trials, with the circle appearing above the fixation point and the triangle appearing below, or the reverse, to prevent confounds. Trial blocks alternated which object was used as the initial cue stimulus, and the constant size of these blocks that was predetermined at the start of each session ensured that trials with each type of cue stimuli would be balanced, thus allowing the learning of the same rule from the spatial and object match-nonmatch tasks—that when trials changed between presenting different stimuli, the rewarded choice target would be change as well—to be examined in a different context as well (Fig 1a-right).

Behavioral Analysis

For all three tasks, performance was calculated for each session by dividing the number of successful trials by the total completed trials, excluding any trials in which fixation was broken. Each monkey was assessed individually.

Behavioral Strategy Analysis

To quantify the behavioral strategies employed by each monkey across training sessions, we modeled each trial as arising from one of five candidate strategies and

estimated the relative contribution of each strategy on a per-session basis. Sessions with fewer than 100 completed trials were excluded from analysis.

Five candidate strategies were defined, each making a deterministic or probabilistic prediction about the probability of choosing the upward saccade target on each trial: 1. Random choice: The monkey saccades upward or downward with equal probability ($p = 0.5$) on every trial, regardless of trial history, stimulus information, or the configuration of the choice targets. 2. Preferred location: The monkey consistently saccades to the same spatial location on every trial, regardless of which choice target appears at this location. 3. Preferred choice target: The monkey consistently chooses to saccade to the same choice target, regardless of whether this target is correct. 4. Win-stay-lose-shift: The monkey repeats the saccade choice that was rewarded on the previous trial (win-stay) and only switches to saccade to the alternative choice target after an unrewarded trial (lose-shift). On the first trial of a session, when no prior outcome was available, this strategy assigned $p(up) = 0.5$. 5. Match-nonmatch rule: The monkey uses the correct task rule, choosing the target stimulus that is associated with the presented stimuli of the current trial. In the case of the spatial and object MNM tasks, this meant a saccade to the “Diamond” or “Green Square” target to indicate a match or the “H” target or “Blue Square” to indicate a non-match. In this case of the object CM task, this meant a saccade to whichever choice target matched the presented cue stimulus. Additional strategies beyond these five candidate strategies may

have been possible in theory, but these would have largely required greater effort and complexity than rule learning itself.

On each trial i, each strategy s generated a prediction for the probability of choosing the upward saccade, denoted $p_s(\text{up} \mid \text{trial i})$. If the monkey chose upward on that trial, the likelihood assigned to that strategy was $p_s(\text{up})$; if the monkey chose downward, it was $1 - p_s(\text{up})$. These raw likelihoods were normalized across strategies on each trial to sum to 1, yielding a per-trial weight for each strategy:

$$w_s(i) = L_s(i) / \Sigma_t L_t(i)$$

where $L_s(i)$ is the likelihood of the observed choice under strategy s on trial i, and the sum in the denominator runs over all five strategies. The session-level weight index for each strategy was then computed as the mean of these per-trial weights across the first 80% of trials in the session:

$$W_s = (1/N) \Sigma_i w_s(i)$$

where N is the number of trials used. This normalization ensures that the five weight indices sum to 1 for each session, allowing these to be interpreted as the relative probability that each strategy best accounts for the monkey's behavior during that session. Sessions in which the animal was excluded due to insufficient trial counts were assigned weight indices of zero and excluded from all subsequent statistical comparisons.

Implant Surgery

All subjects were initially acclimated with the laboratory and trained to maintain fixation on a white dot while visual stimuli appeared on the screen. After this initial stage of training was complete, monkeys were implanted with an MRI-compatible headpost device. Another surgery was used to implant a chronic array of electrodes in their lateral PFCs (Fig. S3a). All cases of surgery were carried out under inhalant anesthesia. Opioid analgesics were administered after the surgery, and the animals were allowed to recover for at least 3 weeks before behavioral sessions began anew.

For the two monkeys assigned to the spatial task (Subjects NI and MA), the chronic implant was designed in-house and encompassed 64 parylene-c insulated, Iridium electrodes that tapered from a 40 mm diameter to an exposed electrode tip that ranged from 5–7 mm long (Spingath, Kang et al. 2011). The implant comprised an 8 × 8 grid of electrodes, with adjacent electrodes spaced 0.75 mm apart from each other, thus covering an area of 5.25 mm × 5.25 mm. This electrode array targeted the dorsolateral PFC, with electrode tracks descending in both banks of the principal sulcus.

For the two monkeys assigned to the object tasks (Subjects FI and IT), the chronic implant was a chronic 128 Grey Matter Research electrode array. This implant comprised an (incomplete) 12x12 grid of tungsten electrodes with a 250 μm diameter, spaced 1.5 mm apart from each other. The electrode array targeted the dorsolateral and ventrolateral PFC, centered across the dorso-ventral and anterior-posterior axis. For the purposes of our analysis, the anterior-posterior axis was

defined as the line connecting the genu of the arcuate sulcus to the frontal pole. The recording coordinates of each neuron was projected onto this line, with position expressed as a proportion of the line's length.

Across all monkeys, each electrode used for recording had a 1 MΩ impedance at 1 kHz. The position of each electrode array was determined based on MRI and then verified during the stereotaxic implantation surgery. Electrodes from both implants could be advanced into the cortex independently of each other and electrode depths were adjusted as needed to optimize placements, up to 5 mm (in the banks of sulci).

Recording Neural Activity

Once electrode positioning was finalized and the monkeys recovered from surgery, task training and neuro-physiological recordings from the array commenced. Neuronal data from each electrode were recorded over the course of training using an unbiased spike selection procedure, with no regard to the response properties of the isolated neurons. For all monkeys, the signal from each electrode was amplified and sampled continuously at 30kHz and stored for off-line analysis. Data was band-pass filtered between 500 Hz and 8 kHz for spike identification.

The two monkeys assigned to the spatial task (Subjects N and M), used a Cerebus system (Blackrock Microsystems, Salt Lake City, UT), with the threshold for spike acquisition was set at 3.5 × RMS of the baseline signal, for each electrode, each day. The two monkeys assigned to the object tasks (Subjects F and I), by

contrast, used a modular data acquisition system (OpenEphys system, FHC, Bowdoin, ME). Semi-automated cluster analysis was carried out through Kilosort, applying principal component analysis of the waveforms to sort recorded spike waveforms into separate units. Neural data was recorded for every training session of every monkey over full span of each of these daily sessions.

To ensure a stable firing rate in the analyzed recordings, we identified the range of trials with stationary firing rate in each neuron. This was necessary because of neurons disappearing or appearing during a run, as we were collecting data from multiple electrodes. We began by examining the mean Fano factor of each cell via the methods of Churchland and colleagues, using the algorithm developed by these authors and made available online at: http://www.stanford.edu/shenoy/GroupCodePacks.htm (Churchland, Yu et al. 2010, Qi and Constantinidis 2012). Data for each neuron was initially treated separately. Spike counts were computed in a 100-ms sliding window moving in 20-ms steps. The method computed the variance and mean of the spike count across trials in each time bin and performed a regression of the variance to the mean. This slope of this regression represents the Fano Factor reported at the given time bin. The trial activity of any neurons that displayed high-variability (mean Fano factor > 2) was then fit to a two-component mixture model (Poisson + Gaussian), using a Bayesian information criterion (BIC) to determine if the distribution is genuinely multimodal, calculating the separation between modes (i.e. peaks in distribution), and then finally, removing all trials from the smaller modes as outliers. Each high-

variability neuron was examined individually to identify and remove these outlier trials. This method was used to remove trials from 0.7% of cells from the spatial task, 16% of cells from the object match-nonmatch task, and 3.6% of cells from the object choose-match task.

Firing Rate Analysis

Analyses of neural firing rate were implemented with the MATLAB computational environment (Mathworks 2019, Natick, MA). Firing rates were assessed separately for each epoch across all cells. This was also repeated for each epoch after subtracting the mean baseline fixation firing rate, thus isolating any activity changes that were caused by the task itself, rather than spontaneous noise (He 2013). Only correct trials were applied in these analyses to reduce possible confounds and to focus on when the monkey's behavior would be most reflective of the rule learning that would be needed for optimized performance.

To begin addressing our question of how parallel activity-based mechanisms may be applied across spatial and object rule learning, we applied a series of generalized additive mixed models (GAMMs). GAMMs are a flexible, semiparametric method for identifying and estimating nonlinear effects of covariates on the outcome variable when observations—in this case referring to the empirically observed data from each session—are not independent. We therefore predicted that this would reveal possible relationships between training and prefrontal activity. To this end, we examined how the single predictor variable of DIP

magnitude may affect the outcome variable of firing rate over the course of training based upon a prior study from our lab that used the following equation (Zhu, Garin et al. 2024):

$$Y_{ij} = \beta_0 + f(DIP_{ij}) + b_i + \varepsilon_{ij}$$

where Y_{ij} represents the mean epoch firing rate for observation (session) j and monkey i, β_0 represents the overall intercept from all subjects. Random subject-specific effects are represented as b_i for monkey i. Random effects included the subject-specific intercepts and slopes for DIP. The outcome of firing rate—which was continuous—was examined via a Gaussian family and identity link function. Statistical significance was then identified from smooth terms via the mgcv package that expanded the smooth into basis coefficients, and for those models in which the smooth term revealed a statistically significant fixed effect, the gratia package was used to conduct exploratory post-hoc analyses to identify significant training effects by approximating the derivatives of each estimated smooth function of DIP via the method of finite differences. A simultaneous 95% confidence, which were calculated by randomizing the labels of different monkeys at each x-axis point (representing mean session DIP) and using that to refit the GAMMs over 100 repetitions. The error term of the given observation j and monkey I is represented as ε_{ij}. Finally, the relationships between each predictor variable and the response variable are represented by the smooth function, f, which is a spline with penalty. Smoothing of the curve has been enforced using up to a maximum of five basis

functions for the sake of preventing overfitting. This smooth function f can be further explained as the following:

$$f(\mathrm{DIP_{ij}}) = \sum_{K=1}^{K} \beta_K B_K\,(\mathrm{DIP_{ij}})$$

where k represents the number of spline basis functions (with a maximum of 5), β_k represents the coefficients estimated by the GAMM, and B_k represents the spline basis function.

Just like our previous firing rate analyses, the firing rates variables that were used in our GAMM analyses were assessed separately for each epoch across all cells, and also across all cells with the mean baseline fixation epoch firing rate was subtracted from the other epochs. Moreover, we once again relied exclusively on correct trials to reduce possible confounds and to focus on when the monkey's behavior would be most reflective of the rule learning that would be needed for optimized performance.

GAMM analyses were implemented using the mgcv package for R with a smooth function of DIP across sessions as a covariate, using a thin plate regression spline basis to estimate this smooth function. Effective degrees of freedom (edf) were estimated from the data using the restricted maximum likelihood approach. We therefore modeled nonlinear effects using a flexible, automatically adaptive curve with theoretically optimal smoothness control, with each datapoint used to construct our splines automatically in lieu of manual knot selection.

Unexplained Variance in Firing Rate

To confront the question of how training may affect the computation efficiency of encoding task relevant stimuli—especially during the reversal between trials with different stimuli—we analyzed changes in firing rate variance that could not be explained by the task factors of the presented stimulus, the direction of the rewarded choice target, and whether the trial was answered correctly or not, examined across trials based on their position relative to the reversal itself. This analysis thus relied upon both correct and error trials, as both were necessary to assess the given task factors, and was evaluated through a linear regression model with the following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Here, y represents the dependent variable of firing rate, modeled as a function of β0, which represents the baseline firing rate as an intercept term, and x1, x2, and x3, which represent the predictors (independent variables of the three examined task factors). These three factors were the presented stimulus, the direction of the rewarded choice target, and whether the trial was answered correctly or not, examined across trials based on their position relative to the reversal itself. Next, β1, β2, and β3 represent the regression coefficients of these task factors and quantify how much each predictor contributes to explaining the firing rate as slopes that show how much the firing rate changes per unit changes with each predictor. Finally, ϵ represents the error term (residuals), which define the unexplained variance as the difference between the observed firing rate and the

model's prediction. To fit the model, spike count data were generated using a moving window of 400ms wide and 100ms steps, from 1.5s before cue onset to 2.5s after in the case of the spatial and object match-nonmatch tasks, or 1.5s before cue onset to 1.5s after in the case of the object choose-match task. The linear model was then fitted to each time bin individually, using all possible trials from each distance relative to the reversal trial. The percentage of unexplained variance ratio (SSE/SST) was then calculated for each relative distance to the reversal, and max-normalized across all distances (i.e. the trial before, during, and after the reversal). Lastly, the results across all time points were averaged for each relative distance to the reversal in each cell. This unexplained variance was then examined across relative distances to the reversal and across early/late training through a series of two-way ANOVAs. Only cells with at least three trials at each of the examined distances relative to the reversal were included in this analysis.

Decoding of Task Variables

To understand how training affects the types of information represented in neural populations, a new set of decoder analyses were added using the supervised max-margin model of a support vector machine (SVM) on the MATLAB environment with a linear kernel function, based upon what has been used in our prior studies of the PFC (Dang, Jaffe et al. 2021). We predicted—just as training was observed to shift firing rate in different directions for different tasks/monkeys—training may shift the separation of task factor conditions in the resulting max-margin boundary (i.e. the

hyperplane) in different directions for different tasks/monkeys as well. This would be expressed through different changes in decoder performance for different tasks/monkeys, ultimately providing new insights on the implications of such diverse effects that could not be inferred from firing rate alone. This led us to apply the fitcecoc function to produce a multiclass, error-correcting output codes (ECOC) model using the predictors that were drawn from our recorded neuronal data.

To obtain the predictor data of a single trial, we created a 3-D matrix that had the dimensions of cell number, by the status of the examined task factor, by the total number of 400 ms time bins. The total number of time bins was decided by using a 400 ms sliding window, with a 100 ms step size (with overlap) spanning the full length of the trial (3s for both of the spatial and object match-nonmatch tasks, or 2.25s for the object choose-match task). This resulted in a total of 27 time bins for the spatial and object match-nonmatch tasks and 19 time bins for the object choose-match task. Each entry of the resulting 3-D matrix consisted of the spike count of the individual neuron in the selected time bin and trial's condition of the examined task variable. This was then used to create a pseudo-population that would provide training data, thus allowing us to test the resulting classifier on the factors of the presented stimulus, the direction of the rewarded choice target (i.e. whether the rewarded target appeared up or down), the configuration of the choice targets (i.e. whether a specific target appeared up or down regardless of whether it was rewarded), the combination of the presented stimulus with the direction of the rewarded choice target, the monkey's choice of saccade direction, and the

monkey's choice of saccade to a match, via the class label input, that is, a scalar variable indicating the task variable that the decoder was designed to examine. This analysis relied upon both correct and error trials, examining the same number of cells at each stage of training.

Ten-fold cross validation then split our training data into 10 sets, 9 of which were used for training, while only the last tenth of the data was applied for testing, thus producing a cross-validated model that could be used to make a prediction of the data at each time point. Training and testing were first carried out using correct trials only, and then again, separately, using correct trials combined with error trials, to result in two cross-validated models that would be examined independently. We reasoned that the encoding of information does not necessarily imply the utilization of such information, and so the inclusion of error trials in this latter cross-validated model would allow us to dissociate behavioral and sensory encoding, thereby revealing possible changes in the encoding of different types of task data that would not be evident from the use of correct trials alone.

This process was then repeated 10 times, thus allowing all data to be used as both training and test data. The average accuracy of these 10 repetitions was counted as the decoding accuracy for a single iteration, and this entire process was then repeated across 10 iterations, with 95% confidence intervals calculated via linear interpolation on these 10 iterations.

Finally, to determine whether the change in decoder performance over the course training was consistently within chance level, the average decoding

accuracy across all iterations was compared across early vs. late training at each timepoint through a nonparametric randomization test—also known as a permutation test—wherein all trials from both the early and late datasets were pooled together, and then randomly shuffled into two surrogate datasets of equal size to the early and late datasets. These pairs of surrogate datasets were used to produce two cross-validated models in the same manner as the original decoders, with the difference in the performance of these surrogate datasets at each timepoint being used to compute a null distribution over the course of 1000 permutations. The values of this null distribution therefore represented the difference between early/late decoders that could be expected from chance. We then compared the empirical difference in performance between the early/late decoders to the corresponding time points of this null distribution. Any time points at which the empirical difference surpassed the top 2.5% of values from the corresponding time points of the null distribution were marked as statistically significant ($p < 0.05$, two-tailed).

Targeted Dimensionality Reduction (TDR)

To characterize how task-relevant variables—specifically matching status (cue shape in the choose-match task) and the saccade direction of the rewarded target—are represented in the neural population, we employed targeted dimensionality reduction (TDR). This method identifies low-dimensional axes within the neural state space that are maximally sensitive to specific task variables while

remaining orthogonal to one another. The neural response matrix R (units × time bins × trials) was first organized by experimental conditions. We defined conditions based on the unique combinations of the two examined task variables: matching status (cue shape in the choose-match task) and choice saccade direction. For each unit, the trial-averaged response $X_c(t)$ was calculated for each condition c. These condition-averaged responses were then z-scored across all time points and conditions to ensure units with high firing rates did not disproportionately dominate the population analysis.

To focus our analysis on the most prominent features of the population response and reduce trial-to-trial noise, we performed Principal Component Analysis (PCA) on the condition-averaged data. We constructed a data matrix by concatenating the z-scored responses across all conditions and time points. The first Np principal components (capped at 12 or the numerical rank of the data) were retained to define a "denoising" subspace. We defined a denoising operator D as: D = V * V', where V is the matrix containing the first Np eigenvectors (loadings). This operator was used to project subsequent regression vectors back into the primary task-related subspace.

To identify the contribution of each task variable to the neural activity, we performed a trial-by-trial linear regression for each unit i and each time bin t: $r_i(t, k) = \beta_{match}(i, t) * var1(k) + \beta_{sac}(i, t) * var2(k) + \beta_0(i, t) + \varepsilon$, where $r_i(t, k)$ is the z-scored response of unit i at time t for trial k, and var1, var2 represent the matching status (cue shape in the choose-match task) and saccade direction variables,

respectively. This yielded time-varying regression vectors, $\beta_{match}(t)$ and $\beta_{sac}(t)$, which point in the direction of the state space that captures the variance associated with each variable.

For both variables regression vectors, we first denoised them by projecting them into the denoise PCA subspace: $\beta_{denoised}(t) = D * \beta(t)$. Then for each variable v, we identified a single time-independent axis by selecting the vector $\beta_{denoised}(t)$ at the time point t where its magnitude (norm) was maximal. To ensure that the resulting task axes were independent, we applied a QR decomposition to the set of regression vectors: [ω_match, ω_sac] = QR([β_match(t), β_sac(t)]). This resulted in an orthonormal basis where each axis represents a specific task variable, purified of contributions from the other. Finally, the condition-averaged population activity was projected onto these orthonormal axes.

Neural data was visualized in the space defined by two task variable related principal components defined as described above, using a 400 ms wide moving window average with a 100 ms step size. Separability was defined as the Euclidean distance between trajectories across both axes of a two-dimensional subspace capturing the variance due to the combination of matching status (cue shape in the choose-match task) and the saccade direction of the rewarded choice target (i.e. up or down) task conditions. To quantify the separability of neural trajectories along each task-relevant axis, we computed a normalized Euclidean distance metric for each dimension separately (matching status and saccade direction). We identified four trial conditions defined by the crossing of matching status (match/non-match)

and saccade direction (up/down): match-up (c_1), match-down (c_2), non-match-up (c_3), and non-match-down (c_4). Match and non-match centroids were then computed by averaging across saccade directions within each matching category:

$$\bar{c}_{match} = (c_1 + c_2) / 2,\ \bar{c}_{nonmatch} = (c_3 + c_4) / 2$$

The separability index was then defined as:

$$S_{match} = 2 \times d(\bar{c}_{match}, \bar{c}_{nonmatch}) \ / \ [d(c_1, c_2) + d(c_3, c_4)]$$

where $d(\cdot,\cdot)$ denotes Euclidean distance in the 2D PCA subspace. This normalization by the mean within-group spread ensures the index reflects between-group separation relative to within-group variability, yielding a value greater than 1 when between-group distance exceeds within-group spread. Early and late training trials were resampled over the course of 100 iterations (with replacement), thus providing 100 early and late training trials for each of the four trial conditions. For each resampling iteration, the neural state for each of the matching status conditions was estimated by averaging the projected coordinates across an interval of 700-1000ms after cue onset which corresponded to the sample presentation period. Separability across the matching status conditions was thus examined through this interval. Likewise for the neural state for each of the saccade direction of the rewarded target condition, was estimated by averaging the projected coordinates across an interval of 1500-1900ms after cue onset, coinciding with the saccade, meaning that separability across these target direction conditions was examined through this interval instead. Finally, for the neural state for each of the match direction conditions for the CM task was estimated by averaging the projected

coordinates across time points corresponding to an interval of 1000-1400ms after cue onset, meaning that separability across match conditions was examined through this interval instead.

The empirical change in separability between conditions was then assessed across early vs late training by pooling all trials from both the early and late datasets, and then randomly shuffling to create two surrogate datasets of equal size to the original early and late datasets. These surrogate datasets were used to plot firing rate trajectories in state space corresponding defined by the same two principal components, and trials were resampled over the course of 100 iterations (with replacement) to plot firing rate trajectories in state space corresponding defined by the same two principal components, representing the change in separability between conditions that could be expected from chance. Finally, we applied a Wilcoxon rank sum test to compare the vector of all 100 empirical differences between the observed early and late neural trajectories against the vector of all 100 shuffled differences from the null distribution. We calculated a test statistic based on how many times the empirical values exceed the null distribution values. Under the null hypothesis (no difference between groups), empirical differences would be randomly distributed, and so this allowed us to determine whether the observed empirical difference between early and late trajectory distances could not plausibly arise from random sampling.

RESULTS

Monkeys acquire sequential task elements via discretized training

Behavioral and neural data were collected from 4 monkeys (three male, one female) as they trained in one spatial and two object working memory tasks. In the spatial match-nonmatch task (Fig. 1a), the monkeys were required to maintain fixation, observe two stimuli appearing in sequence separated by delay periods, and to indicate if the two stimuli appeared at the same location or not by selecting one of two choice targets, defined by their shape ("H" or "Diamond"). The object match-nonmatch task (Fig. 1a) followed the same structure but prompted the monkey to indicate if two stimuli appearing in sequence had the same shape or not, by selecting one of two choice targets, defined by their color ("Green square" or "Blue square"). Finally, in the object choose-match task (Fig. 1a), the monkeys were required to maintain fixation and observe a single cue stimulus (either circle or triangle) that was followed by a delay period. Then a pair of choice targets (both circle and triangle) would appear, prompting the monkey to saccade to whichever matched the cue.

Task paradigms discretized training via alternating trial blocks with different stimulus combinations. Each session of both match-nonmatch tasks was thus assigned a set number of trials that would have to be completed successfully before a reversal between trials that presented either the matching or nonmatching stimulus combination. In the case of the object choose-match task that only

presented a single cue stimulus, the completion of a trial block led to a reversal between which object was used as that initial cue stimulus instead. The size of these trial blocks—also known as the “reversal rate”—was constant and manually predetermined at the start of each session. Alternating blocks of match and nonmatch trials were presented in both versions of the spatial and object match-nonmatch task, with block length decreasing over the course of multiple sessions until they were randomly mixed, thus requiring the monkey to associate the second stimulus location/shape with the corresponding choice target to avoid failure. Alternating blocks of circle and triangle cue presenting trials were applied in each session, with block length decreasing over the course of multiple sessions until they were randomly intermixed, thus requiring the monkey to associate the cue stimulus shape with the corresponding choice target to avoid failure.

Monkeys were able to perform all three tasks with an average accuracy across sessions that was above chance level (Fig. 1b and Fig. S2a, mean performance of spatial match-nonmatch task: 70.30%, mean performance of object match-nonmatch task: 67.96%, mean performance of object choose-match task: 66.50%) even as we progressively reduced the reversal rate over the course of training. During the first training sessions when the monkey had not yet begun to learn the new rules of this experiment, they were expected to perseverate and would thus almost always fail the trial in which there was a reversal between matching and nonmatching stimuli. This led to a consistent Drop In Performance (henceforth abbreviated as DIP) at the reversal trial (Fig. 1d and Fig. S2c). This

forced the monkeys to realize that a new element had been added to their task: the rewarded target could now change over the course of a session. The magnitude of DIP was calculated for each session by subtracting from the mean success rate of the trials where the reversal occurred the mean success rate of trials that occurred directly prior to this reversal.

Monkeys were given no indication for when a reversal occurred, and so DIP—as well as the brief period of behavioral adjustment during the trials that immediately followed from this point—proved to be largely responsible for the error trials that occurred in the course of training. Over the course of additional training sessions, DIP would only decrease in absolute magnitude if the monkeys began to learn that when trials changed from presenting pairs of matching stimuli to presenting pairs of nonmatching stimuli (or the reverse), the rewarded choice target would be change as well, switching from the Diamond and Green targets to the H-shape and Blue targets (or the reverse). Learning this rule allowed the monkeys to adjust their behavior accordingly to avoid failure. As a result, the magnitude of DIP could be used to identify the stage of learning. We relied on DIP to classify sessions into “early” and “late” based on the longest path of monotonic decrease in the running average (with a sliding window of 5 sessions) that crosses the halfway point of sessions (Fig. 1e and Fig S2d).

During training, the reversal rate was gradually reduced across multiple sessions, shortening trial blocks over the course of training until the presentation of match or nonmatch stimulus (i.e. trials presenting stimulus pairs with the same

location/shape. in the spatial/object match-nonmatch tasks, or trials presenting the circle and triangle stimulus in the object choose-match task), were randomly mixed. A complete lack of understanding of the rule would result in failure following a reversal between the presentation of match or nonmatch stimulus, thus causing a consistent drop in performance (DIP). In contrast, full understanding of the rule would result in no DIP when the rewarded choice target reversed. We thus tested for evidence of rule learning as training progressed by determining whether the absolute DIP magnitude decreased as a function of training. This was indeed the case. An example is shown in Fig. 1e (linear regression; $F(1,101) = 135.9$, $p=2.07E-20$)), The rest of the cases are shown in Fig S2d ($p<E-4$ in all cases).

Our training paradigm was designed to incentivize genuine rule learning for success and so our monkeys were forced to use trial-and-error to realize that the stimuli presented prior to the choice targets would determine which of these targets would be rewarded. In behavioral terms, this meant that the optimal strategy—that is, the only strategy that could potentially achieve a consistent 100% success rate—would be to examine the presented stimuli, and then saccade to the associated choice target. However, at least during the start of training, a variety of alternative task strategies could also be used to produce reward, even in the absence of rule learning. For example, longer trial blocks preceding a reversal between the presentation of match or nonmatch stimulus (i.e. trials presenting stimulus pairs with the same location/shape. in the spatial/object match-nonmatch tasks, or trials presenting the circle and triangle stimulus in the object choose-match task) would

allow the monkey to continue receiving reward by saccading to the same reward target for a longer span of time, without needing to learn the difference between trials with different stimulus presentations. Once these trial blocks came to an end, the monkey could then recognize the sudden halt in reward as a signal to shift their saccade to the other choice target instead, thus allowing them to continue receiving reward without genuine rule learning. This was known as the "Win-stay-lose-shift" strategy.

Other strategies that could serve as possible alternatives to genuine rule learning included choosing the target at random, choosing to saccade towards the same location every trial regardless of which target was at that location, or choosing the same target stimulus every trial regardless of the target's location—just as the monkeys had learned to do during the pre-training paradigm.

We assessed the weight index (see Methods) for each alternative strategy across each session by comparing the percentage of trials in which the monkey's choice was the same as the choice that would be dictated under each of the given strategies to determine the probability of each strategy being used, relative to the others (Fig. 1f and Fig. S2e). Raw likelihoods were therefore normalized across strategies on each trial to sum to 1, yielding a per-trial weight for each of the individual strategies. The session-level weight index for each strategy was then computed as the mean of these per-trial weights across the first 80% of trials in the session, to account for the reduced effort that would be expected from the monkey as they approaching satiation for juice reward towards the end of the session. As

expected, the weight index of the optimal rule learning-based strategy increased significantly from early to late training to far surpass all other strategies by the end of training (two-way ANOVA, $F(4,515)=85.95$ for interaction of early/late training and strategy type; $p=6.8E-56$ in Fig. 1f; $p<E-4$ for all other cases shown in Fig. S2e).

Firing rate changes during training

Extracellular neurophysiological recordings were recorded from the lateral PFC of the four monkeys with chronic arrays (Fig. S3A-D). A total of 270 single units were recorded over the course of 159 sessions from two monkeys assigned to the spatial match-nonmatch task, 1348 single units were recorded over the course of 203 sessions from one monkey assigned to the object match-nonmatch task, and 1765 single units were recorded over the course of 127 sessions from two monkeys assigned to the object choose-match task (a total of 3383 single units and 489 sessions across all tasks).

We sought to examine whether rule learning could be characterized by changes in firing rate across all tasks. Previous neurophysiological studies have suggested an overall increase in mean firing rate as learning progresses (Tang, Riley et al. 2022). However, such studies have also revealed decreases in firing rate during acquisition of some task elements (Tang, Qi et al. 2019, Tang, Riley et al. 2022). For this reason, we quantified prefrontal firing rates in each epoch of each task to assess how activity changed as training progressed. Mean firing rates in correct trials across all neurons were plotted (Fig. S3E). We then examined the

possibility that a systematic relationship in firing rate changes between tasks might be non-linear and might progress as a function of training. We therefore relied on a Generalized Additive Mixed Models approach, using DIP as the dependent variable in this analysis (Fig. 2). This analysis serves as a flexible, semiparametric method for identifying and estimating nonlinear effects of covariates on the outcome variable (i.e. firing rate), while accounting for the possibility that differing firing rate intercepts from different monkeys might obscure the relationship that we were seeking to examine. We first examined the baseline fixation epoch firing rate across all available cells, revealing a slight trend of increase (GAMM, baseline fixation: $F = 0.078$, $p = 0.045$). Repeating this analysis in other task epochs, after subtracting the baseline firing rate, revealed a highly significant nonlinear trend of decrease in firing rate of the choice epoch (GAMM: $p = 0.001$, $F = 3.62$). Firing rate changes in other epochs were not significant (GAMM, cue: $p = 0.776$, first delay: $p = 0.696$).

These results suggest that the overall increase in firing rate after training reported in previous studies is not a general phenomenon across different tasks, with only a minor increase in baseline firing rate captured by the GAMM across our experiments. Instead, a decrease in firing rate during the choice epoch period appeared to have a more consistent effect.

In addition to examining changes in overall firing rate, we sought to examine the hypothesis that rule learning would be characterized by changes in the variance of firing rate explained by task variables, as learning takes place. Monkeys trained in the two-armed bandit reversal learning task would display increased unexplained

variance in firing rate during the change between conditions in a reversal learning task (Bartolo and Averbeck 2020). This means that when firing rate was modeled with known task variables, the amount of variance that could not be explained by this model would significantly increase during the trials when the monkeys made an internal decision to reverse their target preference. We therefore sought to determine whether this increased unexplained variance would occur over the process of rule learning as well, or alternatively, if this might be a phenomenon that would only emerge after the completion of training.

We therefore used a linear regression to model firing rate as a combination of three task factors a) the presentation of match or nonmatch stimulus (i.e. trials presenting stimulus pairs with the same location/shape in the spatial/object match-nonmatch tasks, or trials presenting the circle and triangle stimulus in the object choose-match task); b) the direction of the rewarded choice target (i.e. whether the rewarded target appeared up or down), and c) whether the trial was answered correctly or not (Fig. 3; see also the insets of Fig. 4 for an illustration of the match/nonmatch, and target direction conditions). Firing rate was thus expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

where y is firing rate calculated in a 400 ms bin, and x_1, x_2, and x_3 are categorical variables representing the predictors of matching or nonmatching status, up or down direction of the correct target, and correct or error outcome of the trial. The amount of firing rate variance that this model failed to explain across all trials in a

session was calculated as $1-R^2$. Such variance would be implied to come from firing rate variability that is not linearly related to the three variables of our model. Values computed in successive 400 ms bins tiling the choice epoch (and the entire trial) were then averaged to calculate the values depicted in Fig. 3 (and Fig. S4).

Unexplained variance was examined across a range of 3 trials around the reversal (i.e. the trial directly before, during, and after the reversal), and so cells required a minimum of three trials at each position to be included in this analysis. We then used a two-way ANOVA to compare the response vector of this unexplained variance across the grouping vectors (factors) of training level (i.e. early vs late), and trial distance relative to the reversal. We focused particularly on the choice epoch, after all the stimulus information had been revealed and the subject received feedback on its choice, in the form of reward, or absence thereof (Fig. 3). This analysis revealed a significant main effect for training stage where unexplained variance was observed to increase from early to late training (spatial match-nonmatch task: $F(1, 600) = 37.29$, $p<E-4$; object match-nonmatch: $F(1, 1365) = 14.6$, $p = 0.0001$; object choose-match: $F(1, 2937) = 71.3$, $p<E-4$). Likewise, repeating the same two-way ANOVA for unexplained variance for individual monkeys also revealed a significant main effect for training stage ($F(1, 468) = 28.42$, $p<E-4$, and $F(1, 126) = 13.74$, $p = 0.003$, for the two monkeys of the spatial match non match task, respectively).

Unexplained variance also tended to be higher for the trial where the reversal occurred, however the difference between trials arranged relative to the reversal

was not consistent across tasks (spatial match-nonmatch task: $F(2, 600) = 3.11$, $p = 0.045$; object match-nonmatch: $F(2, 1365) = 4.98$, $p = 0.007$; object choose-match: $F(2, 2937) = 2.04$, $p = 0.131$). There was no significant interaction between early/late training and serial position in either of the three tasks (spatial match-nonmatch: $F(2, 600) = 0.32$, $p = 0.726$; object match-nonmatch: $F(2, 1365) = 1.47$, $p = 0.231$; object choose-match: $F(2, 2937) <$ E-4, $p = 0.999$).

Repeating the same two-way ANOVA for unexplained variance averaged across time bins spanning all epochs showed similar results (Fig. S4). A significant main effect for training stage was observed, where unexplained variance increased from early to late training (spatial match-nonmatch: $F(1, 600) = 42.61$, $p<$E-4; object match-nonmatch: $F(1, 1365) = 30.28$, $p<$E-4; object choose-match: $F(1, 2937) = 95.61$, $p<$E-4).

This pattern of unexplained variance increase with training was not tied to systematic firing rate differences between training stages. We examined choice epoch firing rate across the same range of 3 trials around the reversal (Fig. 3C-D), using a similar two-way ANOVA to compare the response vector of firing rate across the grouping vectors (factors) of training level (i.e. early vs late), and trial distance relative to the reversal. This analysis revealed no consistent effects across tasks. There was no significant difference in firing rate for the trials examined in the spatial match-nonmatch task ($F(1, 600) = 0.02$, $p = 0.893$). Lower firing rate was observed for the late training stage in the object match-nonmatch: $F(1, 1365) = 35.8$, $p<$E-4. Higher firing rate was observed the late training stage in the object choose-match:

$F(1, 2937) = 20.53$, $p<E-4$. No significant effect of trial distance from the reversal was observed for any task, either (spatial match-nonmatch: $F(2, 600) = 0.02$, $p = 0.979$; object match-nonmatch: $F(1, 1365) = 0.02$, $p = 0.982$; object choose-match: $F(1, 2937) = 0.04$, $p = 0.963$). Repeating this analysis with firing rate averaged across all epochs yielded similar results as well (Fig. S4C-D). These findings suggest that the consistent increase in unexplained variance we observed after training was not dependent on systematic firing rate changes across tasks/monkeys.

We further wished to test which of the three task variables that we used in our regression model changed the most with training. To evaluate this, we calculated partial R-square separately for each of the three task variables (matching status, up/down direction of the correct target, or the correct/error outcome) that was used in our original regression model. Just as before, the amount of firing rate variance that this model failed to explain across all trials at each time point was calculated as $1-R^2$. Critically, this analysis revealed that the explanatory power of reward status (i.e. whether the trial was rewarded or unrewarded) diminished markedly as training progressed and the animal internalized the task rules, leading to more predictable outcomes. This was true across all task paradigms, while the contributions of the remaining variables remained stable (Fig. S5). These results suggest that reward status account for a greater proportion of firing rate variance early in the training process, when the subjects determine their responses based on these factors, on a trial-by-trial basis. These factors become less important once the monkey understands the task rule, and the appropriate

response becomes deterministic, and particularly the reward outcome does not inform the choice to be selected in the next trial.

Decoding analysis

To understand how training may affect the various types of task-related information that can be represented in prefrontal populations, we performed a decoding analysis. Decoding accuracy has been previously shown to increase for some specific task factors over the course of training, such as trials presenting matching versus nonmatching stimuli, though importantly, decoding has also been shown to remain constant for other specific task factors, such as the shape or location of these presented stimuli (Meyers, Qi et al. 2012).

We first sought to decode the match or nonmatch status of the sample stimulus. We performed this analysis both based on correct trials alone, and including correct and error trials. Decoding performance was compared between early and late training using a nonparametric randomization (permutation) test. We did not observe a systematic pattern of decoding changes across tasks as rule learning took place over the progression from early to late training. Rather, differentiating between a match and a nonmatch stimulus appeared to involve three different decoding patterns. In the spatial match-nonmatch task, we observed that the progression from early to late training led to significant (permutation test: $p < 0.05$) improvements in decoding the presentation of match or nonmatch stimulus (Fig. 4A-left)). This improvement was most evident when the classifier was

trained/tested with exclusively correct trials (solid lines in Fig. 4) than when error trials were also included (dotted lines in Fig. 4). In other words, neural activity after training better differentiated whether the second stimulus was presented on the right or left during the second stimulus presentation epoch. In contrast, no such differentiation was clearly evident in the object match-nonmatch task (Fig. 4A-center), regardless of whether error trials were included or not. Across our sample of neurons from this task, very few conveyed information about whether the second stimulus was a circle or triangle, both early and late in the training process. Finally, the object-choose-match task was characterized by the appearance of a bias in neural activity, early, which disappeared as training progressed (Fig. 4A-right). This meant that neurons represented whether the monkey would indicate the circle or triangle to be the match even before the sample was presented (as it was possible to do, since these trials were presented in blocks with the same cue stimulus). This decoding bias occurred across the full span of the trial, regardless of whether error trials were included or not. Similar patterns were evident for the decoding of the monkey's choice to saccade up or down. In the spatial match-nonmatch tasks, we observed that the progression from early to late training led to significant (permutation test: $p < 0.05$) improvements in decoding the chosen direction of the monkey's saccade (Fig. 4B-left). Decoding accuracy was greater in correct trials, than when both correct and error trials were used to train and test the classifier. By contrast, no improvement in decoding accuracy of the saccade direction was clearly evident in the object match-nonmatch task, regardless of whether error

trials were included or not (Fig. 4B-center). In the choose-match task, a bias was present again (Fig. 4b-right). Neurons represented early whether the monkey would choose to saccade up or down even before the targets were presented. This bias was most evident when the classifier was trained/tested with exclusively correct trials.

Decoding of other variables such as the configuration of the choice targets (i.e. whether a specific choice target appeared up or down regardless of whether this was the rewarded target); the monkey's choice to saccade for a match (regardless of whether this choice was correct); the direction of the rewarded choice target (regardless of whether the monkey chose to saccade to this target); and the combination of a match or nonmatch stimulus presentation with the direction of the rewarded target are depicted in Fig. S6.

To summarize, our results indicate that several different patterns of neuronal activity changes supported learning of different tasks. In the spatial match-nonmatch task, they relied upon representing the match or nonmatch stimulus, which gradually improved over the course of training. In the shape match-nonmatch task, the relative lack of decoding is less clear. Prior studies that observed relatively few cells representing object stimuli in the PFC may suggest this lack of decoding to be due to a so-called "needle-in-the-haystack" problem wherein the neural representations of the match or nonmatch stimulus may rely upon a small number of highly selective neurons that our sampling methods seemed to be largely unable to gather (Dang, Li et al. 2022). Finally, in the object choose-match task, the pre-

training paradigm was suggested to have led our monkeys to develop a strong bias towards a single saccade direction—even prior to the stimulus presentation—which gradually faded over the course of training as reversal rates were shortened and monkeys lost the ability to predict trials in advance.

Principal component analysis

While the decoding analysis discussed in the preceding section was a supervised analysis that identified relationships between neural activity and known task variables, unsupervised methods such as principal component analysis (PCA) can reveal training effects on latent population dynamics and neural trajectories that might not have been evident from our decoder analysis alone (Mante, Sussillo et al. 2013, Mozumder and Constantinidis 2023). We therefore performed a targeted PCA analysis (Mante, Sussillo et al. 2013), which decomposed the matrix of neural activity along the axis coding the match or nonmatch status (or shape of the cue in the choose match task), as well as along the axis coding the direction of the rewarded choice. The average population response for a given condition and time was represented as a point in state space projected into the two-dimensional subspace (Fig. 5A-B). Full neural trajectories from the empirical early and late training data are shown in Fig. S7. We then compared trajectories in the state space defined by these components across early and late training (Fig. 5 and Fig. S7).

We wished to test whether differentiation between task conditions would be represented by increased differences in their respective neural trajectories (Kobak,

Brendel et al. 2016). To perform such tests, we created a distribution of empirical trajectories by subsampling trials from all neurons and creating 100 permutations for the trajectories of the early and late training stage. We then calculated a null distribution, wherein all trials from both the early and late training datasets were pooled together, and then randomly shuffled into two surrogate datasets of equal size to the early and late training datasets. These surrogate datasets were used to plot firing rate trajectories in a state space defined by the same principal components to compute a null distribution over the course of 100 permutations.

We first sought to compare neural trajectories during the presentation of the match and nonmatch stimuli. In the spatial match-nonmatch task, we first grouped trials into two groups, depending on whether the sample stimulus was a match or nonmatch (see inset of Fig. 4A-left). Each trial group contained trials in which the monkey saccaded either up or down (as illustrated in the inset of Fig. 4B-left). We then examined the separation of neural trajectories in state space specifically during the interval of 700-1000ms after cue onset, during the sample presentation period, in the early and late training stages (Fig. 5A and 5B, respectively). For each training stage, we normalized the mean trajectory separation of the trials grouped based on match/nonmatch appearance by dividing this value by the separation of the trials grouped by the saccade-up vs. saccade-down conditions during the same, sample presentation, interval. This normalization by the mean within-group spread ensured the index reflected between-group separation relative to within-group variability, yielding a value greater than 1 when between-group distance exceeds

within-group spread. This allowed trajectory separation to be presented as a ratio that accounted for the different axes that were used for plotting the early and late training stage trajectories, thus controlling for the differences in the traces between early and late training conditions and ensuring a fair comparison.

Trajectory separation was then compared in early and late data to obtain a value we refer to as ΔSeparability (Fig. 5C-E). A distribution of ΔSeparability values was obtained for the empirical data by sampling different trials of each neuron (Blue curves in Fig. 5C), and also for the data obtained by randomizing the early and late labels (orange curves in Fig. 5C). The empirical increase in separability between the early and late training was significantly higher than what would be expected by chance (Wilcoxon test; $p<E-4$; Fig. 5c-left). This result mirrored the increase in match vs. nonmatch decoding accuracy described above (Fig. 4a - left).

We repeated the same analysis for the separability of neural trajectories in the spatial match/nonmatch task, during the interval of 1500-1900ms after cue onset, during the choice epoch (Fig. 5B). We now grouped trials based on saccade direction, examined trajectory separation during the same, choice epoch, interval, and normalized the separability by the mean distance of trials grouped based on match/nonmatch (Fig. 5C right). In this case, the difference in separability between early and late training stages did not reach statistical significance over chance (Wilcoxon test, $p=0.088$). This result mirrored the difference in decoding accuracy described in Fig. 4B-left.

Importantly, the PCA analysis also revealed an increase in separability in the latent space for the object match-nonmatch task at the time of the sample (match/nonmatch) stimulus presentation (Fig 5d-left, Wilcoxon test, $p<E-4$). This was so, even though decoding of the match and nonmatch conditions had failed to reveal a clear improvement (in Fig. 4A-center).

In the case of the object choose-match task, we computed match/nonmatch separability at the time the two targets were presented and the monkey needed to choose one of the two, which corresponded to the interval of 1000-1400ms after cue onset, during the saccade generation in this task. We then normalized the mean trajectory separation of the trials grouped based on match direction by dividing the separation value by the separation value of the trials grouped by the circle vs. triangle cue conditions during the same saccade generation interval. In this case too, separability increased from the early to the late training stage (Wilcoxon test, $p<E-4$). In conclusion, PCA analysis revealed some common patterns of training effects in latent variables across our tasks, even in cases when explicit decoding of task conditions was not evident.

DISCUSSION

Prior studies have established that different aspects of working memory such as processing speed, capacity, or the ability to multitask can be modified through training (Klingberg, Forssberg et al. 2002, Bherer, Kramer et al. 2008, Constantinidis and Klingberg 2016). Such training has been proven to be beneficial particularly for clinical populations, such as stroke, ADHD, and schizophrenia patients (Klingberg, Forssberg et al. 2002, Westerberg, Jacobaeus et al. 2007, Jaeggi, Buschkuehl et al. 2008, Subramaniam, Luks et al. 2012) as well as aged adults (Anguera, Boccanfuso et al. 2013). A debate still exists regarding if and how these benefits may generalize ("transfer") to untrained domains, including everyday life (Cortese, Ferrin et al. 2015, Schwaighofer, Fischer et al. 2015).

A variety of corresponding changes in neural mechanisms, particularly in the prefrontal cortex—which is widely considered to be the seat of WM—have been described (Qi and Constantinidis 2012, Averbeck and O'Doherty 2022). For example, prior studies have demonstrated how the decoding of task variables from the activity of prefrontal populations has provided a strong neural correlate of rule learning (Meyers, Qi et al. 2012). Neurophysiological research—supported by fMRI studies—has also highlighted increases in prefrontal activity after training to indicate increased specialization or strengthened representations of stimuli (Garavan, Kelley et al. 2000, Mendoza-Halliday and Martinez-Trujillo 2017). However, other studies have identified decreases in activity after training that may

indicate possible improvements in efficiency instead (Schneiders, Opitz et al. 2011, Takeuchi, Taki et al. 2013). Studies have also demonstrated more subtle training effects in firing rate, such as modifications to network modularity and correlation (Qi and Constantinidis 2012, Finc, Bonna et al. 2020). The malleability of WM is thus believed to be mediated by the underlying plasticity in neural responses. Such plasticity has been observed over the course of even individual training sessions, and also during the more extended process that is entailed by rule learning—as shown by our current training paradigm (Asaad, Rainer et al. 1998, Bartolo and Averbeck 2020). Rule learning has long been examined through similar reversal tasks as a measure of behavioral flexibility (Schoenbaum, Chiba et al. 2000, Izquierdo, Brigman et al. 2017). However, reported results of training are most often limited to an individual task and it has been difficult to determine common and unique elements, investigated with the same experimental and analytical techniques, other than artificial neural network simulations (Masse, Yang et al. 2019, Xie, Liu et al. 2022).

Prefrontal activity changes

Prior studies have demonstrated strong evidence of prefrontal persistent activity dynamically incorporating multiple types of both spatial and object information (Amengual and Ben Hamed 2021, Curtis and Sprague 2021) with the overall effects of training suggesting increased recruitment of more prefrontal neurons and generation of persistent activity at higher levels (Tang, Riley et al. 2022). Our present

study was therefore guided by experimental and theoretical predictions to examine whether rule learning could be characterized by changes in firing rate across all tasks. Somewhat unexpectedly, we found that the overall increase in firing rate after training was not a general phenomenon across different tasks. Instead, the progression of rule learning seemed to move firing rate in different directions for different monkeys/tasks, both increasing and decreasing. This result mirrors the conflicting fMRI studies that have suggested the former to indicate strengthened neural representations that underlie short-term information maintenance and manipulation (Garavan, Kelley et al. 2000, Hempel, Giesel et al. 2004, Mendoza-Halliday and Martinez-Trujillo 2017), just as the latter has been suggested to indicate improvements in efficiency (Schneiders, Opitz et al. 2011, Takeuchi, Taki et al. 2013). Our observation of both effects across tasks—increases and decreases—would therefore imply that this debate is not mutually exclusive to one side or the other, and instead, either effect may be responsible for mediating training, as determined by the individual monkey/task.

Nevertheless, at least some common changes in firing rate were evident in our study. For example, our observed trend of a significant decrease in choice epoch activity, particularly early in training, seemed to directly parallel the period of behavioral adjustment wherein each monkey gradually learns the rule—through trial and error—that the rewarded choice target is not random, but instead, is determined by the previously presented stimuli. This was the rule learning that allowed our monkeys to better anticipate which choice would be rewarded in their

assigned tasks. Decreased choice epoch activity thus paralleled the observed trend of decreased DIP magnitude that behaviorally characterized the acquisition of the rule until eventually, the rule was mastered and choice epoch firing rate returned to its original state to reflect this new mastery.

The more consistent effect we observed was that of an increase in unexplained variance after training, independent of mean firing rate changes. Prior studies have shown that in real life, animals learn reversal learning tasks via reinforcement learning based on reward outcomes (Kim, Chow et al. 2025). In such tasks, increased unexplained variance in firing rate has been observed during reversals due to the encoding of decisions that represent a belief about the current state of the task—which is then contradicted by such reversals (Bartolo and Averbeck 2020). The change in choice preference reversal that followed from this contradiction is thus predicted by increased unexplained variance as the monkey encoded the task factors that this new preference demanded. Such increases were suggested to be caused by prefrontal neurons dynamically encoding decisions associated with Bayesian subjective values to represent a belief about the current state of the task. To examine how this concept may apply over the process of rule learning, we ran a linear regression to compare how effectively the pre-training model of firing rate could explain variance during early and late training based on the three factors of matching status, rewarded target direction, and reward status, this is, whether the trial was answered correctly or not. This revealed a significant increase in unexplained variance across all monkeys/tasks, suggesting that as the

monkeys progressively acquired the new rule, they increasingly encoded new task factors that were not linearly related to the three variables of our model. Our effect was attributed primarily to less variance being explained during the delivery of reward. In other words, delivery of reward was very informative, and neural activity tracks this variable early in the training when monkeys rely on this feedback to adjust their behavior in each trial but becomes less important when monkeys understand the rule and the outcome becomes more expectable. By contrast, the degree to which the other two predictors of our regression model—matching status and rewarded target direction— were informative to the monkeys did not undergo any consistent changes over the course of training across tasks.

Encoding of new factors

One of the most important issues related to the role of prefrontal activity is how the representation of different types of information may be altered by rule learning to achieve more effective task performance (Rigotti, Barak et al. 2013). We applied a decoding analysis to examine how training affects the decoding of different types of task factors, with the prediction that training would lead to an overall improvement in decoder performance across tasks—and by extension, an improved representation of the information needed for success. Somewhat unexpectedly, each of our tasks revealed a different pattern of decoding, with only the spatial match-nonmatch task displaying the improved decoding of match/nonmatch information that might have been expected. When we tested the two object tasks to

determine if the same pattern would occur, the object match-nonmatch task revealed no clear training effects, while the object choose-match task revealed a strong bias in neural activity, early, which disappeared as training progressed.

Yet when we used PCA analysis to examine latent population dynamics (Mante, Sussillo et al. 2013, Mozumder and Constantinidis 2023), we found an increased separation of population response trajectories in state space during the match/nonmatch presentation across all tasks. This suggests that rule learning leads to at least a subtle differentiation between task conditions in prefrontal activity that was not evident from the decoder analysis which examined each explicit task variable separately.

Confronting the object-spatial divide

Our findings demonstrate that although both spatial and object information may be represented in prefrontal activity, these modalities seem to present a variety of key differences that would imply at least some degree of task-specificity in the neural mechanisms of rule learning (Nystrom, Braver et al. 2000, Ester, Serences et al. 2009). Most importantly, a large proportion of neurons represented spatial location, were modulated by the task, and differentiated between a left and right stimulus, when these represented a match or nonmatch condition. In contrast, few neurons represented the object stimuli, both before and after training, resulting in a relative paucity of information about object shape when it signaled a match or nonmatch.

Such differences are emphasized by studies that identified dynamic interactions with lower cortices, such as the primary visual cortex, during training of object WM tasks as well (Fyall, El-Shamayleh et al. 2017, Skirzewski, Molotchnikoff et al. 2021). Any debate on the generalization of mechanisms across spatial and object rule learning should not be seen as an all or nothing divide. Instead, both theories seem to work in tandem to an extent that is determined by context. The object-spatial divide may not be absolute, but its existence is evident in our observations of the clear contrast between the neural activity of our spatial task compared to the parallel object tasks.

While there is some evidence of prefrontal persistent activity dynamically incorporating multiple types of both spatial and object information, thus implying that this may not be an exclusive mechanism to spatial WM alone (Amengual and Ben Hamed 2021), the extent to which this may overlap between modalities is highly debated. A lack of shared training effects in our tasks would imply at least some degree of separation of spatial and object rule learning mechanisms, consistent with prior theories that object stimuli are maintained via lower cortices—such as the primary visual cortex—while the PFC plays a more supervisory role instead, among other possibilities (Ester, Serences et al. 2009, Lawrence, van Mourik et al. 2018). Further research will be necessary to identify the precise limits of this object-spatial divide, and should therefore investigate beyond the PFC alone, to confront these lower cortices to provide an improved understanding of if and how such areas may be involved with rule learning.

Likewise, our current training paradigm would be unable to fully assess the potential influence of phenomena such as match suppression, where activity elicited after repeated presentation of the same stimulus is reduced, which may be informative regarding the matching status of a given stimulus with a prior stimulus (Grill-Spector, Henson et al. 2006). Identifying which factors determine when a given rule learning mechanism is either generalizable or task-specific may also provide further explanation for previously nebulous phenomena, such as the previously noted "special status" of spatial information in WM via the bump attractor model (Foster, Bsales et al. 2017). Furthermore, such investigations will allow new insights on the overall role of persistent activity in the PFC and how the generalization of this mechanism may compare to the generalization of alternative mechanisms that have been proposed for WM, such as the modulation of synaptic resources (Mongillo, Barak et al. 2008). To that end, our results provide a framework for exploring how prefrontal mechanisms parallel the progression of rule learning, ultimately laying the foundation for addressing each of these questions and beyond.

ACKNOWLEDGEMENTS

Supported by NIH award number R01 EY017077, and F31 EY035546. We wish to thank Chrissy Suell, Kayla Yetman, Kris Clifft, and Jaela Bills for technical help and Emilio Salinas, Terry Stanford, Ben Rowland and Robert Hampson for helpful comments on an earlier version of this manuscript.

AUTHOR CONTRIBUTIONS

CC designed the experiment. RJ, WD, TG, and JZ performed data analyses. CC, WD, and RJ wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DATA AVAILABILITY

Data for the current study will be made available through Zenodo upon acceptance of the paper.

CODE AVAILABILITY

The code used to process the results and generate the figures will be made available at Github upon acceptance of the paper.

FIGURES

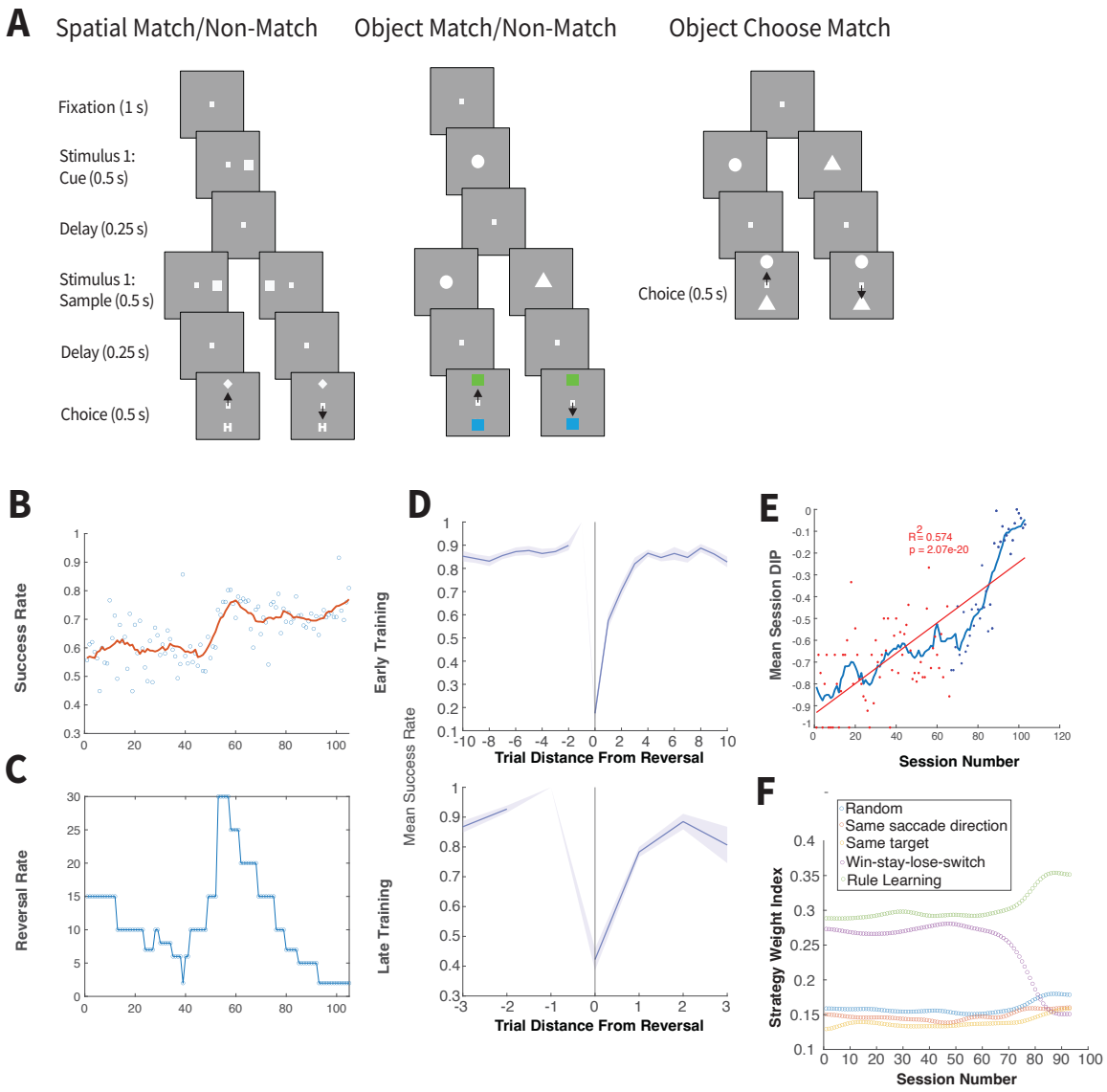


Figure 1: Experimental Tasks and Behavior. A. Schematic of the spatial match-nonmatch, object match-nonmatch, and object choose-match tasks. In the match-nonmatch tasks, the first stimulus (cue) remained fixed throughout training (i.e. a square on the right side of the screen for the spatial task, and a circle at the center of the screen for the object task). After a delay, this was followed by a match

stimulus identical to the cue, or a nonmatch stimulus (i.e. square on the left side of the screen for the spatial task, and a triangle at the center of the screen for the object task). After a second delay, the monkey had to saccade to the “Diamond” or “Green Square” target to indicate a match or the “H” target or “Blue Square” to indicate a non-match. In the object choose-match task, the cue could be either a circle or triangle at the center of the screen. After a delay, two targets appeared (i.e. a circle and triangle) and the monkey had to saccade to whichever matched this cue. B. Mean performance of an example monkey at each session—indicated via dots—across training. Red lines indicate centered running average with a symmetric window of 5. C. Reversal rate between trial blocks for an example monkey at each session across training. Reversal rates of 2 indicate randomized reversals, by convention. D. Example monkey performance (y-axis) across trials relative to their distance from the reversal (x-axis). Shaded areas indicate standard error. The last trial that preceded the reversal was—by definition—always correct, and were thus removed from these plots as artifactual. Upper, early training; lower, late training. E. Mean session DIP (y-axis) of an example monkey across training, with early/late training highlighted in red/blue, respectively. Thin lines represent centered running average with a symmetric window of 5. Thick lines represent linear regression of DIP on session number. F. Weight indices (y-axis) of an example monkey’s strategies across training. Each color dot represents a different strategy in one session.

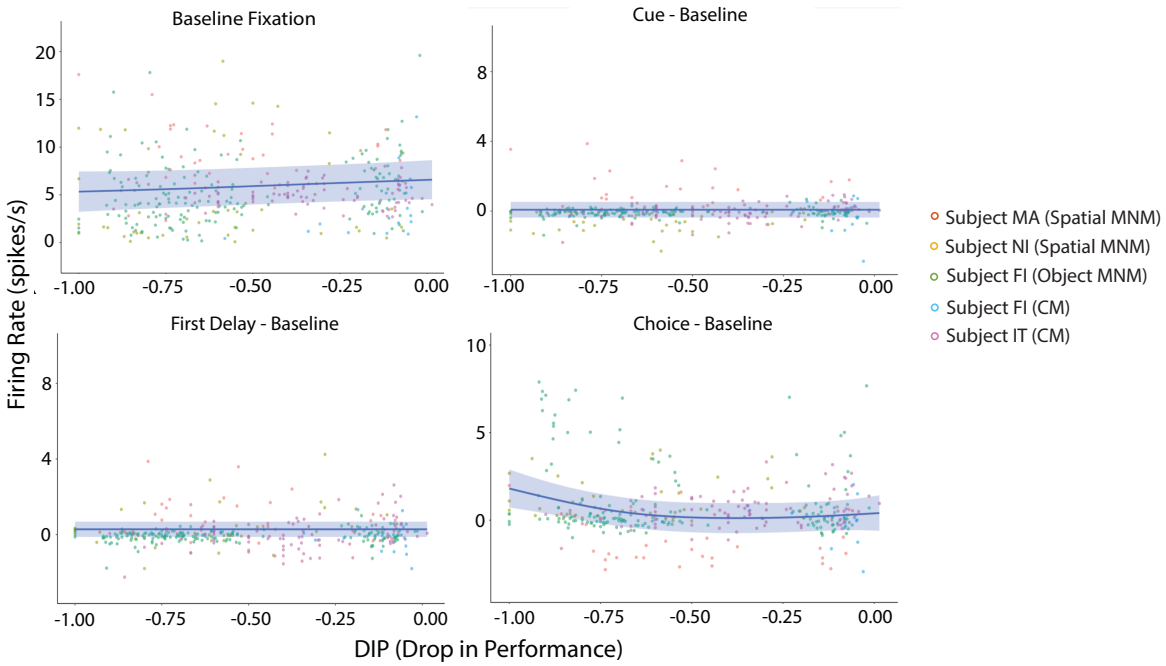


Figure 2: Nonlinear training effects on firing rate. The trajectory of the mean session firing rate (y-axis) in different epochs (baseline fixation, as well as cue epoch, first delay epoch, and choice epoch after subtracting the mean baseline fixation activity) across mean session DIP (x-axis). Solid lines indicate training effects of the GAMM fittings. Shaded ribbons denote the 95% confidence intervals (CIs). Each dot represents a single session, and individual monkeys are indicated by different colors. Outliers were removed from the visual plots by scaling the y-axis, with a total of 2 outliers removed from the fixation epoch plot, 1 outlier removed from the cue epoch plot, 1 outlier removed from the first delay epoch plot, and 2 outliers removed from the choice epoch plot for better visualization, though these points were still used in the curve estimation.

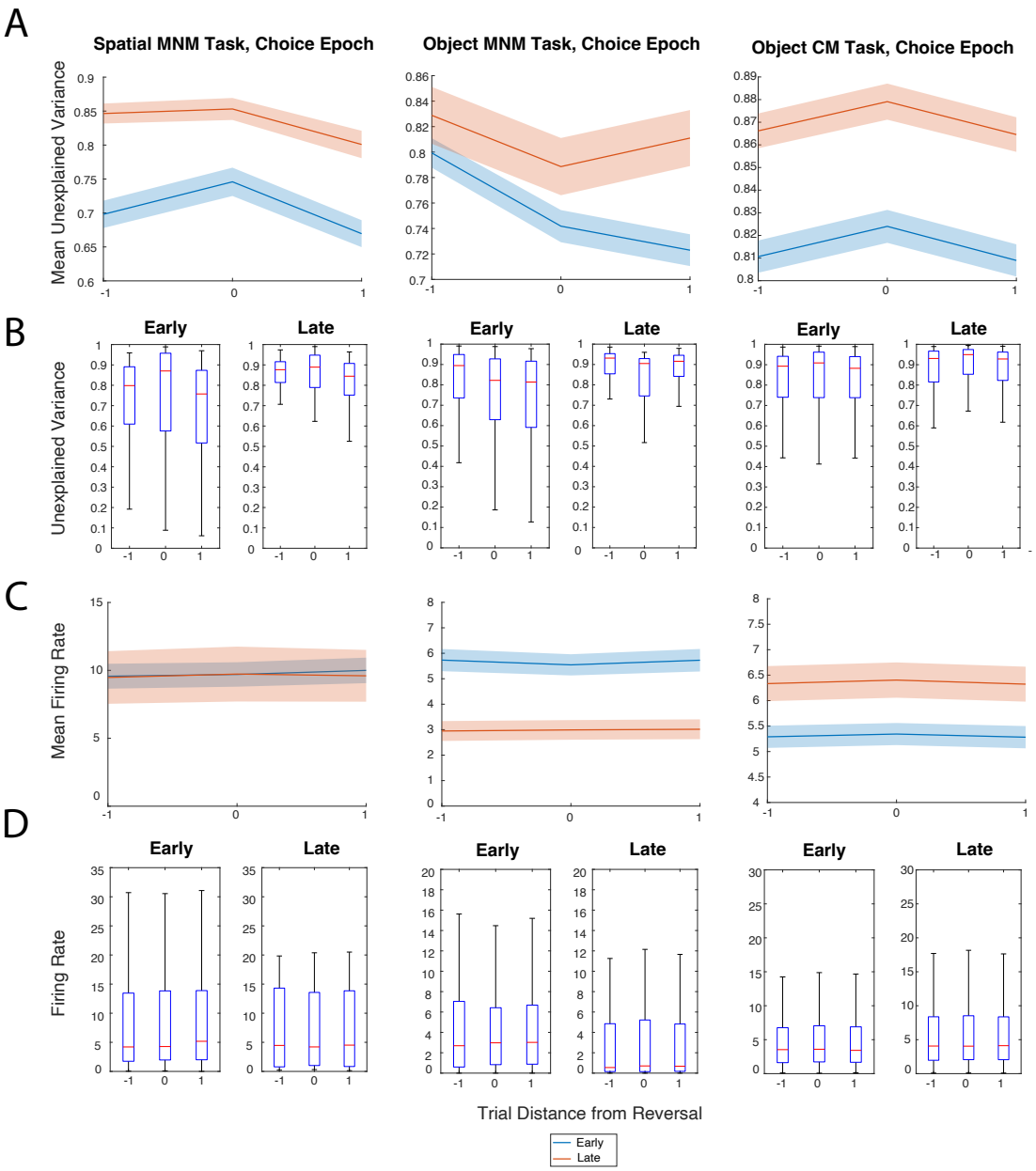


Figure 3: Training increases unexplained variance across tasks. A. Mean unexplained variance of firing rate in the choice epoch across trials relative to their distance from the reversal. Shaded areas indicate standard error. Blue indicates early training while red indicates late training stage. B. The same data as in A represented as boxplots. On each box, the central mark indicates the median, and

the bottom and top edges indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers—defined as any values that were more than 1.5 * the interquartile range (i.e. the distance between the 25th and 75th percentiles) below the 25th percentile, or above the 75th percentile. C-D. Same conventions as A-B but examining choice epoch firing rate instead of unexplained variance.

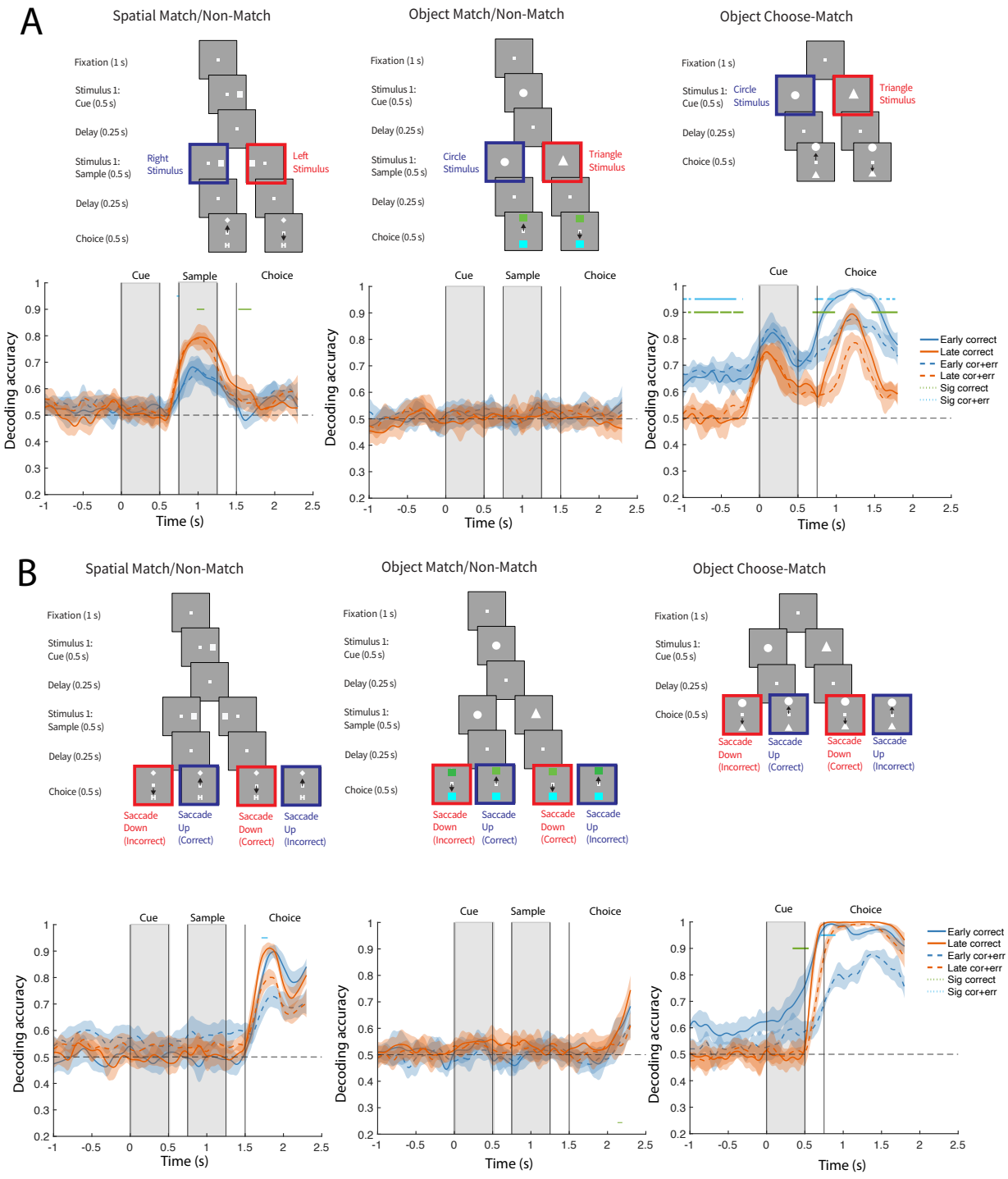


Figure 4: Training has no systematic pattern in decoding effects across tasks.

Top diagrams illustrate A. the presentation of match or nonmatch stimulus (i.e. trials presenting stimulus pairs with the same location/shape in the spatial/object

match-nonmatch tasks, or trials presenting the circle and triangle stimulus in the object choose-match task), and B. the monkey's chosen saccade direction (up or down) for each task. Bottom plots illustrate the accuracy of decoding this task factor from the early training (blue) versus late training (red) neurons over the time course (x-axis) of each task. Vertical black lines indicate each task epoch. Solid blue and red lines indicate the use of exclusively correct trials for training and testing data. Dashed blue and red lines indicate the use of both success and error trials for training and testing data. Shaded ribbons represent 95% confidence intervals. Significant differences between early and late training decoding are marked overhead in green and cyan (two tailed permutation test, $p < 0.05$). We performed our decoding analyses using an equal number of cells drawn from early and late training sessions (Fig. S6). This was a total of 100 early and late training cells from the spatial match-nonmatch task (left), 400 early and late training cells from the object match-nonmatch task (middle), and 400 early and late training cells from the object choose-match task (right).

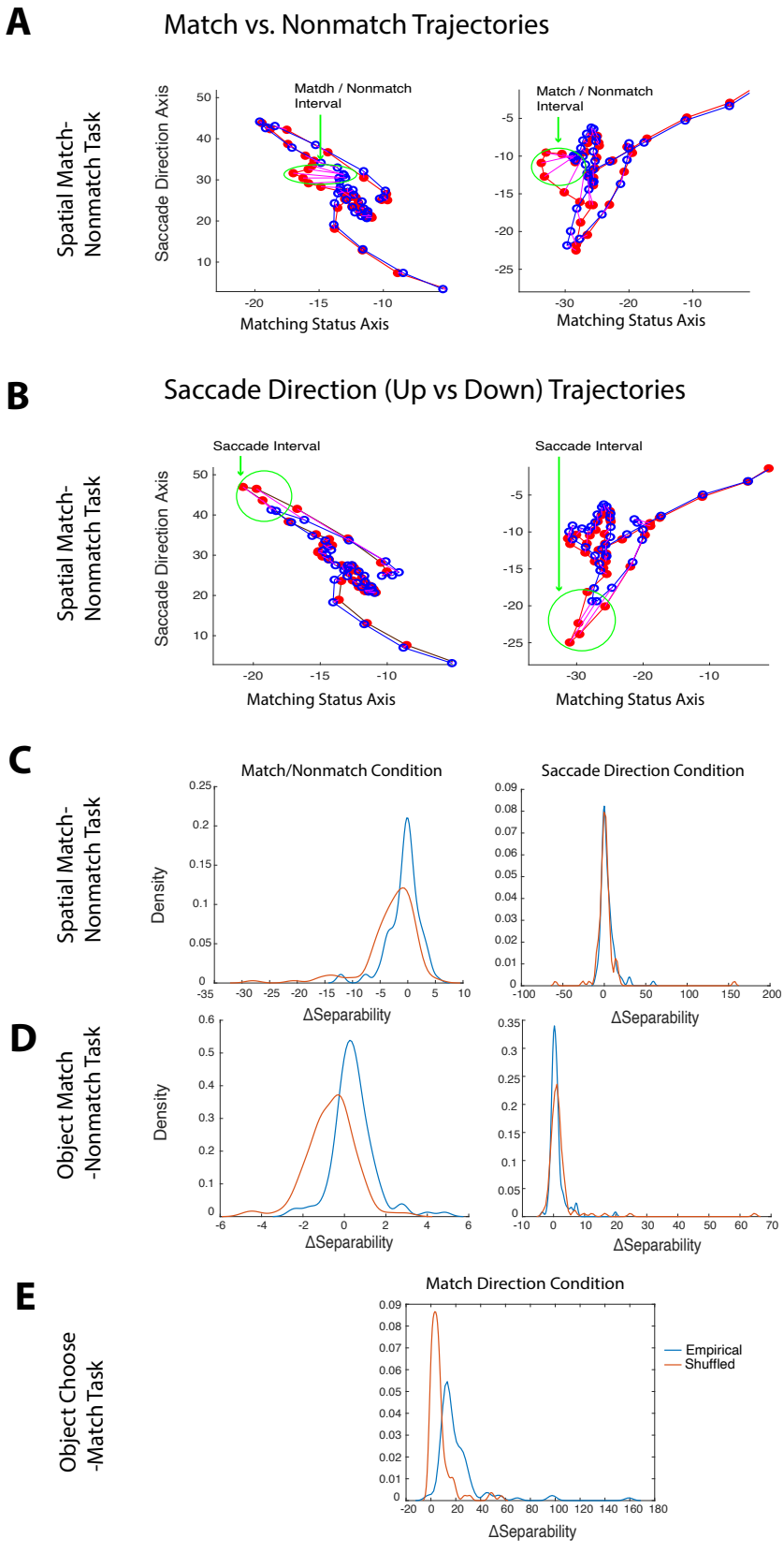
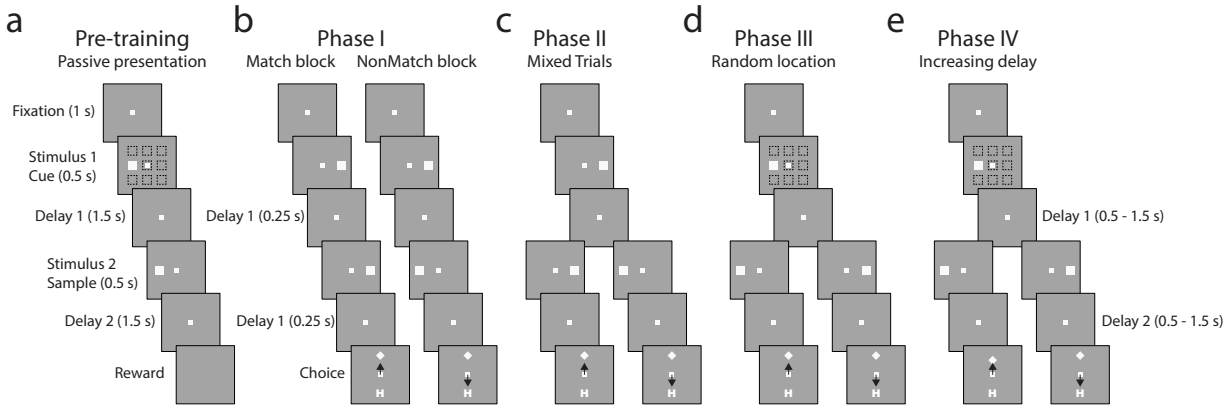

A
Match vs. Nonmatch Trajectories
Spatial Match-Nonmatch Task
Saccade Direction Axis
Matdh / Nonmatch Interval
Matching Status Axis
Match / Nonmatch Interval
Matching Status Axis
B
Saccade Direction (Up vs Down) Trajectories
Spatial Match-Nonmatch Task
Saccade Direction Axis
Saccade Interval
Matching Status Axis
Saccade Interval
Matching Status Axis
C
Spatial Match-Nonmatch Task
Match/Nonmatch Condition
Density
ΔSeparability
Saccade Direction Condition
ΔSeparability
D
Object Match -Nonmatch Task
Density
ΔSeparability
ΔSeparability
E
Object Choose -Match Task
Match Direction Condition
Empirical
Shuffled
ΔSeparability

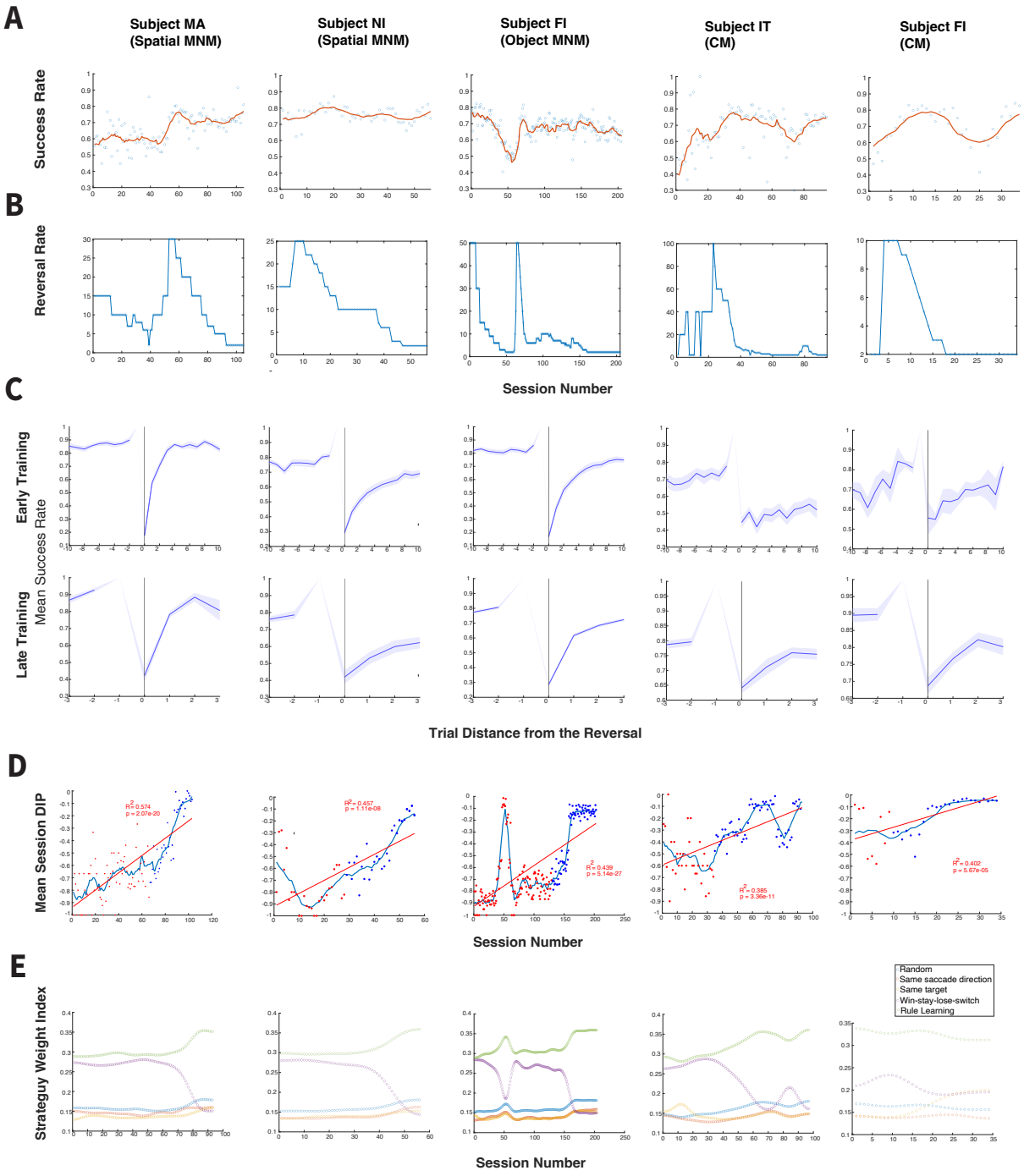
Figure 5: Dynamics of populations responses in PFC. A. Comparison of the complete neural trajectories of early (left) and late (right) training across the match (blue) and non-match (red) conditions of the exemplar spatial match-nonmatch task. Separability is compared using the interval that is highlighted in green. These intervals were determined a based on the time of different task events. B. Same as Part A for the rewarded target up (blue) and down (red) conditions of the spatial match-nonmatch task. C-E. The distribution (y-axis) of the change in separability (x-axis) between empirical early and late training neural trajectories (blue) versus a null distribution (orange) that was calculated from a pair of randomly shuffled surrogate datasets. Separability was defined as the Euclidean distance between trajectories across both axes of a two-dimensional subspace capturing the variance due to the combination of the presented stimulus (i.e. match or nonmatch in the spatial and object match-nonmatch tasks, and circle or triangle stimulus presentation in the object choose-match task) and the saccade direction of the rewarded choice target (i.e. up or down) task conditions. This was examined for the match and nonmatch conditions (left) and the target direction conditions (right). Experimental tasks were examined in the order of C. spatial match-nonmatch task, D. object match-nonmatch task, and E. object choose-match task.

SUPPLEMENTARY INFORMATION



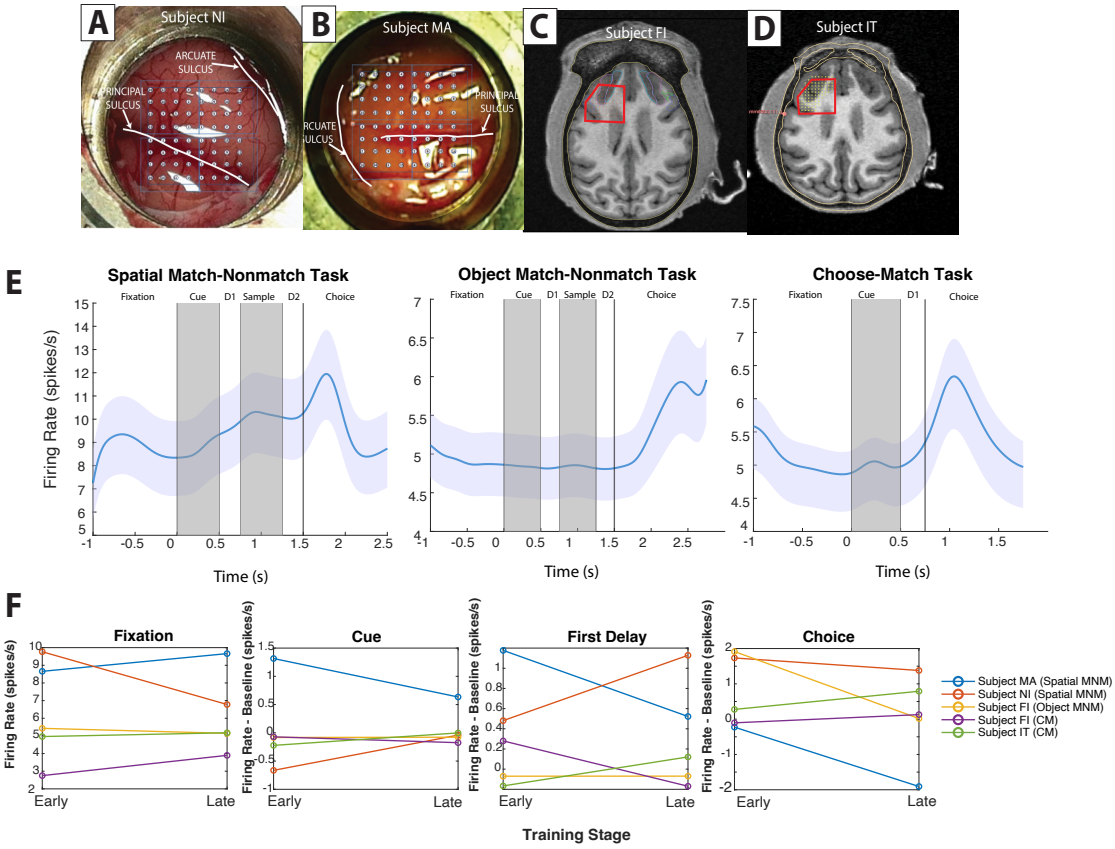
Supplementary Figure S1: Full training paradigm. A schematic of all training phases from the previous multi-phase spatial match-nonmatch task training paradigm that our current training paradigm was founded on (Tang, Riley et al. 2022). A-E Successive frames illustrate the sequence of events in the tasks used in progressing training phases. A. During the passive pre-training phase, the monkey only had to fixate while the stimuli were displayed at any one of the nine locations on the screen. B. In Phase I, a stimulus was always presented to the right, followed by a match stimulus in a block of trials and by a nonmatch stimulus in another block of trials. At the end of the trial, two choice targets appeared, and the monkey had to choose the "Diamond" target in match blocks and the "H" target in nonmatch blocks to get a reward. C. In Phase II, match and nonmatch trials were mixed in a block. This was identical to the spatial match-nonmatch task that was used in our current experiment. D. In Phase III, the stimulus location of the first stimulus could vary. E. In Phase IV, the duration of the delay period increased. The

passive stimulus set continued to be presented at the beginning of each session throughout all four phases of training in this paradigm.



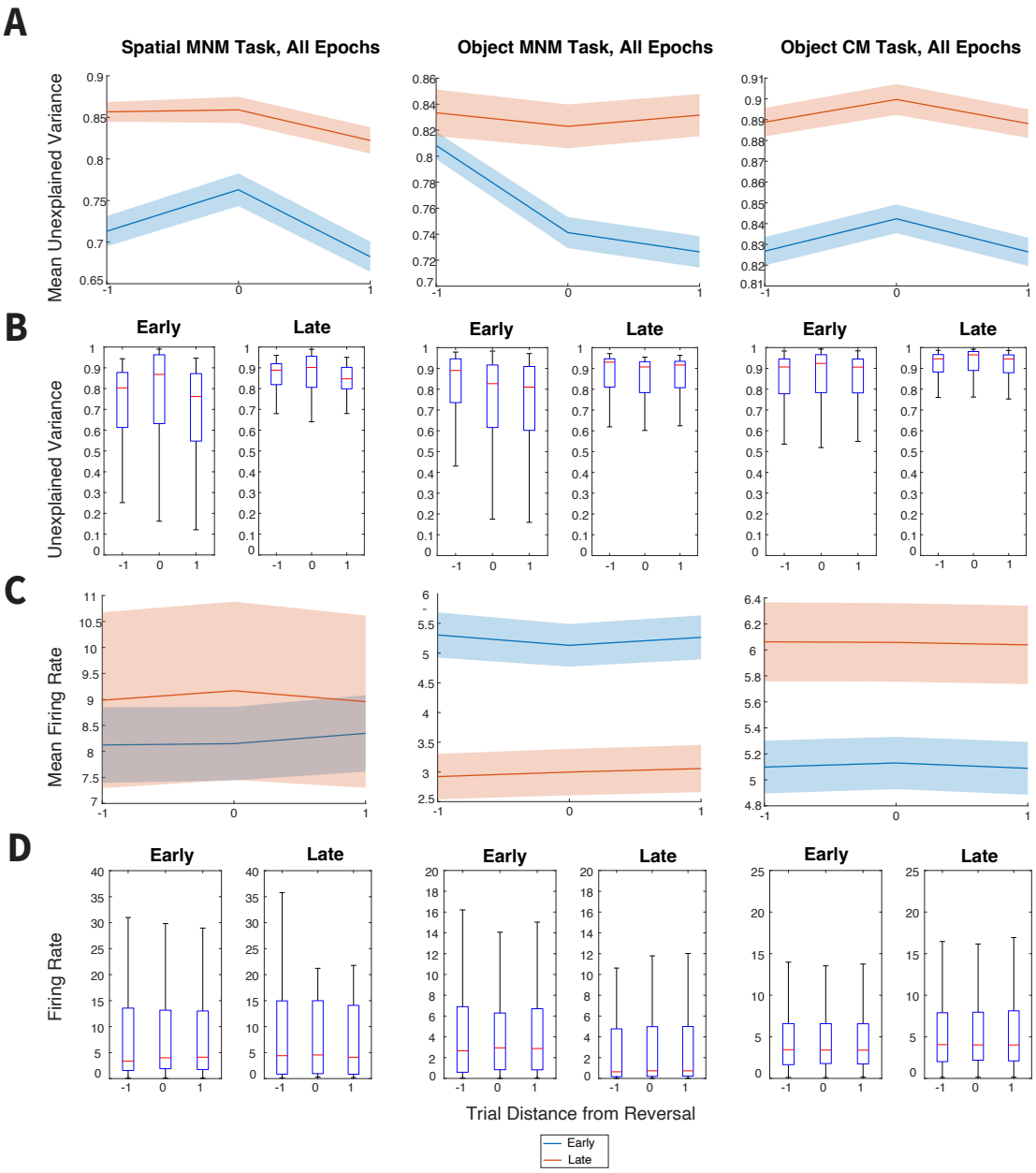
Supplementary Figure S2: Expanded task behavior. Expanded behavioral results. This includes the mean success rate (A), the reversal rate (B) between the presentation of match and nonmatch stimulus (or circle versus triangle cue

stimulus in the case of the choose-match task), mean performance across trials relative to their distance (C) from the reversal between the presentation of match and nonmatch stimulus (or circle versus triangle cue stimulus in the case of the choose-match task) for early versus late training, the mean DIP (D) of each session over the course of training, and the weight index of five possible strategies (E) over the course of training, across all monkeys/tasks. All conventions are the same as Fig. 1.

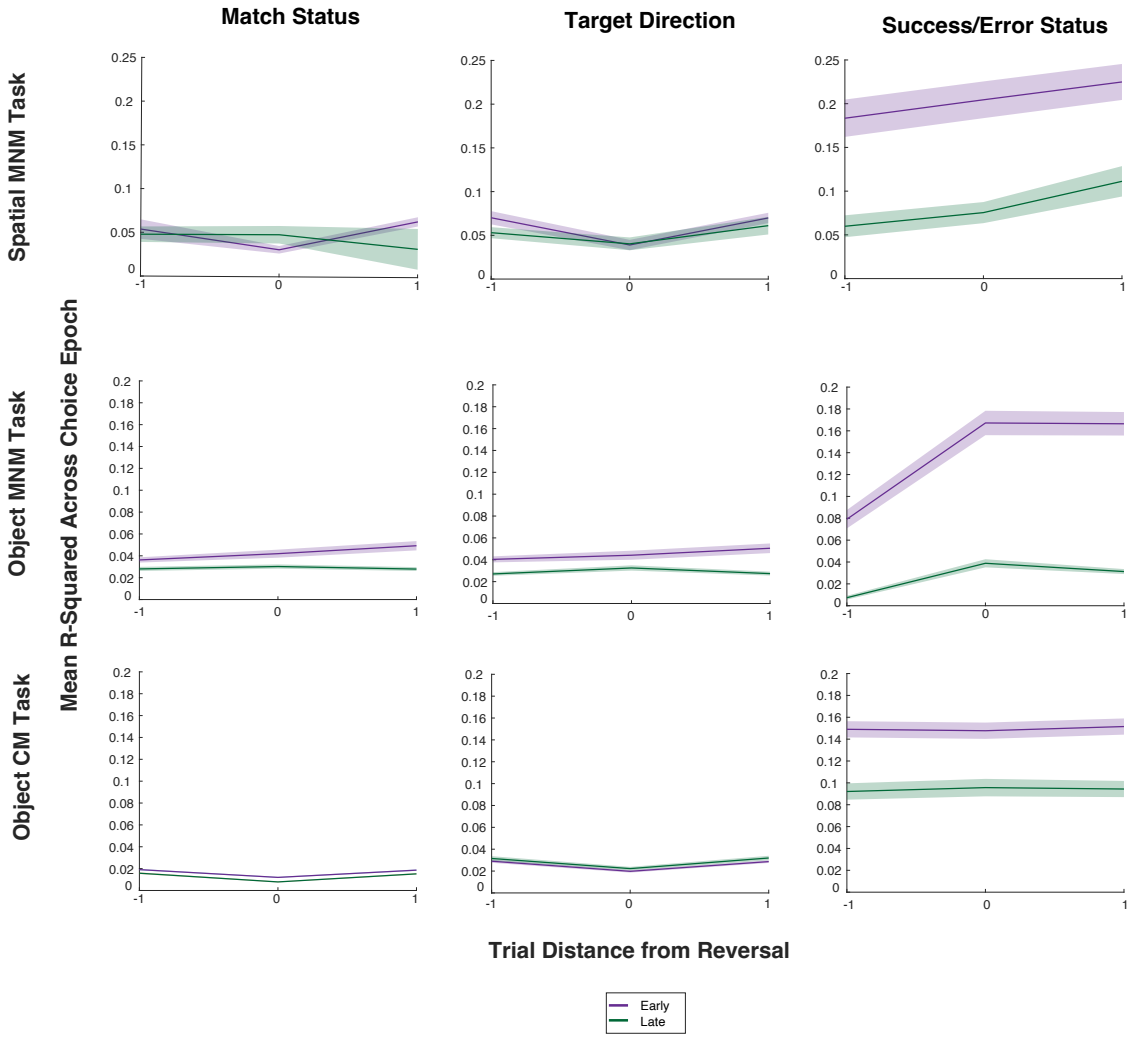


Supplementary Figure S3: Neural recordings and activity. A. Position of the electrode array that was implanted into the dorsolateral and ventrolateral prefrontal cortex of Subject FI indicated in red. This was centered across the dorso-ventral and anterior-posterior axis—as defined by the line connecting the genu of the arcuate sulcus to the frontal pole. B. Position of the electrode array that was implanted into the dorsolateral prefrontal cortex of Subject MA, as indicated relative to the principal sulcus and arcuate sulcus. C. Position of the electrode array that was implanted into the dorsolateral and ventrolateral prefrontal cortex of Subject IT indicated in red. This was centered across the dorso-ventral and anterior-posterior

axis—as defined by the line connecting the genu of the arcuate sulcus to the frontal pole. D. Position of the electrode array in the dorsolateral prefrontal cortex of Subject NI, as indicated relative to the principal sulcus and arcuate sulcus. E. Peri-stimulus time histogram (PSTH) plots depicting mean firing rate (y-axis) over the time course of the trial (x-axis) across all cells from each of our three experimental tasks. Gray bars indicate stimulus presentations. Time 0 represents the cue presentation. F. Mean baseline fixation and firing rates (y-axis) for after subtracting the mean baseline fixation activity for the cue epoch, first delay epoch, and choice epoch of each monkey/task (indicated by colors), across early vs late training (x-axis).

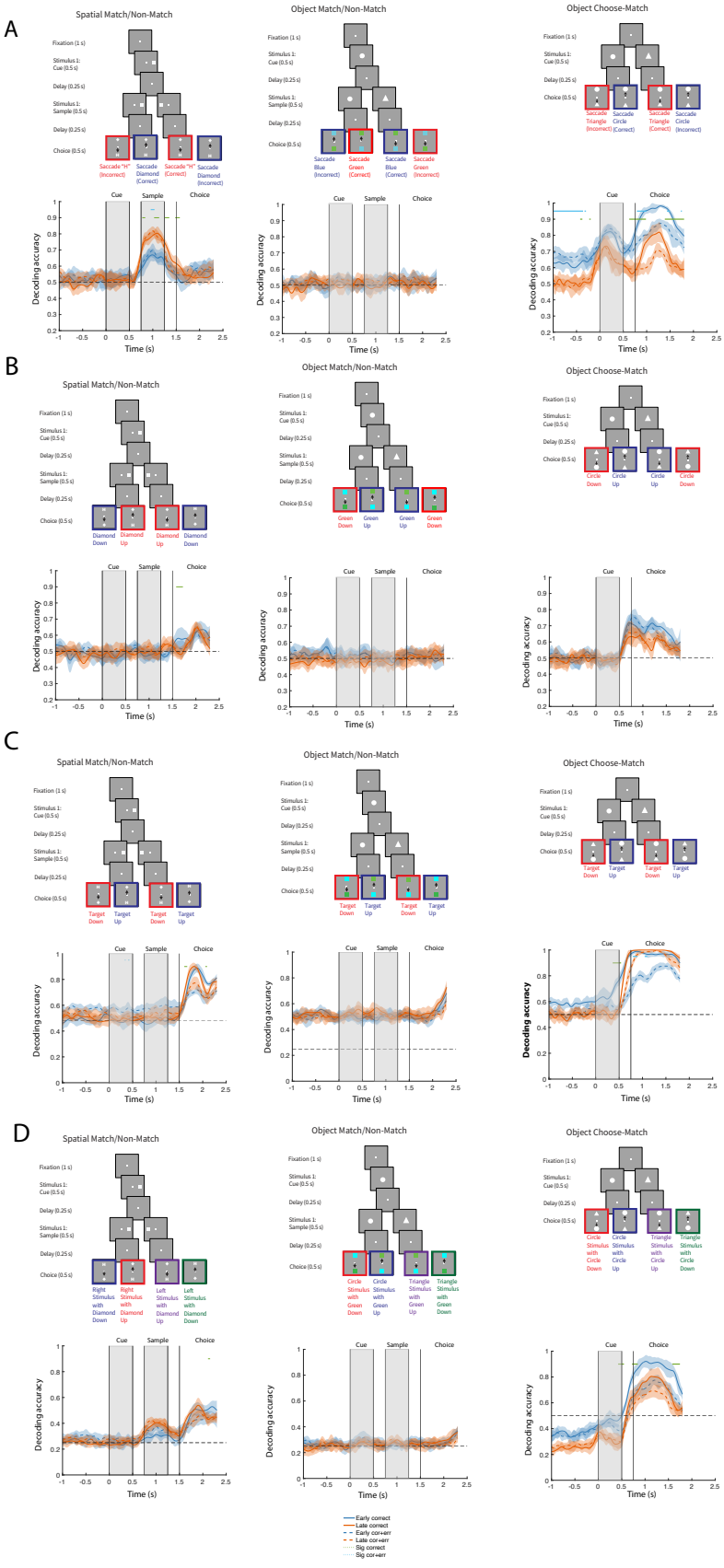


Supplementary Figure S4: Expanded Unexplained Variance. Unexplained variance of firing rate (A-B) and the firing rate (C-D) in the spatial match-nonmatch (left), object match-nonmatch (middle), and object choose-match (right) tasks averaged across all epochs. Conventions are the same as in Fig. 3.

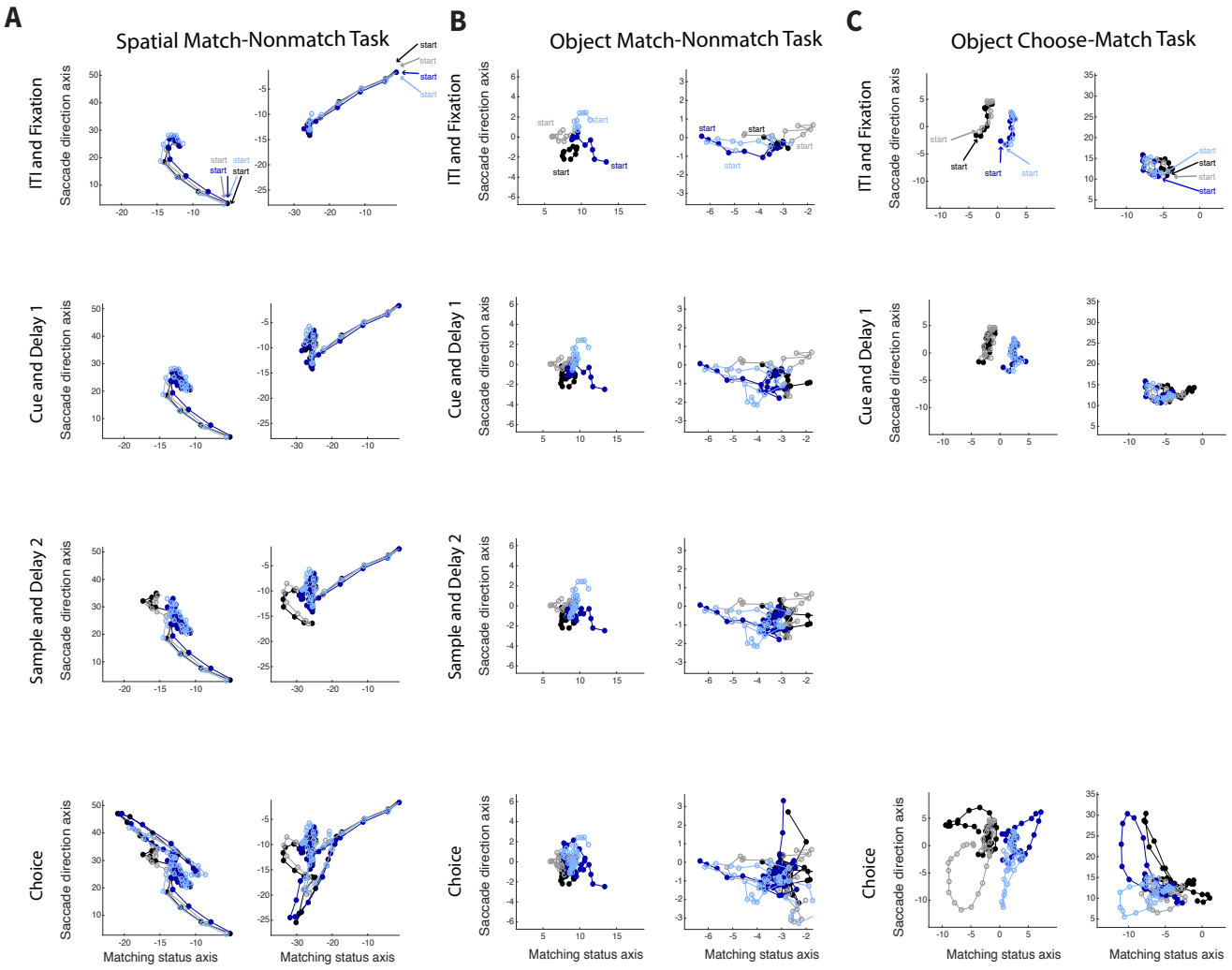


Supplementary Figure S5: Partial correlation of task variables with firing rate.

Mean partial R^2 of firing rate from each task variable that was used in our regression model during the choice epoch across trials relative to their distance from the reversal. Shaded areas indicate standard error. Purple indicates early training while green indicates late training stage.


A
Spatial Match/Non-Match
Object Match/Non-Match
Object Choose-Match
B
Spatial Match/Non-Match
Object Match/Non-Match
Object Choose-Match
C
Spatial Match/Non-Match
Object Match/Non-Match
Object Choose-Match
D
Spatial Match/Non-Match
Object Match/Non-Match
Object Choose-Match
Cue
Sample
Choice
Decoding accuracy
Time (s)
Early correct
Late correct
Early cor+err
Late cor+err
Sig correct
Sig cor+err

Supplementary Figure S6: Expanded decoder analyses. Decoder performance for all task factors that were not included in Fig. 4. We performed our decoding analyses using an equal number of cells drawn from early and late training sessions. This was a total of 100 early and late training cells from the spatial match-nonmatch task (left), 400 early and late training cells from the object match-nonmatch task (middle), and 400 early and late training cells from the object choose-match task (right). In each task, top diagrams illustrate the A. the monkey's choice to saccade for a match regardless of whether this choice was correct, B. the configuration of the choice targets (i.e. whether a specific choice target appeared up or down regardless of whether this was the rewarded target), C. the direction of the rewarded target (regardless of whether the monkey chose to saccade to this target), and D. the combination of the presentation of match or nonmatch stimulus with the direction of the rewarded target. Bottom plots illustrate the accuracy of decoding (y-axis) this task factor from the early (blue) versus late (red) neurons over the time course (x-axis) of each task. Vertical black lines indicate each task epoch. Solid blue and red lines indicate the use of exclusively correct trials for training and testing data. Dashed blue and red lines indicate the use of both success and error trials for training and testing data. Shaded ribbons represent 95% confidence intervals. Significant differences between early and late training decoding are marked overhead in green and cyan (two tailed permutation test, object match-nonmatch and choose-match tasks, $p < 0.05$, spatial match-nonmatch task, $p < 0.05$).



Supplementary Figure S7: Expanded dynamics of populations responses in PFC. The average population response for a given condition and time is represented as a point in state space for A. the spatial match-nonmatch task, B. the object match-nonmatch task, and C. the object choose-match task. Responses from correct trials only are shown from 1000 ms before the start of the trial (inter-trial interval) to the end of the trial using a 400 ms sliding window, with a 100 ms step size (with overlap), and are projected into the two-dimensional subspace capturing

the variance due to the combination of the presented stimulus (i.e. match or nonmatch in the spatial and object match-nonmatch tasks, and circle or triangle stimulus presentation in the object choose-match task) and the saccade direction of the rewarded choice target (up or down) task conditions for early (left) and late (right) training phases. Units are arbitrary. Each color indicates a different combination of the two pairs of possible task conditions. Starting points for each of the resulting neural trajectories is indicated by text and arrows in the same color.

Chapter 4: General Discussion

Overview: What are the mechanisms of training in working memory?

Working memory (WM)—referring to the ability to maintain and manipulate information in the conscious mind over a timescale of seconds—is a fundamental concept in the field of neuroscience that has been developed over the course of generations (Baddeley 1992). This concept was originally created through the distinctions between short-term memory and long-term memory and thus progressed through an extensive history of research and revisions from the classic Atkinson-Shiffrin multi-store model to the Baddeley model. (Peterson and Peterson 1959, Atkinson and Shiffrin 1968, Baddeley 1997, Hebb 2002, Baddeley 2003). Early lesion studies first revealed the key role of the prefrontal cortex (PFC) in WM, directly leading to the examination of various neural mechanisms at the single-neuron level through electrophysiological investigations in non-human primates (Fuster and Alexander 1971). The question of training and the acquisition of new rules is of particular interest in this regard, as these have been linked to improvement in a variety of ailments, such as stroke, aging, and PTSH, among others, which may indicate the possibility of further malleability for WM, even in healthy adults (Klingberg, Forssberg et al. 2002, Westerberg, Jacobaeus et al. 2007, Saunders, Downham et al. 2015, Vermeij, Kessels et al. 2016). Such studies provided a revolutionary insight that overturned the previous belief that WM was both constant and predetermined by the limits of the individual's prefrontal and parietal cortices (Li, Schmiedek et al. 2008, Brehmer, Westerberg et al. 2012). However, the underlying mechanisms in the PFC that may be responsible for

training effects remain under debate, and this question serves as the foundation upon that this current dissertation seeks to elaborate.

To provide a framework for our investigation, we attempted to upgrade the prior models of WM by proposing a new model that supports the relative independence of different WM modalities through the activation of different ensembles of neurons (Jaffe and Constantinidis 2021). This representation of different parameters in response properties is known as neural selectivity, and has been demonstrated a wide range of imaging and electrophysiological studies that were conducted across a variety of WM tasks (Fuster and Alexander 1971, Funahashi, Bruce et al. 1989, Jansma, Ramsey et al. 2001, Olesen, Westerberg et al. 2004, Badre and D'Esposito 2009, Qi and Constantinidis 2013). However, our proposed independence of WM modalities is tempered by neurons with nonlinear interaction of stimulus properties, or “mixed selectivity” that has been observed in previous experiments (Dang, Jaffe et al. 2021).

In the specific context of training effects, selectivity—especially various types of mixed selectivity—has been proven to play an important role based upon its high degree of plasticity, undergoing significant increases in the PFC over the course of training in WM tasks (Riley, Qi et al. 2017, Riley, Qi et al. 2018, Tang, Riley et al. 2022). This would seem to serve as an underlying mechanism to support the increased functional connectivity and widespread reorganization that a variety of fMRI studies have observed across prefrontal neural networks with training, thus providing a deeper explanation for how training strengthens the neural

representations that underlie short-term information maintenance and manipulation (Takeuchi, Taki et al. 2013, Gordon, Breeden et al. 2014). Understanding how selectivity emerges over the course of training may therefore provide the key for unlocking the prefrontal mechanisms that are ultimately responsible for corresponding behavioral improvements.

The cellular basis of neural selectivity arises from recurrent excitatory connections within neural circuits, stabilized by inhibitory interneurons and neuromodulators such as dopamine, in order to form self-sustaining loops that represent WM information based on specific patterns of activity (Constantinidis, Funahashi et al. 2018, Froudist-Walsh, Bliss et al. 2021, Kim, Chow et al. 2025). Importantly, this would seem to provide a link to the dopaminergic circuits whose the reward prediction errors that govern rule learning—the process of learning to inhibit previously rewarded actions in reversal tasks (Costa, Tran et al. 2015, Averbeck and O'Doherty 2022). This serves as the framework for flexible cognitive behavior and has been observed to involve a variety of cortical and subcortical brain areas, most notably including the PFC—which is known as the seat of WM (Qi and Constantinidis 2012, Averbeck and O'Doherty 2022, Groman, Thompson et al. 2022). However, while rule learning has been examined extensively through studies of spatial memory, there is still a significant debate on whether the same mechanisms apply to object memory as well (Tang, Riley et al. 2022). Confronting the object-spatial divide in the specific case of rule learning and identifying patterns

in training effects across different tasks will thus serve as the first steps in eventually identifying generalized prefrontal mechanisms.

In the present studies, we report nonlinear mixed selectivity that emerges even prior to training, with different roles implied for different types of WM. We also examine parallel spatial and object task rule learning to reveal new trends in training effects on both prefrontal activity and neural dynamics. Such findings are aimed to understand the basis of treatments for related disorders, such as Alzheimer's, ADHD, and traumatic brain injury (TBI), ultimately serving as a deeper glimpse into the nature of cognition itself (Morris and Baddeley 1988, Curtis and D'Esposito 2004, Martinussen, Hayden et al. 2005, Rossi, Bichot et al. 2007, Forbes, Carrick et al. 2009).

Selectivity Emerging through Training

Much of the heterogeneity in the activity of prefrontal neurons comes from how the persistent activity of an individual neuron can encode either single or multiple parameters, with a portion of PFC populations even exhibiting nonlinear mixed selectivity (NMS) for different parameters (Kobak, Brendel et al. 2016). More specifically, NMS means that a neuron's response to a combination of parameters cannot be predicted by the linear summation of their responses to single parameters (Rigotti, Barak et al. 2013, Parthasarathy, Herikstad et al. 2017). This may theoretically be used by the PFC to produce reliable WM representations using otherwise unreliable neurons (Kaufman, Benna et al. 2022).

More specifically, NMS has previously been shown to maximize computational power by adding versatility to the functional role of each neuron, and by extension, increasing the dimensionality of any resulting representations (Fusi, Miller et al. 2016, Tye, Miller et al. 2024). This is critical given that higher dimensionality of neural representations has been shown to predict improved behavior in WM tasks (Rigotti, Barak et al. 2013). However, such flexibility of decoding also presents the drawback of generalization and noise sensitivity, with relatively modest changes in firing rates corresponding to the encoding very different types of information (Kaufman, Benna et al. 2022). This has led to NMS being considered as one of the potential mechanisms through which persistent activity may be able to implement training effects in WM, ultimately serving to control the trade-off between discrimination and generalization, with increased NMS leading to increased information coding (Barak, Rigotti et al. 2013, Johnston, Palmer et al. 2020).

Studies have hypothesized that NMS emerges over the course training in various WM tasks, and the magnitude of NMS has been shown to increase as a function of time even within the span of single sessions (Dang, Li et al. 2022). This would imply NMS as a potential medium for both short- and long-term training effects. Moreover, by providing linear readouts of flexible, arbitrary combinations of variables, NMS has been suggested to serve as a key mechanism for training effects by reducing the otherwise significant metabolic cost that persistent activity would otherwise pose for maintaining information in WM (Laughlin, de Ruyter van

Steveninck et al. 1998, Rigotti, Barak et al. 2013). On the other side of the debate however, such claims have been critically disputed by studies that failed to detect NMS, even after training in WM tasks (Cavanagh, Towers et al. 2018). The relative scarcity of NMS—especially in object WM—has further undermined any claims to its central role in related training effects and has led to significant difficulties in investigating this phenomenon in an experimental context.

Some researchers have attempted to explain these apparent inconsistencies by proposing that NMS may emerge with training due to multiple types of tasks and information being combined (Johnston, Palmer et al. 2020). Other researchers have proposed that NMS in the PFC may be exclusive for spatial WM, with object WM being represented by persistent activity in lower cortices instead, emphasizing the decoding of this information from the primary visual cortex as further evidence of separation between modalities (Ester, Serences et al. 2009, Lawrence, van Mourik et al. 2018). Until now however, these possibilities have remained unconfirmed, and so the true role of NMS in WM—as well as how this may emerge through training in different task types and contexts—has remained a significant gap in research. *Chapter II of this dissertation thus aimed to resolve this debate by exploring the hypothesis that NMS may emerge as a result of training to perform both spatial and object tasks in different contexts* (Dang, Jaffe et al. 2021).

To this end, we examined NMS across various prefrontal areas using a series of tasks that relied on different modalities, training levels, and degrees of complexity. However, our findings ultimately failed to uncover substantial NMS that

could be generalized across contexts. Training effects documented in Chapter II were still able to demonstrate a modest increase in the proportion of PFC neurons with NMS exclusively for spatial WM, but these effects did not extend to object WM tasks. Furthermore, NMS was not predictive of performance, thus pointing to additional mechanisms governing with the representation of complex information about stimulus and task context in neuronal activity, which will require further research to fully investigate. Our findings thus stand in sharp contrast to prior studies that emphasized the potential importance of NMS by pointing to how the high dimensionality of NMS neurons allow simple readouts such as linear classifiers to implement a large set of input–output relations (Rigotti, Barak et al. 2013). Instead, NMS seems to be selectively engaged under specific representational demands and does not arise automatically from all forms of training, which could potentially explain why some studies have observed a role for NMS while others have not. Determining the limits of how NMS engages—or does not engage—with specific representational demands will likely serve as the next step to resolving this longstanding debate.

Our conclusions are further reenforced by the fact that even though the emergence of NMS has previously been linked to task complexity, our own experiments failed to reveal higher rates of NMS when comparing a simple object task with a more complex version, referred to as a conjunction task, which incorporated spatial elements as well. This negative result is highly informative, as it implies that complexity alone is not the key driver of NMS in the PFC. However, this

does not dismiss the possibility that NMS may be driven by specific types of complexity; for example, the read-out rule (match/nonmatch) of this conjunction task remained constant and may thus limit the development of NMS if this were driven specifically by the complexity of the task rules, rather than the complexity of the related WM load. Further research will thus be needed to assess this possibility by controlling for these factors separately.

New insights were also obtained into the underlying mechanisms that are responsible for the functional anatomy of the PFC. For example, prior studies have highlighted posterior areas to be relatively responsive for object stimuli (Badre and D'Esposito 2009, Bahlmann, Blumenfeld et al. 2015). More recent research has further oriented stimulus information selectivity along two functional axes, with a dorsal-ventral division representing spatial and object information, respectively (Riley, Qi et al. 2017, Riley, Qi et al. 2018). Our focus on the mid-dorsal region—and the subsequent lack of observed increases in object selectivity that occurred here—would thus seem to conform to established predictions. There was also a much lower incidence of NMS in the object task compared to the spatial task, despite the fact that the latter was no more complex or difficult for the monkeys to perform (Meyer, Qi et al. 2011, Riley, Qi et al. 2018). Training effects were revealed to be strongest in area 9/46 and thus suggest that different PFC subregions may be differentially engaged in the emergence of NMS, reflecting possible differences in connectivity, intrinsic network architecture or task engagement that should be further investigated in future research.

An alternative explanation for our observed absence of NMS across tasks could attempt to argue for training bringing about an increased encoding of information from NMS cells, rather than an increase in the proportion of these cells in prefrontal populations, pointing to how NMS has previously been highlighted as a potential means of increasing the efficiency of WM representations (Rigotti, Barak et al. 2013, Johnston, Palmer et al. 2020). However, our findings refuted this possibility, as we did not observe strong evidence that individual neurons with NMS perform more effectively at decoding task variables than classical or linear mixed-selectivity neurons, except in the spatial task, thus implying a lack of generalization in this mechanism. Importantly, prior studies have highlighted spatial representations to have a special status in WM, being decodable from prefrontal activity, even when irrelevant to the given task (Foster, Bsales et al. 2017). The exclusivity of our observed NMS training effects to spatial WM may therefore present a possible mechanism through which this special status may be implemented. As a result, NMS may still play a significant role in contributing to WM representations—especially in terms of spatial WM and its differentiation from object WM. However, our results firmly demonstrate that training is not a prerequisite of NMS, nor does NMS seem to be either the sole or sufficient mechanism underlying training effects across different types of tasks.

Our observed lack of increases in NMS with task complexity raised the question of whether NMS would even be necessary for performing complex tasks, and thus sparked a significant debate with prior studies that hypothesized high

important for NMS in this regard (Rigotti, Barak et al. 2013). However, subsequent studies would seem to support our side of this debate by revealing that prefrontal NMS was present in only a small percentage of neurons, which did not result in a substantial rotation of the population responses in PCA space, and was also not predictive of task performance (Dang, Li et al. 2022). This indicates a more limited role for NMS, with other mechanisms being responsible for representing complex tasks in neural activity. Likewise, while our present study failed to detect appreciable numbers of neurons with NMS, subsequent fMRI research has since been able to highlight NMS neurons clustering at a much larger scale than previously observed on the single neuron level, and may thus offer new brain regions—as well as alternative ways to assess NMS in these regions—that may be fruitful for future investigations into this phenomenon (Taylor and Xu 2024). Finally, to better clarify the role of NMS, this has recently been correlated with improved conjunction learning, as well as subsequent advantages in performing certain types of untrained tasks (Karami and Schwiedrzik 2024). Examining a wider range of task contexts may thus be necessary to reveal the factors that are necessary for NMS to emerge, and likewise, we also must confront the possibility that NMS representations only or additionally arise in higher areas, such as the PFC, by expanding future anatomical studies beyond the limits of this area alone.

Bridging the Object-Spatial Divide

Considering how our findings in Chapter II revealed both parallels and differences in the training mechanisms of object and spatial WM, we were motivated to investigate further into the overarching question of the object-spatial divide. At least in terms of anatomy in the PFC, a quantitative difference in specialization seems to be evident between training effects in object and spatial selectivity, with the former being most prevalent in the ventrolateral PFC, while the latter is most prevalent in the dorsolateral PFC (Meyer, Qi et al. 2011, Dang, Jaffe et al. 2021). The role of persistent activity in training and rule learning has been examined in detail through prior studies of spatial WM, but object WM remains far less explored (Tang, Riley et al. 2022). Determining the extent over which such effects may transfer between modalities and into tasks that were not part of the training itself (such as everyday life), is the key to providing tangible treatments to patients in need, and is thus considered to be among the most important questions in the field of neuroscience (Cortese, Ferrin et al. 2015, Peijnenborgh, Hurks et al. 2016, Al-Saad, Al-Jabri et al. 2021).

There are a variety of reasons that may account for the current disparity between object and spatial WM research in the PFC. For example, object stimuli have a virtually infinite range of shapes, and even the stimulus sets that were applied in our own object memory tasks in this dissertation would represent only a fraction of the available possibilities. The relatively lower proportion of neurons that respond to object stimuli compared to spatial stimuli may thus reflect the increased specificity that would be needed for the former (Dang, Jaffe et al. 2021).

This means that object mechanisms may be much more difficult to identify through changes in activity alone. In more practical terms too, the virtually infinite range of object stimuli would imply that controlling for all possible confounds in object memory tasks—such as the relative complexity of different shape stimuli—can often be much more challenging than in spatial memory tasks. The result is that many paradigms inevitably conflate binding (i.e. whether the aspects of a given object, such as shape and color, are stored as integrated objects vs. independent attributes) and identity encoding, making it difficult to determine whether observed deficits reflect feature loss, binding failure, or attentional lapses, among other possibilities (Xu 2002, Allen, Hitch et al. 2012).

Moreover, there has been debate on whether the mechanism of prefrontal persistent activity even applies to object memory, with some researchers suggesting a merely supervisory role for the PFC, wherein spatial locations would be highlighted by persistent activity, while object information would be maintained in the sensory cortices instead (Nystrom, Braver et al. 2000, Ester, Serences et al. 2009). Such supervisory models of the PFC are supported by the decoding of object WM information from the primary visual cortex, and would imply that seemingly selective responses to (single) stimuli may represent not the properties of stimuli themselves, but rather categorical judgements about trained task associations (Ester, Serences et al. 2009, Lawrence, van Mourik et al. 2018). By contrast however, there is strong evidence of prefrontal persistent activity dynamically incorporating multiple types of both spatial and object information, thus refuting

claims that this mechanism may be an exclusive to spatial WM alone (Amengual and Ben Hamed 2021, Curtis and Sprague 2021). Further research on the mechanisms of object WM—and how these may overlap with the mechanisms of spatial WM—are thus needed for the advancement of our understanding.

Chapter III of this dissertation thus aimed to establish a new foundation for identifying generalized mechanisms of rule learning by exploring the hypothesis that rule learning across spatial and feature memory tasks may be mediated by a variety of activity-based mechanisms that parallel the related behavioral improvements. Any trends in training effects across tasks would imply the possibility of more specific and generalized mechanisms that would govern rule learning across different contexts, thus providing new targets for investigation in subsequent research. We were thus led to examine changes in prefrontal activity in monkeys conducting an object memory task that was specifically designed to parallel a previously published spatial memory task (Tang, Riley et al. 2022). An additional object memory task was then applied to demonstrate the same rule being learned in different contexts. Ultimately, our hypothesis was supported by a variety of rule learning mechanisms that followed the same trends across all three of our experimental tasks, displaying increases or decreases of both firing rate and decoding, increases in firing rate variance that was unexplained by sensory stimuli and motor actions, and the increased separation of population response trajectories in state space.

Training effects documented in Chapter III provide direct evidence to further refute the claims of studies that have argued against the role of prefrontal activity in rule learning by proposing alternative mechanisms instead, including rhythmic-burst models that posit prefrontal gamma band rhythmicity as the neural correlate of memory, as well as activity silent models, that do not depend on discharge rates modulation (Stokes 2015, Lundqvist, Rose et al. 2016, Buschman and Miller 2022). Likewise, the modulation of synaptic resources, such as the availability of neurotransmitter and presynaptic calcium, is considered to be the another alternative mechanism for implementing WM representations and related training effects (Mongillo, Barak et al. 2008, Masse, Rosen et al. 2020). As with activity-based models, there has been an ongoing debate regarding if and how these alternative mechanisms may apply in different contexts, and our present findings are not meant to necessarily refute the possibility of these having their own applications in rule learning (Oberauer and Awh 2022). In either case however, we demonstrate that activity-based mechanisms seem to be utilized across both spatial and object modalities, thus providing evidence for prefrontal activity serving significant role in rule learning in addition to any other potential mechanisms that may be present.

More specifically, the progression of rule learning seemed to move firing rate in different directions for different monkeys/tasks, both increasing and decreasing. This result mirrors the conflicting fMRI studies that have suggested the former to indicate strengthened neural representations that underlie short-term information

maintenance and manipulation (Garavan, Kelley et al. 2000, Hempel, Giesel et al. 2004, Mendoza-Halliday and Martinez-Trujillo 2017), just as the latter has been suggested to indicate improvements in efficiency (Schneiders, Opitz et al. 2011, Takeuchi, Taki et al. 2013). Our observation of both effects across tasks—increases and decreases—would therefore imply that this debate is not mutually exclusive to one side or the other, and instead, either effect may be responsible for mediating training, as determined by the individual monkey/task.

The other training effects that we observed—such as the increased variance of firing rate that was unexplained by sensory stimuli and motor actions, and increased separation of population response trajectories in state space—provide an additional perspective as well, since these findings can be directly correlated to training-induced improvements in behavioral task performance. More specifically, our observations of increased unexplained variance in firing rate elaborate upon the prior finding that fully trained monkeys display increased unexplained variance during the change between conditions in a reversal learning task, demonstrating that these increases seem to be implemented as a direct result of training (Bartolo and Averbeck 2020). Likewise, our population dynamic findings seem to build upon prior observations of how, in real life, animals most likely learn a reversal learning task via reinforcement learning based on reward outcomes (Kim, Chow et al. 2025). Intervening behaviors—such as introducing an intruder stimulus or accumulating reward across trials—are known to produce temporary deviations in neural trajectories (Sylwestrak, Jo et al. 2022, Nair, Karigo et al. 2023). Our own reversal-

learning tasks thus allowed this phenomenon to be examined in the context of the object-spatial divide, using the intervention of the reversal between matching stimulus presentations, and then comparing neural trajectories over the course of rule learning to reveal an increased separation of population response trajectories in state space over the course of training.

Together, our findings do not refute the totality of all separations between spatial and object mechanisms in the PFC, and instead, seem to confirm that there are at least limited segregations of spatial and object rule learning (Hoshi, Shima et al. 1998, Ninokura, Mushiake et al. 2004). These segregations would expand even beyond the long-established dorsal-ventral divide in the functional architecture of the PFC (Adcock, Constable et al. 2000, Volle, Kinkingnehun et al. 2008). Yet even so, we were also able to demonstrate that this divide is not absolute, and there exists a degree of flexibility through which individual neurons may integrate different types of information that are distributed throughout the PFC. This provides strong support for our Jaffe-Constantinidis model of WM (Chapter 1, Fig. 1C), and reinforces prior claims of neuronal responses being shaped by cognitive demands imposed by the task rather than being subject exclusively to selectivity for specific domains (Rao, Harrington et al. 1997, Jaffe and Constantinidis 2021).

Our present findings may have established specific trends in training effects that have been able to bridge the object-spatial divide, but additional studies will still be needed to determine whether such trends are due to generalized or task-specific mechanisms. For example, our current training paradigm would be unable

to fully assess the potential influence of phenomena such as match suppression, where activity elicited after repeated presentation of the same stimulus is reduced, which may be informative regarding the matching status of a given stimulus with a prior stimulus (Grill-Spector, Henson et al. 2006). Likewise, the spatial and object stimulus sets that were used in our current study represent only a minimal fraction of what is possible. These were also familiar to our animals, by virtue of exposure during our pre-training paradigms, and thus posed a high potential for interference (Stern, Sherman et al. 2001). Future studies that examine a wider range of both familiar and novel stimuli through additional rule learning paradigms will therefore be necessary to overcome these limitations and ultimately reveal the extent to which our results may be generalized across all forms of rule learning.

Conclusion and future directions

In the present studies, we have described the emergence of various types of selectivity before and following training in a working memory task, with a particular focus on NMS—which has been hypothesized as a signature for high-dimensional neural representations. Surprisingly however, our findings indicate a more limited role for NMS in WM training than previously hypothesized, and only spatial tasks revealed significant increases over the course of training. NMS could also appear even without training, but was not ubiquitous, implying different roles for different types of WM. Moreover, NMS was not predictive of either task complexity or performance. Other factors were thus implied to be responsible for WM

representation instead, and so our findings did not support a causal relationship between dimensionality and task success.

Yet additional questions regarding the training process and the exact timeframe of training effects in the PFC remain. The results of our study from Chapter II could capture only two contexts in our examined tasks: prior to training and after completion of training. Our experiments did not determine the rate at which these changes occur in the training process. The changes observed in firing rate, selectivity, and information may occur rapidly as the rule is acquired or may occur gradually throughout the training process. Further studies were therefore necessary to examine the changes that daily training induces on PFC activity using recordings from chronically implanted electrodes.

Thus, to begin confronting this gap and to build upon our first foray into the field of prefrontal training effects, our next study examined how prefrontal activity predicts rule learning across parallel spatial and object memory tasks. This led to the identification of a variety of trends in training effects that could be observed across tasks, including increases in both unexplained variance and the separation of population response trajectories in state space that could linearly correlated to behavioral improvements, as well as increases or decreases of both firing rate and decoding. Such findings provide potential targets that should be investigated as specific and generalized prefrontal mechanisms of rule learning.

Given that our studies only examined training effects through limited and familiar stimulus sets, additional research will be needed to determine whether our

findings may be applied to other types of stimuli, especially novel objects/locations. Our experimental tasks were also limited, with both studies relying on variations of the classic match-nonmatch task. Additional training paradigms will thus be needed to determine how our findings generalize to other contexts. Moreover, while the study from Chapter III included an object choose-match task to examine the acquisition of the same rule from an additional point of view, subsequent studies must examine the acquisition of other rules as well, to determine if and how our observed prefrontal mechanisms may be generalized across all forms of rule learning. Finally, none of our experiments sampled from the Frontopolar region—an area linked to the acquisition of new rules—and so investigations into possible anatomical differences between anterior and posterior frontal neurons will be necessary for understanding how training affects the PFC as a whole (Boschin, Piekema et al. 2015). The more we can improve our understanding of the neurophysiological basis of rule learning, the more we may be able to optimize cognitive training and related therapies, ultimately providing new therapies for related deficits and improving human capabilities in both clinical and healthy populations alike.

REFERENCES

Adcock, R. A., R. T. Constable, J. C. Gore and P. S. Goldman-Rakic (2000). "Functional neuroanatomy of executive processes involved in dual-task performance." Proc Natl Acad Sci U S A **97**(7): 3567–3572.

Al-Saad, M. S. H., B. Al-Jabri and A. F. Almarzouki (2021). "A Review of Working Memory Training in the Management of Attention Deficit Hyperactivity Disorder." Front Behav Neurosci **15**: 686873.

Albers, A. M., P. Kok, I. Toni, H. C. Dijkerman and F. P. de Lange (2013). "Shared representations for working memory and mental imagery in early visual cortex." Curr Biol **23**(15): 1427–1431.

Alexander, G. E., M. R. DeLong and P. L. Strick (1986). "Parallel organization of functionally segregated circuits linking basal ganglia and cortex." Annu Rev Neurosci **9**: 357–381.

Allen, R. J., G. J. Hitch, J. Mate and A. D. Baddeley (2012). "Feature binding and attention in working memory: a resolution of previous contradictory findings." Q J Exp Psychol (Hove) **65**(12): 2369–2383.

Alloway, T. P., S. E. Gathercole and S. J. Pickering (2006). "Verbal and visuospatial short-term and working memory in children: are they separable?" Child Dev **77**(6): 1698–1716.

Amengual, J. L. and S. Ben Hamed (2021). "Revisiting Persistent Neuronal Activity During Covert Spatial Attention." Front Neural Circuits **15**: 679796.

Anguera, J. A., J. Boccanfuso, J. L. Rintoul, O. Al-Hashimi, F. Faraji, J. Janowich, E. Kong, Y. Larraburo, C. Rolle, E. Johnston and A. Gazzaley (2013). "Video game training enhances cognitive control in older adults." Nature **501**(7465): 97–101.

Arce, E., R. Balice-Gordon, S. Duvvuri, M. Naylor, Z. Xie, B. Harel, R. Kozak, D. L. Gray and N. DeMartinis (2019). "A novel approach to evaluate the pharmacodynamics of a selective dopamine D1/D5 receptor partial agonist (PF-06412562) in patients with stable schizophrenia." J Psychopharmacol **33**(10): 1237–1247.

Arnsten, A. F. (2013). "The neurobiology of thought: the groundbreaking discoveries of Patricia Goldman-Rakic 1937-2003." Cereb Cortex **23**(10): 2269–2281.

Asaad, W. F., G. Rainer and E. K. Miller (1998). "Neural activity in the primate prefrontal cortex during associative learning." Neuron **21**(6): 1399–1407.

Asaad, W. F., G. Rainer and E. K. Miller (1998). "Neural activity in the primate prefrontal cortex during associative learning." Neuron **21**(6): 1399–1407.

Asaad, W. F., G. Rainer and E. K. Miller (2000). "Task-specific neural activity in the primate prefrontal cortex." J Neurophysiol **84**(1): 451–459.

Atkinson, A. L., E. D. J. Berry, A. H. Waterman, A. D. Baddeley, G. J. Hitch and R. J. Allen (2018). "Are there multiple ways to direct attention in working memory?" Ann N Y Acad Sci **1424**(1): 115–126.

Atkinson, R. C. and R. M. Shiffrin (1968). Human memory: A proposed system and its control processes. The psychology of learning and motivation. K. W. Spence and J. T. Spence. London, Academic Press. **2:** 89–195.

Averbeck, B. and J. P. O'Doherty (2022). "Reinforcement-learning in fronto-striatal circuits." Neuropsychopharmacology **47**(1): 147–162.

Averbeck, B. B., M. V. Chafee, D. A. Crowe and A. P. Georgopoulos (2002). "Parallel processing of serial movements in prefrontal cortex." Proc Natl Acad Sci U S A **99**(20): 13172–13177.

Averbeck, B. B., M. V. Chafee, D. A. Crowe and A. P. Georgopoulos (2003). "Neural activity in prefrontal cortex during copying geometrical shapesI. Single cells encode shape, sequence, and metric parameters." Exp Brain Res **150**(2): 127–141.

Awh, E., E. K. Vogel and S. H. Oh (2006). "Interactions between attention and working memory." Neuroscience **139**(1): 201–208.

Backman, L., L. Nyberg, A. Soveri, J. Johansson, M. Andersson, E. Dahlin, A. S. Neely, J. Virta, M. Laine and J. O. Rinne (2011). "Effects of working-memory training on striatal dopamine release." Science **333**(6043): 718.

Baddeley, A. (1992). "Working memory." Science **255**(5044): 556–559.

Baddeley, A. (2000). "The episodic buffer: a new component of working memory?" Trends Cogn Sci **4**(11): 417–423.

Baddeley, A. (2003). "Working memory: looking back and looking forward." Nat Rev Neurosci **4**(10): 829–839.

Baddeley, A. (2012). "Working memory: theories, models, and controversies." Annu Rev Psychol **63**: 1–29.

Baddeley, A. D. (1997). Human memory: Theory and practice, Psychology Press.

Baddeley, A. D. and J. Andrade (2000). "Working memory and the vividness of imagery." J Exp Psychol Gen **129**(1): 126–145.

Badre, D. and M. D'Esposito (2009). "Is the rostro-caudal axis of the frontal lobe hierarchical?" Nat Rev Neurosci **10**(9): 659–669.

Bae, G. Y. and S. J. Luck (2018). "Dissociable Decoding of Spatial Attention and Working Memory from EEG Oscillations and Sustained Potentials." J Neurosci **38**(2): 409–422.

Bae, G. Y. and S. J. Luck (2019). "Decoding motion direction using the topography of sustained ERPs and alpha oscillations." Neuroimage **184**: 242–255.

Baeg, E. H., Y. B. Kim, K. Huh, I. Mook-Jung, H. T. Kim and M. W. Jung (2003). "Dynamics of population code for working memory in the prefrontal cortex." Neuron **40**(1): 177–188.

Bahlmann, J., R. S. Blumenfeld and M. D'Esposito (2015). "The Rostro-Caudal Axis of Frontal Cortex Is Sensitive to the Domain of Stimulus Information." Cereb Cortex **25**(7): 1815–1826.

Balakhonov, D. and J. Rose (2017). "Crows Rival Monkeys in Cognitive Capacity." Sci Rep **7**(1): 8809.

Balan, P. F. and V. P. Ferrera (2003). "Effects of spontaneous eye movements on spatial memory in macaque periarcuate cortex." J Neurosci **23**(36): 11392–11401.

Barak, O., M. Rigotti and S. Fusi (2013). "The sparseness of mixed selectivity neurons controls the generalization-discrimination trade-off." J Neurosci **33**(9): 3844–3856.

Barbas, H. (2015). "General cortical and special prefrontal connections: principles from structure to function." Annu Rev Neurosci **38**: 269–289.

Barbosa, J., H. Stein, R. L. Martinez, A. Galan-Gadea, S. Li, J. Dalmau, K. C. S. Adam, J. Valls-Sole, C. Constantinidis and A. Compte (2020). "Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory." Nat Neurosci **23**(8): 1016–1024.

Barraclough, D. J., M. L. Conroy and D. Lee (2004). "Prefrontal cortex and decision making in a mixed-strategy game." Nat Neurosci **7**(4): 404–410.

Bartolo, R. and B. B. Averbeck (2020). "Prefrontal Cortex Predicts State Switches during Reversal Learning." Neuron **106**(6): 1044–1054 e1044.

Bartus, R. T. (2000). "On neurodegenerative diseases, models, and treatment strategies: lessons learned and lessons forgotten a generation following the cholinergic hypothesis." Exp Neurol **163**(2): 495–529.

Bastos, A. M., R. Loonis, S. Kornblith, M. Lundqvist and E. K. Miller (2018). "Laminar recordings in frontal cortex suggest distinct layers for maintenance and control of working memory." Proc Natl Acad Sci U S A **115**(5): 1117–1122.

Bathelt, J., S. E. Gathercole, A. Johnson and D. E. Astle (2018). "Differences in brain morphology and working memory capacity across childhood." Dev Sci **21**(3): e12579.

Bauer, M., R. Oostenveld, M. Peeters and P. Fries (2006). "Tactile spatial attention enhances gamma-band activity in somatosensory cortex and reduces low-frequency activity in parieto-occipital areas." J Neurosci **26**(2): 490–501.

Bays, P. M. and M. Husain (2008). "Dynamic shifts of limited working memory resources in human vision." Science **321**(5890): 851–854.

Berdyyeva, T. K. and C. R. Olson (2010). "Rank signals in four areas of macaque frontal cortex during selection of actions and objects in serial order." J Neurophysiol **104**(1): 141–159.

Bergmann, J., E. Genc, A. Kohler, W. Singer and J. Pearson (2014). "Neural Anatomy of Primary Visual Cortex Limits Visual Working Memory." Cereb Cortex.

Bherer, L., A. F. Kramer, M. S. Peterson, S. Colcombe, K. Erickson and E. Becic (2008). "Transfer effects in task-set cost and dual-task cost after dual-task training in older and younger adults: further evidence for cognitive plasticity in attentional control in late adulthood." Exp Aging Res **34**(3): 188–219.

Bichot, N. P., M. T. Heard, E. M. DeGennaro and R. Desimone (2015). "A Source for Feature-Based Attention in the Prefrontal Cortex." Neuron **88**(4): 832–844.

Bigelow, J., B. Rossi and A. Poremba (2014). "Neural correlates of short-term memory in primate auditory cortex." Front Neurosci **8**: 250.

Blake, R., N. J. Cepeda and E. Hiris (1997). "Memory for visual motion." J Exp Psychol Hum Percept Perform **23**(2): 353–369.

Bliss, D. P. and M. D'Esposito (2017). "Synaptic augmentation in a cortical circuit model reproduces serial dependence in visual working memory." PLoS One **12**(12): e0188927.

Bliss, D. P., J. J. Sun and M. D'Esposito (2017). "Serial dependence is absent at the time of perception but increases in visual working memory." Sci Rep **7**(1): 14739.

Bodner, M., J. Kroger and J. M. Fuster (1996). "Auditory memory cells in dorsolateral prefrontal cortex." Neuroreport **7**(12): 1905–1908.

Bolkan, S. S., J. M. Stujenske, S. Parnaudeau, T. J. Spellman, C. Rauffenbart, A. I. Abbas, A. Z. Harris, J. A. Gordon and C. Kellendonk (2017). "Thalamic projections sustain prefrontal activity during working memory maintenance." Nat Neurosci **20**(7): 987–996.

Borra, E., C. G. Ferroni, M. Gerbella, V. Giorgetti, C. Mangiaracina, S. Rozzi and G. Luppino (2017). "Rostro-caudal Connectional Heterogeneity of the Dorsal Part of the Macaque Prefrontal Area 46." Cereb Cortex.

Boschin, E. A., C. Piekema and M. J. Buckley (2015). "Essential functions of primate frontopolar cortex in cognition." Proc Natl Acad Sci U S A **112**(9): E1020–1027.

Brehmer, Y., H. Westerberg and L. Backman (2012). "Working-memory training in younger and older adults: training gains, transfer, and maintenance." Front Hum Neurosci **6**: 63.

Brincat, S. L. and E. K. Miller (2015). "Frequency-specific hippocampal-prefrontal interactions during associative learning." Nat Neurosci **18**(4): 576–581.

Brody, C. D., A. Hernandez, A. Zainos and R. Romo (2003). "Timing and Neural Encoding of Somatosensory Parametric Working Memory in Macaque Prefrontal Cortex." Cereb. Cortex **13**(11): 1196–1207.

Brooks, D. N. (1975). "Long and short term memory in head injured patients." Cortex **11**(4): 329–340.

Buckley, M. J., F. A. Mansouri, H. Hoda, M. Mahboubi, P. G. Browning, S. C. Kwok, A. Phillips and K. Tanaka (2009). "Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions." Science **325**(5936): 52–58.

Bunge, S. A., N. M. Dudukovic, M. E. Thomason, C. J. Vaidya and J. D. Gabrieli (2002). "Immature frontal lobe contributions to cognitive control in children: evidence from fMRI." Neuron **33**(2): 301–311.

Bunge, S. A., K. N. Ochsner, J. E. Desmond, G. H. Glover and J. D. Gabrieli (2001). "Prefrontal regions involved in keeping information in and out of mind." Brain **124**(Pt 10): 2074–2086.

Buonomano, D. V. and W. Maass (2009). "State-dependent computations: spatiotemporal processing in cortical networks." Nat Rev Neurosci **10**(2): 113–125.

Burgess, G. C., B. E. Depue, L. Ruzic, E. G. Willcutt, Y. P. Du and M. T. Banich (2010). "Attentional control activation relates to working memory in attention-deficit/hyperactivity disorder." Biol Psychiatry **67**(7): 632–640.

Burgund, E. D., H. M. Lugar, F. M. Miezin, B. L. Schlaggar and S. E. Petersen (2006). "The development of sustained and transient neural activity." Neuroimage **29**(3): 812–821.

Burt, J. B., M. Demirtas, W. J. Eckner, N. M. Navejar, J. L. Ji, W. J. Martin, A. Bernacchia, A. Anticevic and J. D. Murray (2018). "Hierarchy of transcriptomic specialization across human cortex captured by structural neuroimaging topography." Nat Neurosci **21**(9): 1251–1259.

Buschman, T. J., E. L. Denovellis, C. Diogo, D. Bullock and E. K. Miller (2012). "Synchronous oscillatory neural ensembles for rules in the prefrontal cortex." Neuron **76**(4): 838–846.

Buschman, T. J. and E. K. Miller (2022). "Working Memory Is Complex and Dynamic, Like Your Thoughts." J Cogn Neurosci **35**(1): 17–23.

Buschman, T. J., M. Siegel, J. E. Roy and E. K. Miller (2011). "Neural substrates of cognitive capacity limitations." Proc Natl Acad Sci U S A **108**(27): 11252–11255.

Butler, A. B. and R. M. Cotterill (2006). "Mammalian and avian neuroanatomy and the question of consciousness in birds." Biol Bull **211**(2): 106–127.

Buzsaki, G. (2010). "Neural syntax: cell assemblies, synapsembles, and readers." Neuron **68**(3): 362–385.

Buzsaki, G. (2020). "The Brain-Cognitive Behavior Problem: A Retrospective." eNeuro **7**(4).

Calvert, G. A., E. T. Bullmore, M. J. Brammer, R. Campbell, S. C. Williams, P. K. McGuire, P. W. Woodruff, S. D. Iversen and A. S. David (1997). "Activation of auditory cortex during silent lipreading." Science **276**(5312): 593–596.

Carmichael, S. T., M. C. Clugnet and J. L. Price (1994). "Central olfactory connections in the macaque monkey." J Comp Neurol **346**(3): 403–434.

Cavada, C., T. Company, J. Tejedor, R. J. Cruz-Rizzolo and F. Reinoso-Suarez (2000). "The anatomical connections of the macaque monkey orbitofrontal cortex. A review." Cereb Cortex **10**(3): 220–242.

Cavada, C. and P. S. Goldman-Rakic (1989). "Posterior parietal cortex in rhesus monkey: II. Evidence for segregated corticocortical networks linking sensory and limbic areas with the frontal lobe." Journal of Comparative Neurology **287**(4): 422–445.

Cavanagh, S. E., J. P. Towers, J. D. Wallis, L. T. Hunt and S. W. Kennerley (2018). "Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex." Nat Commun **9**(1): 3498.

Chafee, M. V. and P. S. Goldman-Rakic (1998). "Matching patterns of activity in primate prefrontal area 8a and parietal area 7ip neurons during a spatial working memory task." Journal of Neurophysiology **79**(6): 2919–2940.

Chafee, M. V. and P. S. Goldman-Rakic (2000). "Inactivation of Parietal and Prefrontal Cortex Reveals Interdependence of Neural Activity During Memory Guided-Saccades." Journal of Neurophysiology **83**: 1550–1566.

Chafee, M. V. and P. S. Goldman-Rakic (2000). "Inactivation of parietal and prefrontal cortex reveals interdependence of neural activity during memory-guided saccades." J Neurophysiol **83**(3): 1550–1566.

Chang, L., P. Bao and D. Y. Tsao (2017). "The representation of colored objects in macaque color patches." Nat Commun **8**(1): 2064.

Chang, M. H., K. M. Armstrong and T. Moore (2012). "Dissociation of response variability from firing rate effects in frontal eye field neurons during visual stimulation, working memory, and attention." J Neurosci **32**(6): 2204–2216.

Chase, W. and K. A. Erickson (1982). "Skill and Working Memory." The Psychology of Learning and Modivation **16**.

Chelazzi, L., E. K. Miller, J. Duncan and R. Desimone (1993). "A neural basis for visual search in inferior temporal cortex." Nature **363**(6427): 345–347.

Chen, G., P. Greengard and Z. Yan (2004). "Potentiation of NMDA receptor currents by dopamine D1 receptors in prefrontal cortex." Proc Natl Acad Sci U S A **101**(8): 2596–2600.

Chen, T. and M. Naveh-Benjamin (2012). "Assessing the associative deficit of older adults in long-term and short-term/working memory." Psychol Aging **27**(3): 666–682.

Christophel, T. B., P. C. Klink, B. Spitzer, P. R. Roelfsema and J. D. Haynes (2017). "The Distributed Nature of Working Memory." Trends Cogn Sci **21**(2): 111–124.

Chugani, H. T., M. E. Phelps and J. C. Mazziotta (1987). "Positron emission tomography study of human brain functional development." Ann Neurol **22**(4): 487–497.

Churchland, M. M., B. M. Yu, J. P. Cunningham, L. P. Sugrue, M. R. Cohen, G. S. Corrado, W. T. Newsome, A. M. Clark, P. Hosseini, B. B. Scott, D. C. Bradley, M. A. Smith, A. Kohn, J. A. Movshon, K. M. Armstrong, T. Moore, S. W. Chang, L. H. Snyder, S. G. Lisberger, N. J. Priebe, I. M. Finn, D. Ferster, S. I. Ryu, G. Santhanam, M. Sahani and K. V. Shenoy (2010). "Stimulus onset quenches neural variability: a widespread cortical phenomenon." Nat Neurosci **13**(3): 369–378.

Clark, K. L., B. Noudoost and T. Moore (2012). "Persistent spatial information in the frontal eye field during object-based short-term memory." J Neurosci **32**(32): 10907–10914.

Clark, K. L., B. Noudoost and T. Moore (2014). "Persistent spatial information in the FEF during object-based short-term memory does not contribute to task performance." J Cogn Neurosci **26**(6): 1292–1299.

Cohen, M. A., D. C. Dennett and N. Kanwisher (2016). "What is the Bandwidth of Perceptual Experience?" Trends Cogn Sci **20**(5): 324–335.

Cohen, Y. E., M. D. Hauser and B. E. Russ (2006). "Spontaneous processing of abstract categorical information in the ventrolateral prefrontal cortex." Biol Lett **2**(2): 261–265.

Colflesh, G. J. and A. R. Conway (2007). "Individual differences in working memory capacity and divided attention in dichotic listening." Psychon Bull Rev **14**(4): 699–703.

Compte, A., N. Brunel, P. S. Goldman-Rakic and X. J. Wang (2000). "Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model." Cerebral Cortex **10**(9): 910–923.

Compte, A., C. Constantinidis, J. Tegner, S. Raghavachari, M. V. Chafee, P. S. Goldman-Rakic and X. J. Wang (2003). "Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task." J Neurophysiol **28**: 3441–3454.

Constantinidis, C. (2016). "A Code for Cross-Modal Working Memory." Neuron **89**(1): 3–5.

Constantinidis, C., M. N. Franowicz and P. S. Goldman-Rakic (2001). "Coding specificity in cortical microcircuits: a multiple electrode analysis of primate prefrontal cortex." Journal of Neuroscience **21**(10): 3646–3655.

Constantinidis, C., M. N. Franowicz and P. S. Goldman-Rakic (2001). "The sensory nature of mnemonic representation in the primate prefrontal cortex." Nature Neuroscience **4**(3): 311–316.

Constantinidis, C., S. Funahashi, D. Lee, J. D. Murray, X. L. Qi, M. Wang and A. F. T. Arnsten (2018). "Persistent Spiking Activity Underlies Working Memory." J Neurosci **38**(32): 7020–7028.

Constantinidis, C. and P. S. Goldman-Rakic (2002). "Correlated discharges among putative pyramidal neurons and interneurons in the primate prefrontal cortex." Journal of Neurophysiology **88**: 3487–3497.

Constantinidis, C. and T. Klingberg (2016). "The neuroscience of working memory capacity and training." Nat Rev Neurosci **17**(7): 438–449.

Constantinidis, C. and E. Procyk (2004). "The primate working memory networks." Cogn Affect Behav Neurosci **4**(4): 444–465.

Constantinidis, C. and X. L. Qi (2018). "Representation of Spatial and Feature Information in the Monkey Dorsal and Ventral Prefrontal Cortex." Front Integr Neurosci **12**: 31.

Constantinidis, C. and M. A. Steinmetz (1996). "Neuronal activity in posterior parietal area 7a during the delay periods of a spatial memory task." Journal of Neurophysiology **76**(2): 1352–1355.

Constantinidis, C. and M. A. Steinmetz (2001). "Neuronal responses in area 7a to multiple stimulus displays: I. Neurons encode the location of the salient stimulus." Cerebral Cortex **11**: 581–591.

Constantinidis, C. and X. J. Wang (2004). "A neural circuit basis for spatial working memory." Neuroscientist **10**(6): 553–565.

Constantinidis, C., G. V. Williams and P. S. Goldman-Rakic (2002). "A role for inhibition in shaping the temporal flow of information in prefrontal cortex." Nature Neuroscience **5**(2): 175–180.

Cortese, S., M. Ferrin, D. Brandeis, J. Buitelaar, D. Daley, R. W. Dittmann, M. Holtmann, P. Santosh, J. Stevenson, A. Stringaris, A. Zuddas, E. J. Sonuga-Barke and A. G. G. European (2015). "Cognitive training for attention-deficit/hyperactivity disorder: meta-analysis of clinical and neuropsychological outcomes from randomized controlled trials." J Am Acad Child Adolesc Psychiatry **54**(3): 164–174.

Costa, V. D., V. L. Tran, J. Turchi and B. B. Averbeck (2015). "Reversal learning and dopamine: a bayesian perspective." J Neurosci **35**(6): 2407–2416.

Courtney, S. M., L. G. Ungerleider, K. Keil and J. V. Haxby (1997). "Transient and sustained activity in a distributed neural system for human working memory." Nature **386**(6625): 608–611.

Cowan, N. (1988). "Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system." Psychol Bull. **104**(2): 163–191.

Cowan, N. (2001). "The magical number 4 in short-term memory: a reconsideration of mental storage capacity." Behav Brain Sci **24**(1): 87–114; discussion 114–185.

Cowan, N. (2005). "Working Memory Capacity." Psychol. Press 260.

Cowan, N. and A. M. AuBuchon (2008). "Short-term memory loss over time without retroactive stimulus interference." Psychon Bull Rev **15**(1): 230–235.

Crone, E. A., C. Wendelken, S. Donohue, L. van Leijenhorst and S. A. Bunge (2006). "Neurocognitive development of the ability to manipulate information in working memory." Proc Natl Acad Sci U S A **103**(24): 9315–9320.

Crowe, D. A., B. B. Averbeck and M. V. Chafee (2008). "Neural ensemble decoding reveals a correlate of viewer- to object-centered spatial transformation in monkey parietal cortex." J Neurosci **28**(20): 5218–5228.

Crowe, D. A., S. J. Goodwin, R. K. Blackman, S. Sakellaridi, S. R. Sponheim, A. W. Macdonald, 3rd and M. V. Chafee (2013). "Prefrontal neurons transmit signals to parietal neurons that reflect executive control of cognition." Nat Neurosci **16**(10): 1484–1491.

Cueva, C. J., A. Saez, E. Marcos, A. Genovesio, M. Jazayeri, R. Romo, C. D. Salzman, M. N. Shadlen and S. Fusi (2020). "Low-dimensional dynamics for working memory and time encoding." Proc Natl Acad Sci U S A **117**(37): 23021–23032.

Curtis, C. E. and M. D'Esposito (2004). "The effects of prefrontal lesions on working memory performance and theory." Cogn Affect Behav Neurosci **4**(4): 528–539.

Curtis, C. E. and T. C. Sprague (2021). "Persistent Activity During Working Memory From Front to Back." Front Neural Circuits **15**: 696060.

D'Esposito, M. (2007). "From cognitive to neural models of working memory." Philos Trans R Soc Lond B Biol Sci. **362**(1481): 761–772.

D'Esposito, M., D. Ballard, E. Zarahn and G. K. Aguirre (2000). "The role of prefrontal cortex in sensory memory and motor preparation: an event-related fMRI study." Neuroimage **11**(5 Pt 1): 400–408.

D'Esposito, M. and B. R. Postle (2015). "The cognitive neuroscience of working memory." Annu Rev Psychol **66**: 115–142.

Dahlin, E., A. S. Neely, A. Larsson, L. Backman and L. Nyberg (2008). "Transfer of learning after updating training mediated by the striatum." Science **320**(5882): 1510–1512.

Dang, W., R. J. Jaffe, X. L. Qi and C. Constantinidis (2020). "Emergence of Non-Linear Mixed Selectivity in Prefrontal Cortex after Training." BioRxiv **doi.org/10.1101/2020.08.02.233247**.

Dang, W., R. J. Jaffe, X. L. Qi and C. Constantinidis (2021). "Emergence of Nonlinear Mixed Selectivity in Prefrontal Cortex after Training." J Neurosci **41**(35): 7420–7434.

Dang, W., S. Li, S. Pu, X. L. Qi and C. Constantinidis (2022). "More Prominent Nonlinear Mixed Selectivity in the Dorsolateral Prefrontal than Posterior Parietal Cortex." eNeuro **9**(2).

Dash, P. K., A. N. Moore, N. Kobori and J. D. Runyan (2007). "Molecular activity underlying working memory." Learn Mem **14**(8): 554–563.

Defelipe, J., M. C. Gonzalez-Albo, M. R. Del Rio and G. N. Elston (1999). "Distribution and patterns of connectivity of interneurons containing calbindin, calretinin, and parvalbumin in visual areas of the occipital and temporal lobes of the macaque monkey." J Comp Neurol **412**(3): 515–526.

Desimone, R., T. D. Albright, C. G. Gross and C. Bruce (1984). "Stimulus-selective properties of inferior temporal neurons in the macaque." J Neurosci **4**(8): 2051–2062.

Desimone, R. and J. Duncan (1995). "Neural mechanisms of selective visual attention." Annu Rev Neurosci **18**: 193–222.

Deubel, H. and W. X. Schneider (1996). "Saccade target selection and object recognition: evidence for a common attentional mechanism." Vision Res **36**(12): 1827–1837.

di Pellegrino, G. and S. P. Wise (1993). "Effects of attention on visuomotor activity in the premotor and prefrontal cortex of a primate." Somatosensory & Motor Research **10**(3): 245–262.

Dias, E. C. and M. A. Segraves (1999). "Muscimol-induced inactivation of monkey frontal eye field: effects on visually and memory-guided saccades." Journal of Neurophysiology **81**(5): 2191–2214.

Diekamp, B., A. Gagliardo and O. Gunturkun (2002). "Nonspatial and subdivision-specific working memory deficits after selective lesions of the avian prefrontal cortex." J Neurosci **22**(21): 9573–9580.

Diekamp, B., T. Kalt and O. Gunturkun (2002). "Working memory neurons in pigeons." J Neurosci **22**(4): RC210.

Doherty, J. M. and R. H. Logie (2016). "Resource-sharing in multiple-component working memory." Mem Cognit **44**(8): 1157–1167.

Donahue, C. H. and D. Lee (2015). "Dynamic routing of task-relevant signals for decision making in dorsolateral prefrontal cortex." Nat Neurosci **18**(2): 295–301.

Druckmann, S. and D. B. Chklovskii (2012). "Neuronal circuits underlying persistent representations despite time varying activity." Curr Biol **22**(22): 2095–2103.

Durstewitz, D., J. K. Seamans and T. J. Sejnowski (2000). "Neurocomputational models of working memory." Nat Neurosci **3**(Suppl): 1184–1191.

Dux, P. E., M. N. Tombu, S. Harrison, B. P. Rogers, F. Tong and R. Marois (2009). "Training improves multitasking performance by increasing the speed of information processing in human prefrontal cortex." Neuron **63**(1): 127–138.

Edin, F., T. Klingberg, P. Johansson, F. McNab, J. Tegner and A. Compte (2009). "Mechanism for top-down control of working memory capacity." Proc Natl Acad Sci U S A **106**(16): 6802–6807.

Elliott, L. A. and R. J. Strawhorn (1976). "Interference in short-term memory from vocalization: aural versus visual modality differences." J Exp Psychol Hum Learn **2**(6): 705–711.

Elston, G. N. (2000). "Pyramidal cells of the frontal lobe: all the more spinous to think with." J Neurosci **20**(18): RC95.

Elston, G. N. (2003). "The pyramidal neuron in occipital, temporal and prefrontal cortex of the owl monkey (Aotus trivirgatus): regional specialization in cell structure." Eur J Neurosci **17**(6): 1313–1318.

Elston, G. N. and M. C. Gonzalez-Albo (2003). "Parvalbumin-, calbindin-, and calretinin-immunoreactive neurons in the prefrontal cortex of the owl monkey (Aotus trivirgatus): a standardized quantitative comparison with sensory and motor areas." Brain Behav Evol **62**(1): 19–30.

Emery, N. J. and N. S. Clayton (2004). "The mentality of crows: convergent evolution of intelligence in corvids and apes." Science **306**(5703): 1903–1907.

Emery, N. J. and N. S. Clayton (2005). "Evolution of the avian brain and intelligence." Curr Biol **15**(23): R946–950.

Eng, H. Y., D. Chen and Y. Jiang (2005). "Visual working memory for simple and complex visual stimuli." Psychon Bull Rev **12**(6): 1127–1133.

Erlich, J. C., M. Bialek and C. D. Brody (2011). "A cortical substrate for memory-guided orienting in the rat." Neuron **72**(2): 330–343.

Esmaeili, V. and M. E. Diamond (2019). "Neuronal Correlates of Tactile Working Memory in Prefrontal and Vibrissal Somatosensory Cortex." Cell Rep **27**(11): 3167–3181 e3165.

Ester, E. F., D. E. Anderson, J. T. Serences and E. Awh (2013). "A Neural Measure of Precision in Visual Working Memory." Journal of Cognitive Neuroscience **25**(5): 754–761.

Ester, E. F., J. T. Serences and E. Awh (2009). "Spatially global representations in human primary visual cortex during working memory maintenance." J Neurosci **29**(48): 15258–15265.

Ester, E. F., T. C. Sprague and J. T. Serences (2015). "Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory." Neuron **87**(4): 893–905.

Felleman, D. J. and D. C. Van Essen (1991). "Distributed hierarchical processing in the primate cerebral cortex." Cerebral Cortex **1**(1): 1–47.

Fiebig, F. and A. Lansner (2017). "A Spiking Working Memory Model Based on Hebbian Short-Term Potentiation." J Neurosci **37**(1): 83–96.

Finc, K., K. Bonna, X. He, D. M. Lydon-Staley, S. Kuhn, W. Duch and D. S. Bassett (2020). "Dynamic reconfiguration of functional brain networks during working memory training." Nat Commun **11**(1): 2435.

Fischer, J. and D. Whitney (2014). "Serial dependence in visual perception." Nat Neurosci **17**(5): 738–743.

Fish, K. N., B. R. Rocco and D. A. Lewis (2018). "Laminar Distribution of Subsets of GABAergic Axon Terminals in Human Prefrontal Cortex." Front Neuroanat **12**: 9.

Forbes, N. F., L. A. Carrick, A. M. McIntosh and S. M. Lawrie (2009). "Working memory in schizophrenia: a meta-analysis." Psychol Med **39**(6): 889–905.

Foster, J. J., E. M. Bsales, R. J. Jaffe and E. Awh (2017). "Alpha-Band Activity Reveals Spontaneous Representations of Spatial Position in Visual Working Memory." Curr Biol **27**(20): 3216–3223 e3216.

Freedman, D. J., M. Riesenhuber, T. Poggio and E. K. Miller (2001). "Categorical representation of visual stimuli in the primate prefrontal cortex." Science **291**(5502): 312–316.

Freeman, J. and E. P. Simoncelli (2011). "Metamers of the ventral stream." Nat Neurosci **14**(9): 1195–1201.

Froudist-Walsh, S., D. P. Bliss, X. Ding, L. Rapan, M. Niu, K. Knoblauch, K. Zilles, H. Kennedy, N. Palomero-Gallagher and X. J. Wang (2021). "A dopamine gradient controls access to distributed working memory in the large-scale monkey cortex." Neuron **109**(21): 3500–3520 e3513.

Fry, A. F. and S. Hale (2000). "Relationships among processing speed, working memory, and fluid intelligence in children." Biol Psychol **54**(1-3): 1–34.

Fujita, I., K. Tanaka, M. Ito and K. Cheng (1992). "Columns for visual features of objects in monkey inferotemporal cortex." Nature **360**(6402): 343–346.

Fukuda, K., E. Awh and E. K. Vogel (2010). "Discrete capacity limits in visual working memory." Curr Opin Neurobiol **20**(2): 177–182.

Funahashi, S. (2015). "Functions of delay-period activity in the prefrontal cortex and mnemonic scotomas revisited." Front Syst Neurosci **9**: 2.

Funahashi, S., C. J. Bruce and P. S. Goldman-Rakic (1989). "Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex." Journal of Neurophysiology **61**(2): 331–349.

Funahashi, S., C. J. Bruce and P. S. Goldman-Rakic (1993). "Dorsolateral prefrontal lesions and oculomotor delayed-response performance: evidence for mnemonic "scotomas"." Journal of Neuroscience **13**(4): 1479–1497.

Funahashi, S., M. V. Chafee and P. S. Goldman-Rakic (1993). "Prefrontal neuronal activity in rhesus monkeys performing a delayed anti-saccade task." Nature **365**(6448): 753–756.

Fusi, S., E. K. Miller and M. Rigotti (2016). "Why neurons mix: high dimensionality for higher cognition." Curr Opin Neurobiol **37**: 66–74.

Fuster, J. M. (2002). "Frontal lobe and cognitive development." J Neurocytol **31**(3-5): 373–385.

Fuster, J. M. and G. E. Alexander (1971). "Neuron activity related to short-term memory." Science **173**(997): 652–654.

Fuster, J. M. and G. E. Alexander (1973). "Firing changes in cells of the nucleus medialis dorsalis associated with delayed response behavior." Brain Res **61**: 79–91.

Fuster, J. M., M. Bodner and J. K. Kroger (2000). "Cross-modal and cross-temporal association in neurons of frontal cortex." Nature **405**(6784): 347–351.

Fuster, J. M. and J. P. Jervey (1981). "Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli." Science **212**(4497): 952–955.

Fuster, J. M. and J. P. Jervey (1982). "Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task." Journal of Neuroscience **2**(3): 361–375.

Fyall, A. M., Y. El-Shamayleh, H. Choi, E. Shea-Brown and A. Pasupathy (2017). "Dynamic representation of partially occluded objects in primate prefrontal and visual cortex." Elife **6**.

Gabbott, P. L. and S. J. Bacon (1997). "Vasoactive intestinal polypeptide containing neurones in monkey medial prefrontal cortex (mPFC): colocalisation with calretinin." Brain Res **744**(1): 179–184.

Gallace, A. and C. Spence (2009). "The cognitive and neural correlates of tactile memory." Psychol Bull **135**(3): 380–406.

Garavan, H., D. Kelley, A. Rosen, S. M. Rao and E. A. Stein (2000). "Practice-related functional activation changes in a working memory task." Microsc Res Tech **51**(1): 54–63.

Garavan, H., D. Kelley, A. Rosen, S. M. Rao and E. A. Stein (2000). "Practice-related functional activation changes in a working memory task." Microsc.Res.Tech. **51**(1): 54–63.

Garcia-Cabezas, M. A., M. K. P. Joyce, Y. J. John, B. Zikopoulos and H. Barbas (2017). "Mirror trends of plasticity and stability indicators in primate prefrontal cortex." Eur J Neurosci **46**(8): 2392–2405.

Gardiner, J. M., B. Gawlik and A. Richardson-Klavehn (1994). "Maintenance rehearsal affects knowing, not remembering; elaborative rehearsal affects remembering, not knowing." Psychon Bull Rev **1**(1): 107–110.

Gathercole, S. E., S. J. Pickering, B. Ambridge and H. Wearing (2004). "The structure of working memory from 4 to 15 years of age." Dev Psychol **40**(2): 177–190.

Genovesio, A., S. Tsujimoto and S. P. Wise (2009). "Feature- and order-based timing representations in the frontal cortex." Neuron **63**(2): 254–266.

Gerbella, M., E. Borra, S. Tonelli, S. Rozzi and G. Luppino (2013). "Connectional heterogeneity of the ventral part of the macaque area 46." Cereb Cortex **23**(4): 967–987.

Gibson, E. M., D. Purger, C. W. Mount, A. K. Goldstein, G. L. Lin, L. S. Wood, I. Inema, S. E. Miller, G. Bieri, J. B. Zuchero, B. A. Barres, P. J. Woo, H. Vogel and M. Monje (2014). "Neuronal activity promotes oligodendrogenesis and adaptive myelination in the mammalian brain." Science **344**(6183): 1252304.

Giguere, M. and P. S. Goldman-Rakic (1988). "Mediodorsal nucleus: areal, laminar, and tangential distribution of afferents and efferents in the frontal lobe of rhesus monkeys." J Comp Neurol **277**: 195–213.

Glasser, M. F. and D. C. Van Essen (2011). "Mapping human cortical areas in vivo based on myelin content as revealed by T1- and T2-weighted MRI." J Neurosci **31**(32): 11597–11616.

Gnadt, J. W. and R. A. Andersen (1988). "Memory related motor planning activity in posterior parietal cortex of macaque." Experimental Brain Research **70**(1): 216–220.

Goelet, P., V. F. Castellucci, S. Schacher and E. R. Kandel (1986). "The long and the short of long-term memory--a molecular framework." Nature **322**(6078): 419–422.

Goldman-Rakic, P. S. (1984). "Modular organization of prefrontal cortex." Trends in Neurosciences **7**(11): 419–424.

Goldman-Rakic, P. S. and L. J. Porrino (1985). "The primate mediodorsal (MD) nucleus and its projection to the frontal lobe." J Comp Neurol **242**(4): 535–560.

Gonzalez-Burgos, G., T. Miyamae, D. E. Pafundo, H. Yoshino, D. C. Rotaru, G. Hoftman, D. Datta, Y. Zhang, M. Hammond, A. R. Sampson, K. N. Fish, G. B. Ermentrout and D. A. Lewis (2015). "Functional Maturation of GABA Synapses During Postnatal Development of the Monkey Dorsolateral Prefrontal Cortex." Cereb Cortex **25**(11): 4076–4093.

Goodwin, S. J., R. K. Blackman, S. Sakellaridi and M. V. Chafee (2012). "Executive control over cognition: stronger and earlier rule-based modulation of spatial category signals in prefrontal cortex relative to parietal cortex." J Neurosci **32**(10): 3499–3515.

Gordon, E. M., A. L. Breeden, S. E. Bean and C. J. Vaidya (2014). "Working memory-related changes in functional connectivity persist beyond task disengagement." Hum Brain Mapp **35**(3): 1004–1017.

Gottlieb, Y., E. Vaadia and M. Abeles (1989). "Single unit activity in the auditory cortex of a monkey performing a short term memory task." Exp Brain Res **74**(1): 139–148.

Goulas, A., P. Stiers, R. M. Hutchison, S. Everling, M. Petrides and D. S. Margulies (2017). "Intrinsic functional architecture of the macaque dorsal and ventral lateral frontal cortex." J Neurophysiol **117**(3): 1084–1099.

Granon, S., J. Hardouin, A. Courtier and B. Poucet (1998). "Evidence for the involvement of the rat prefrontal cortex in sustained attention." Q J Exp Psychol B **51**(3): 219–233.

Grill-Spector, K., R. Henson and A. Martin (2006). "Repetition and the brain: neural models of stimulus-specific effects." Trends Cogn Sci **10**(1): 14–23.

Groman, S. M., S. L. Thompson, D. Lee and J. R. Taylor (2022). "Reinforcement learning detuned in addiction: integrative and translational approaches." Trends Neurosci **45**(2): 96–105.

Gross, C. G., C. E. Rocha-Miranda and D. B. Bender (1972). "Visual properties of neurons in inferotemporal cortex of the Macaque." J Neurophysiol **35**(1): 96–111.

Grossman, C. D., V. Man and J. P. O'Doherty (2026). "The representation and valuation of subgoals in the human brain during model-based hierarchical behavior." Neuron.

Gunturkun, O. (2005). "The avian 'prefrontal cortex' and cognition." Curr Opin Neurobiol **15**(6): 686–693.

Guo, K., N. Yamawaki, K. Svoboda and G. M. G. Shepherd (2018). "Anterolateral Motor Cortex Connects with a Medial Subdivision of Ventromedial Thalamus through Cell Type-Specific Circuits, Forming an Excitatory Thalamo-Cortico-Thalamic Loop via Layer 1 Apical Tuft Dendrites of Layer 5B Pyramidal Tract Type Neurons." J Neurosci **38**(41): 8787–8797.

Haber, S. N., K. Kunishio, M. Mizobuchi and E. Lynd-Balta (1995). "The orbital and medial prefrontal circuit through the primate basal ganglia." J Neurosci **15**(7 Pt 1): 4851–4867.

Hackett, T. A., I. Stepniewska and J. H. Kaas (1999). "Prefrontal connections of the parabelt auditory cortex in macaque monkeys." Brain Research **817**(1-2): 45–58.

Hadj-Bouziane, F., M. Meunier and D. Boussaoud (2003). "Conditional visuo-motor learning in primates: a key role for the basal ganglia." J Physiol Paris **97**(4-6): 567–579.

Hamid, A. A., M. J. Frank and C. I. Moore (2021). "Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment." Cell **184**(10): 2733–2749 e2716.

Hampson, R. E., T. P. Pons, T. R. Stanford and S. A. Deadwyler (2004). "Categorization in the monkey hippocampus: a possible mechanism for encoding information into memory." Proc Natl Acad Sci U S A **101**(9): 3184–3189. Epub 2004 Feb 3120.

Hampson, R. E., J. D. Simeral and S. A. Deadwyler (1999). "Distribution of spatial and nonspatial information in dorsal hippocampus." Nature **402**(6762): 610–614.

Hannula, H., T. Neuvonen, P. Savolainen, J. Hiltunen, Y. Y. Ma, H. Antila, O. Salonen, S. Carlson and A. Pertovaara (2010). "Increasing top-down suppression from prefrontal cortex facilitates tactile working memory." Neuroimage **49**(1): 1091–1098.

Hansel, D. and G. Mato (2013). "Short-term plasticity explains irregular persistent activity in working memory tasks." J Neurosci **33**(1): 133–149.

Harrison, S. A. and F. Tong (2009). "Decoding reveals the contents of visual working memory in early visual areas." Nature **458**(7238): 632–635.

Hart, E. and A. C. Huk (2020). "Recurrent circuit dynamics underlie persistent activity in the macaque frontoparietal network." Elife **9**.

Hazy, T. E., Frank, M.J., O'Reilly, R.C. (2006). "Banishing the homunculus: making working memory work." Neuroscience **139**(1): 105–118.

He, B. J. (2013). "Spontaneous and task-evoked brain activity negatively interact." J Neurosci **33**(11): 4672–4682.

Hebb, D. O. (2002). The Organization of Behavior: A Neuropsychological Theory, Taylor & Francis.

Hempel, A., F. L. Giesel, N. M. Garcia Caraballo, M. Amann, H. Meyer, T. Wustenberg, M. Essig and J. Schroder (2004). "Plasticity of cortical activation related to working memory during training." Am J Psychiatry **161**(4): 745–747.

Hempel, A., F. L. Giesel, N. M. Garcia Caraballo, M. Amann, H. Meyer, T. Wustenberg, M. Essig and J. Schroder (2004). "Plasticity of cortical activation related to working memory during training." Am.J Psychiatry **161**(4): 745–747.

Hempel, C. M., K. H. Hartman, X. J. Wang, G. G. Turrigiano and S. B. Nelson (2000). "Multiple forms of short-term plasticity at excitatory synapses in rat medial prefrontal cortex." J Neurophysiol **83**(5): 3031–3041.

Herman, R. A., J. L. Zehr and K. Wallen (2006). "Prenatal androgen blockade accelerates pubertal development in male rhesus monkeys." Psychoneuroendocrinology **31**(1): 118–130.

Hernandez, A., A. Zainos and R. Romo (2002). "Temporal evolution of a decision-making process in medial premotor cortex." Neuron **33**(6): 959–972.

Heyselaar, E., K. Johnston and M. Pare (2011). "A change detection approach to study visual working memory of the macaque monkey." J Vis **11**(3).

Hikosaka, O., M. Sakamoto and S. Usui (1989). "Functional properties of monkey caudate neurons. I. Activities related to saccadic eye movements." J Neurophysiol **61**(4): 780–798.

Hikosaka, O., M. Sakamoto and S. Usui (1989). "Functional properties of monkey caudate neurons. III. Activities related to expectation of target and reward." J Neurophysiol **61**(4): 814–832.

Hoftman, G. D. and D. A. Lewis (2011). "Postnatal developmental trajectories of neural circuits in the primate prefrontal cortex: identifying sensitive periods for vulnerability to schizophrenia." Schizophr Bull **37**(3): 493–503.

Holmes, C. D., C. Papadimitriou and L. H. Snyder (2018). "Dissociation of LFP Power and Tuning in the Frontal Cortex during Memory." J Neurosci **38**(38): 8177–8186.

Hoshi, E., K. Shima and J. Tanji (1998). "Task-dependent selectivity of movement-related neuronal activity in the primate prefrontal cortex." Journal of Neurophysiology **80**(6): 3392–3397.

Houk, J. C. and S. P. Wise (1995). "Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: their role in planning and controlling action." Cereb Cortex **5**(2): 95–110.

Hulme, C., Maughan, S., Brown, G. (1991). "Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span." Journal of Memory and Language **30**(6): 685–701.

Huntenburg, J. M., P. L. Bazin, A. Goulas, C. L. Tardif, A. Villringer and D. S. Margulies (2017). "A Systematic Relationship Between Functional Connectivity and Intracortical Myelin in the Human Cerebral Cortex." Cereb Cortex **27**(2): 981–997.

Hussar, C. and T. Pasternak (2010). "Trial-to-trial variability of the prefrontal neurons reveals the nature of their engagement in a motion discrimination task." Proc Natl Acad Sci U S A **107**(50): 21842–21847.

Hussar, C. R. and T. Pasternak (2012). "Memory-guided sensory comparisons in the prefrontal cortex: contribution of putative pyramidal cells and interneurons." J Neurosci **32**(8): 2747–2761.

Hussar, C. R. and T. Pasternak (2013). "Common rules guide comparisons of speed and direction of motion in the dorsolateral prefrontal cortex." J Neurosci **33**(3): 972–986.

Hwang, J. and L. M. Romanski (2015). "Prefrontal neuronal responses during audiovisual mnemonic processing." J Neurosci **35**(3): 960–971.

Ibos, G., J. R. Duhamel and S. Ben Hamed (2013). "A functional hierarchy within the parietofrontal network in stimulus selection and attention control." J Neurosci **33**(19): 8359–8369.

Ifuku, H., S. Hirata, T. Nakamura and H. Ogawa (2003). "Neuronal activities in the monkey primary and higher-order gustatory cortices during a taste discrimination delayed GO/NOGO task and after reversal." Neurosci Res **47**(2): 161–175.

Inoue, M. and A. Mikami (2006). "Prefrontal activity during serial probe reproduction task: encoding, mnemonic, and retrieval processes." J Neurophysiol **95**(2): 1008–1041.

Isbell, E., K. Fukuda, H. J. Neville and E. K. Vogel (2015). "Visual working memory continues to develop through adolescence." Front Psychol **6**: 696.

Izquierdo, A., J. L. Brigman, A. K. Radke, P. H. Rudebeck and A. Holmes (2017). "The neural basis of reversal learning: An updated perspective." Neuroscience **345**: 12–26.

Jacob, S. N. and A. Nieder (2014). "Complementary roles for primate frontal and parietal cortex in guarding working memory from distractor stimuli." Neuron **83**(1): 226–237.

Jacob, S. N., M. Stalter and A. Nieder (2016). "Cell-type-specific modulation of targets and distractors by dopamine D1 receptors in primate prefrontal cortex." Nat Commun **7**: 13218.

Jaeggi, S. M., M. Buschkuehl, J. Jonides and W. J. Perrig (2008). "Improving fluid intelligence with training on working memory." Proc Natl Acad Sci U S A **105**(19): 6829–6833.

Jaeggi, S. M., M. Buschkuehl, J. Jonides and W. J. Perrig (2008). "Improving fluid intelligence with training on working memory." Proc.Natl.Acad.Sci.U.S.A **105**(19): 6829–6833.

Jaffe, R. J. and C. Constantinidis (2021). "Working Memory: From Neural Activity to the Sentient Mind." Compr Physiol **11**(4): 2547–2587.

Jansma, J. M., N. F. Ramsey, H. A. Slagter and R. S. Kahn (2001). "Functional anatomical correlates of controlled and automatic processing." J Cogn Neurosci **13**(6): 730–743.

Janssen, P., S. Srivastava, S. Ombelet and G. A. Orban (2008). "Coding of shape and position in macaque lateral intraparietal area." J Neurosci **28**(26): 6679–6690.

Jernigan, T. L., S. L. Archibald, M. T. Berhow, E. R. Sowell, D. S. Foster and J. R. Hesselink (1991). "Cerebral structure on MRI, Part I: Localization of age-related changes." Biol Psychiatry **29**(1): 55–67.

Johnston, M., C. Anderson and M. Colombo (2017). "Neural correlates of sample-coding and reward-coding in the delay activity of neurons in the entopallium and nidopallium caudolaterale of pigeons (Columba livia)." Behav Brain Res **317**: 382–392.

Johnston, M., B. Porter and M. Colombo (2019). "Delay activity in pigeon nidopallium caudolaterale during a variable-delay memory task." Behav Neurosci **133**(6): 563–568.

Johnston, W. J., S. E. Palmer and D. J. Freedman (2020). "Nonlinear mixed selectivity supports reliable neural computation." PLoS Comput Biol **16**(2): e1007544.

Jolles, D. D., M. A. van Buchem, E. A. Crone and S. A. Rombouts (2013). "Functional brain connectivity at rest changes after working memory training." Hum Brain Mapp **34**(2): 396–406.

Jones, E. G., M. E. Dell'Anna, M. Molinari, E. Rausell and T. Hashikawa (1995). "Subdivisions of macaque monkey auditory cortex revealed by calcium-binding protein immunoreactivity." J Comp Neurol **362**(2): 153–170.

Jonides, J., E. E. Smith, R. A. Koeppe, E. Awh, S. Minoshima and M. A. Mintun (1993). "Spatial working memory in humans as revealed by PET." Nature **363**(6430): 623–625.

Jun, H., J. Y. Lee, N. R. Bleza, A. Ichii, J. D. Donohue and K. M. Igarashi (2024). "Prefrontal and lateral entorhinal neurons co-dependently learn item-outcome rules." Nature **633**(8031): 864–871.

Kaas, J. H. and T. A. Hackett (2000). "Subdivisions of auditory cortex and processing streams in primates." Proc Natl Acad Sci U S A **97**(22): 11793–11799.

Kadohisa, M., M. Kusunoki, P. Petrov, N. Sigala, M. J. Buckley, D. Gaffan and J. Duncan (2015). "Spatial and temporal distribution of visual information coding in lateral prefrontal cortex." Eur J Neurosci **41**(1): 89–96.

Kahn, J. B., R. D. Ward, L. W. Kahn, N. M. Rudy, E. R. Kandel, P. D. Balsam and E. H. Simpson (2012). "Medial prefrontal lesions in mice impair sustained attention but spare maintenance of information in working memory." Learn Mem **19**(11): 513–517.

Kalt, T., B. Diekamp and O. Gunturkun (1999). "Single unit activity during a Go/NoGo task in the "prefrontal cortex" of pigeons." Brain Res **839**(2): 263–278.

Kamigaki, T. and Y. Dan (2017). "Delay activity of specific prefrontal interneuron subtypes modulates memory-guided behavior." Nat Neurosci **20**(6): 854–863.

Kaminski, J., S. Sullivan, J. M. Chung, I. B. Ross, A. N. Mamelak and U. Rutishauser (2017). "Persistently active neurons in human medial frontal and medial temporal lobe support working memory." Nat Neurosci **20**(4): 590–601.

Karami, B. and C. M. Schwiedrzik (2024). "Visual perceptual learning of feature conjunctions leverages non-linear mixed selectivity." NPJ Sci Learn **9**(1): 13.

Katsuki, F. and C. Constantinidis (2012). "Unique and shared roles of the posterior parietal and dorsolateral prefrontal cortex in cognitive functions." Frontiers in Integrative Neuroscience **6**: 17.

Katsuki, F., X. L. Qi, T. Meyer, P. M. Kostelic, E. Salinas and C. Constantinidis (2014). "Differences in intrinsic functional organization between dorsolateral prefrontal and posterior parietal cortex." Cereb Cortex **24**(9): 2334–2349.

Kaufman, M. T., M. K. Benna, M. Rigotti, F. Stefanini, S. Fusi and A. K. Churchland (2022). "The implications of categorical and category-free mixed selectivity on representational geometries." Curr Opin Neurobiol **77**: 102644.

Kawagoe, R., Y. Takikawa and O. Hikosaka (1998). "Expectation of reward modulates cognitive signals in the basal ganglia." Nat Neurosci **1**(5): 411–416.

Kermadi, I. and J. P. Joseph (1995). "Activity in the caudate nucleus of monkey during spatial sequencing." J Neurophysiol **74**(3): 911–933.

Kilpatrick, Z. P. (2018). "Synaptic mechanisms of interference in working memory." Sci Rep **8**(1): 7879.

Kim, C. M., C. C. Chow and B. B. Averbeck (2025). "Neural dynamics of reversal learning in the prefrontal cortex and recurrent neural networks." Elife **13**.

Kim, J. N. and M. N. Shadlen (1999). "Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque." Nature Neuroscience **2**(2): 176–185.

Kim, R. and T. J. Sejnowski (2021). "Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks." Nat Neurosci **24**(1): 129–139.

Kirsch, J. A., O. Gunturkun and J. Rose (2008). "Insight without cortex: lessons from the avian brain." Conscious Cogn **17**(2): 475–483.

Kliegl, R. J., S. Heckhausen, J. Baltes, P. (1987). "Mnemonic Training for the Acquisition of Skilled Digit Memory." Cognition and Instruction **4**(4).

Klingberg, T., E. Fernell, P. Olesen, M. Johnson, P. Gustafsson, K. Dahlstrîm, C. G. Gillberg, H. Forssberg and H. Westerberg (2005). "Computerized Training of Working Memory in Children with ADHD - a Randomized, Controlled Trial." Journal of the American Academy of Child and Adolescent Psychiatry **44**(2): 177–186.

Klingberg, T., E. Fernell, P. J. Olesen, M. Johnson, P. Gustafsson, K. Dahlstrom, C. G. Gillberg, H. Forssberg and H. Westerberg (2005). "Computerized training of working memory in children with ADHD--a randomized, controlled trial." J Am Acad Child Adolesc Psychiatry **44**(2): 177–186.

Klingberg, T., H. Forssberg and H. Westerberg (2002). "Training of working memory in children with ADHD." J Clin Exp Neuropsychol **24**(6): 781–791.

Klingberg, T., H. Forssberg and H. Westerberg (2002). "Training of working memory in children with ADHD." Journal of Clinical and Experimental Neuropsychology **24**(6): 781–791.

Kobak, D., W. Brendel, C. Constantinidis, C. E. Feierstein, A. Kepecs, Z. F. Mainen, X. L. Qi, R. Romo, N. Uchida and C. K. Machens (2016). "Demixed principal component analysis of neural population data." Elife **5**.

Kobak, D., W. Brendel, C. Constantinidis, C. E. Feierstein, A. Kepecs, Z. F. Mainen, R. Romo, X. L. Qi, N. Uchida and C. K. Machens (2014). "Demixed principal component analysis of population activity in higher cortical areas reveals independent representation of task parameters." arXiv: 1410.6031.

Koch, C. and J. M. Fuster (1989). "Unit activity in monkey parietal cortex related to haptic perception and temporary memory." Experimental Brain Research **76**(2): 292–306.

Konecky, R. O., M. A. Smith and C. R. Olson (2017). "Monkey prefrontal neurons during Sternberg task performance: full contents of working memory or most recent item?" J Neurophysiol **117**(6): 2269–2281.

Kosaki, H., T. Hashikawa, J. He and E. G. Jones (1997). "Tonotopic organization of auditory cortical fields delineated by parvalbumin immunoreactivity in macaque monkeys." J Comp Neurol **386**(2): 304–316.

Kritzer, M. F. and P. S. Goldman-Rakic (1995). "Intrinsic circuit organization of the major layers and sublayers of the dorsolateral prefrontal cortex in the rhesus monkey." Journal of Comparative Neurology **359**(1): 131–143.

Kroner, S. and O. Gunturkun (1999). "Afferent and efferent connections of the caudolateral neostriatum in the pigeon (Columba livia): a retro- and anterograde pathway tracing study." J Comp Neurol **407**(2): 228–260.

Kuboshima-Amemori, S. and T. Sawaguchi (2007). "Plasticity of the primate prefrontal cortex." Neuroscientist **13**(3): 229–240.

Kubota, K. and H. Niki (1971). "Prefrontal cortical unit activity and delayed alternation performance in monkeys." Journal of Neurophysiology **34**(3): 337–347.

Kuhn, S., F. Schmiedek, H. Noack, E. Wenger, N. C. Bodammer, U. Lindenberger and M. Lovden (2013). "The dynamics of change in striatal activity following updating training." Hum Brain Mapp **34**(7): 1530–1541.

Kumar, S., S. Joseph, P. E. Gander, N. Barascud, A. R. Halpern and T. D. Griffiths (2016). "A Brain System for Auditory Working Memory." J Neurosci **36**(16): 4492–4505.

Kuo, B. C., M. G. Stokes, A. M. Murray and A. C. Nobre (2014). "Attention biases visual activity in visual short-term memory." J Cogn Neurosci **26**(7): 1377–1389.

Kwon, H., A. L. Reiss and V. Menon (2002). "Neural basis of protracted developmental changes in visuo-spatial working memory." Proc Natl Acad Sci U S A **99**(20): 13336–13341.

Lafer-Sousa, R. and B. R. Conway (2013). "Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex." Nat Neurosci **16**(12): 1870–1878.

Lara, A. H., S. W. Kennerley and J. D. Wallis (2009). "Encoding of gustatory working memory by orbitofrontal neurons." J Neurosci **29**(3): 765–774.

Lara, A. H. and J. D. Wallis (2012). "Capacity and precision in an animal model of visual short-term memory." J Vis **12**(3).

Lara, A. H. and J. D. Wallis (2014). "Executive control processes underlying multi-item working memory." Nat Neurosci **17**(6): 876–883.

Laroche, S., S. Davis and T. M. Jay (2000). "Plasticity at hippocampal to prefrontal cortex synapses: dual roles in working memory and consolidation." Hippocampus **10**(4): 438–446.

Laubach, M., L. M. Amarante, K. Swanson and S. R. White (2018). "What, If Anything, Is Rodent Prefrontal Cortex?" eNeuro **5**(5).

Laughlin, S. B., R. R. de Ruyter van Steveninck and J. C. Anderson (1998). "The metabolic cost of neural information." Nat Neurosci **1**(1): 36–41.

Lauwereyns, J., M. Sakagami, K. Tsutsui, S. Kobayashi, M. Koizumi and O. Hikosaka (2001). "Responses to task-irrelevant visual features by primate prefrontal neurons." J Neurophysiol **86**(4): 2001–2010.

Lawrence, S. J. D., T. van Mourik, P. Kok, P. J. Koopmans, D. G. Norris and F. P. de Lange (2018). "Laminar Organization of Working Memory Signals in Human Visual Cortex." Curr Biol **28**(21): 3435–3440 e3434.

Leavitt, M. L., D. Mendoza-Halliday and J. C. Martinez-Trujillo (2017). "Sustained Activity Encoding Working Memories: Not Fully Distributed." Trends Neurosci **40**(6): 328–346.

Leavitt, M. L., F. Pieper, A. J. Sachs and J. C. Martinez-Trujillo (2017). "A Quadrantic Bias in Prefrontal Representation of Visual-Mnemonic Space." Cereb Cortex: 1–17.

Lebedev, M. A., A. Messinger, J. D. Kralik and S. P. Wise (2004). "Representation of attended versus remembered locations in prefrontal cortex." PLoS Biol **2**(11): e365. Epub 2004 Oct 2026.

LeCun, Y., Y. Bengio and G. Hinton (2015). "Deep learning." Nature **521**(7553): 436–444.

Lee, E. Y., N. Cowan, E. K. Vogel, T. Rolan, F. Valle-Inclan and S. A. Hackley (2010). "Visual working memory deficits in patients with Parkinson's disease are due to both reduced storage capacity and impaired ability to filter out irrelevant information." Brain **133**(9): 2677–2689.

Lemus, L., A. Hernandez and R. Romo (2009). "Neural codes for perceptual discrimination of acoustic flutter in the primate auditory cortex." Proc Natl Acad Sci U S A **106**(23): 9471–9476.

Leon, M. I. and M. N. Shadlen (1999). "Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque." Neuron **24**(2): 415–425.

Levitt, J. B., D. A. Lewis, T. Yoshioka and J. S. Lund (1993). "Topography of pyramidal neuron intrinsic connections in macaque monkey prefrontal cortex (areas 9 and 46)." Journal of Comparative Neurology **338**(3): 360–376.

Levitt, P., P. Rakic and P. Goldman-Rakic (1984). "Region-specific distribution of catecholamine afferents in primate cerebral cortex: a fluorescence histochemical analysis." Journal of Comparative Neurology **227**(1): 23–36.

Lewis, D. A. (1997). "Development of the prefrontal cortex during adolescence: insights into vulnerable neural circuits in schizophrenia." Neuropsychopharmacology **16**(6): 385–398.

Li, C. S., P. Mazzoni and R. A. Andersen (1999). "Effect of reversible inactivation of macaque lateral intraparietal area on visual and memory saccades." Journal of Neurophysiology **81**(4): 1827–1838.

Li, D., C. Constantinidis and J. D. Murray (2020). "Trial-to-trial variability of spiking delay activity in prefrontal cortex constrains burst-coding models of working memory." bioRxiv **10.1101/2021.01.30.428962**.

Li, S., X. Zhou, C. Constantinidis and X. L. Qi (2020). "Plasticity of Persistent Activity and Its Constraints." Front Neural Circuits **14**: 15.

Li, S. C., F. Schmiedek, O. Huxhold, C. Rocke, J. Smith and U. Lindenberger (2008). "Working memory plasticity in old age: practice gain, transfer, and maintenance." Psychol Aging **23**(4): 731–742.

Liebe, S., G. M. Hoerzer, N. K. Logothetis and G. Rainer (2012). "Theta coupling between V4 and prefrontal cortex predicts visual short-term memory performance." Nat Neurosci **15**(3): 456–462, S451–452.

Lindsay, G. W., M. Rigotti, M. R. Warden, E. K. Miller and S. Fusi (2017). "Hebbian Learning in a Random Network Captures Selectivity Properties of the Prefrontal Cortex." J Neurosci **37**(45): 11021–11036.

Lisman, J. E. and A. A. Grace (2005). "The hippocampal-VTA loop: controlling the entry of information into long-term memory." Neuron **46**(5): 703–713.

Lisman, J. E. and M. A. Idiart (1995). "Storage of 7 +/- 2 short-term memories in oscillatory subcycles." Science **267**(5203): 1512–1515.

Liu, R., J. Crawford, P. M. Callahan, A. V. Terry, Jr., C. Constantinidis and D. T. Blake (2017). "Intermittent Stimulation of the Nucleus Basalis of Meynert Improves Working Memory in Adult Monkeys." Curr Biol **7**(17): 2640–2646.

Liu, Y., E. A. Yttri and L. H. Snyder (2010). "Intention and attention: different functional roles for LIPd and LIPv." Nat Neurosci **13**(4): 495–500.

Llinas, R., U. Ribary, D. Contreras and C. Pedroarena (1998). "The neuronal basis for consciousness." Philos Trans R Soc Lond B Biol Sci **353**(1377): 1841–1849.

Llinas, R. R., E. Leznik and F. J. Urbano (2002). "Temporal binding via cortical coincidence detection of specific and nonspecific thalamocortical inputs: a

voltage-dependent dye-imaging study in mouse brain slices." Proc Natl Acad Sci U S A **99**(1): 449–454.

Lloyd, K., N. Becker, M. W. Jones and R. Bogacz (2012). "Learning to use working memory: a reinforcement learning gating model of rule acquisition in rats." Front Comput Neurosci **6**: 87.

Logothetis, N. K. and B. A. Wandell (2004). "Interpreting the BOLD signal." Annu Rev Physiol **66**: 735–769.

Lowet, E., B. Gomes, K. Srinivasan, H. Zhou, R. J. Schafer and R. Desimone (2018). "Enhanced Neural Processing by Covert Attention only during Microsaccades Directed toward the Attended Stimulus." Neuron **99**(1): 207–214 e203.

Luck, S. J. and E. K. Vogel (1997). "The capacity of visual working memory for features and conjunctions." Nature **390**(6657): 279–281.

Luck, S. J. and E. K. Vogel (2013). "Visual working memory capacity: from psychophysics and neurobiology to individual differences." Trends Cogn Sci **17**(8): 391–400.

Luna, B., K. E. Garver, T. A. Urban, N. A. Lazar and J. A. Sweeney (2004). "Maturation of cognitive processes from late childhood to adulthood." Child Dev **75**(5): 1357–1372.

Lundqvist, M., P. Herman and E. K. Miller (2018). "Working Memory: Delay Activity, Yes! Persistent Activity? Maybe Not." J Neurosci **38**(32): 7013–7019.

Lundqvist, M., P. Herman, M. R. Warden, S. L. Brincat and E. K. Miller (2018). "Gamma and beta bursts during working memory readout suggest roles in its volitional control." Nat Commun **9**(1): 394.

Lundqvist, M., J. Rose, P. Herman, S. L. Brincat, T. J. Buschman and E. K. Miller (2016). "Gamma and Beta Bursts Underlie Working Memory." Neuron **90**(1): 152–164.

Lundqvist, M., J. Rose, P. Herman, S. L. Brincat, T. J. Buschman and E. K. Miller (2016). "Gamma and Beta Bursts Underlie Working Memory." Neuron **90**(1): 152–164.

Luria, R., P. Sessa, A. Gotler, P. Jolicoeur and R. Dell'Acqua (2010). "Visual short-term memory capacity for simple and complex objects." J Cogn Neurosci **22**(3): 496–512.

Lyamzin, D. and A. Benucci (2019). "The mouse posterior parietal cortex: Anatomy and functions." Neurosci Res **140**: 14–22.

Machens, C. K., R. Romo and C. D. Brody (2010). "Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex." J Neurosci **30**(1): 350–360.

Machens, C. K., R. Romo and C. D. Brody (2010). "Functional, But Not Anatomical, Separation of "What" and "When" in Prefrontal Cortex." Journal of Neuroscience **30**(1): 350–360.

Magnussen, S. and M. W. Greenlee (1992). "Retention and disruption of motion information in visual short-term memory." J Exp Psychol Learn Mem Cogn **18**(1): 151–156.

Magnussen, S., E. Idas and S. H. Myhre (1998). "Representation of orientation and spatial frequency in perception and memory: a choice reaction-time analysis." J Exp Psychol Hum Percept Perform **24**(3): 707–718.

Malkova, L., E. Heuer and R. C. Saunders (2006). "Longitudinal magnetic resonance imaging study of rhesus monkey brain development." Eur J Neurosci **24**(11): 3204–3212.

Malmberg, K. J., J. G. W. Raaijmakers and R. M. Shiffrin (2019). "50 years of research sparked by Atkinson and Shiffrin (1968)." Mem Cognit **47**(4): 561–574.

Mansouri, F. A., D. J. Freedman and M. J. Buckley (2020). "Emergence of abstract rules in the primate brain." Nat Rev Neurosci **21**(11): 595–610.

Mansouri, F. A., K. Matsumoto and K. Tanaka (2006). "Prefrontal cell activities related to monkeys' success and failure in adapting to rule changes in a Wisconsin Card Sorting Test analog." J Neurosci **26**(10): 2745–2756.

Mante, V., D. Sussillo, K. V. Shenoy and W. T. Newsome (2013). "Context-dependent computation by recurrent dynamics in prefrontal cortex." Nature **503**(7474): 78–84.

Markowitz, D. A., C. E. Curtis and B. Pesaran (2015). "Multiple component networks support working memory in prefrontal cortex." Proc Natl Acad Sci U S A **112**(35): 11084–11089.

Marshuetz, C., E. E. Smith, J. Jonides, J. DeGutis and T. L. Chenevert (2000). "Order information in working memory: fMRI evidence for parietal and prefrontal mechanisms." J Cogn Neurosci **12**(Suppl 2): 130–144.

Martinussen, R., J. Hayden, S. Hogg-Johnson and R. Tannock (2005). "A meta-analysis of working memory impairments in children with attention-deficit/hyperactivity disorder." J Am Acad Child Adolesc Psychiatry **44**(4): 377–384.

Masse, N. Y., J. M. Hodnefield and D. J. Freedman (2017). "Mnemonic Encoding and Cortical Organization in Parietal and Prefrontal Cortices." J Neurosci **37**(25): 6098–6112.

Masse, N. Y., M. C. Rosen and D. J. Freedman (2020). "Reevaluating the Role of Persistent Neural Activity in Short-Term Memory." Trends Cogn Sci **24**(3): 242–258.

Masse, N. Y., G. R. Yang, H. F. Song, X. J. Wang and D. J. Freedman (2019). "Circuit mechanisms for the maintenance and manipulation of information in working memory." Nat Neurosci **22**(7): 1159–1167.

McEwen, B. S. and J. H. Morrison (2013). "The brain on stress: vulnerability and plasticity of the prefrontal cortex over the life course." Neuron **79**(1): 16–29.

McNab, F., A. Varrone, L. Farde, A. Jucaite, P. Bystritsky, H. Forssberg and T. Klingberg (2009). "Changes in cortical dopamine D1 receptor binding associated with cognitive training." Science **323**(5915): 800–802.

Mejias, J. F. and X. J. Wang (2019). "Mechanisms of distributed working memory in a large-scale model of the macaque neocortex." BioRxiv: 760231.

Melchitzky, D. S. and D. A. Lewis (2008). "Dendritic-targeting GABA neurons in monkey prefrontal cortex: comparison of somatostatin- and calretinin-immunoreactive axon terminals." Synapse **62**(6): 456–465.

Mendoza-Halliday, D. and J. C. Martinez-Trujillo (2017). "Neuronal population coding of perceived and memorized visual features in the lateral prefrontal cortex." Nat Commun **8**: 15471.

Mendoza-Halliday, D., S. Torres and J. C. Martinez-Trujillo (2014). "Sharp emergence of feature-selective sustained activity along the dorsal visual pathway." Nat Neurosci **17**(9): 1255–1262.

Meskenaite, V. (1997). "Calretinin-immunoreactive local circuit neurons in area 17 of the cynomolgus monkey, Macaca fascicularis." J Comp Neurol **379**(1): 113–132.

Meyer, T. and C. Constantinidis (2005). "A software solution for the control of visual behavioral experimentation." J Neurosci Methods **142**(1): 27–34.

Meyer, T. and C. Constantinidis (2005). "A software solution for the control of visual behavioral experimentation." J Neurosci Methods **142**(1): 27–34.

Meyer, T., X. L. Qi and C. Constantinidis (2007). "Persistent discharges in the prefrontal cortex of monkeys naive to working memory tasks." Cereb Cortex **17 Suppl 1**: i70–76.

Meyer, T., X. L. Qi, T. R. Stanford and C. Constantinidis (2011). "Stimulus selectivity in dorsal and ventral prefrontal cortex after training in working memory tasks." J Neurosci **31**(17): 6266–6276.

Meyers, E. M., X. L. Qi and C. Constantinidis (2012). "Incorporation of new information into prefrontal cortical activity after learning working memory tasks." Proc Natl Acad Sci U S A **109**(12): 4651–4656.

Michalareas, G., J. Vezoli, S. van Pelt, J. M. Schoffelen, H. Kennedy and P. Fries (2016). "Alpha-Beta and Gamma Rhythms Subserve Feedback and Feedforward Influences among Human Visual Cortical Areas." Neuron **89**(2): 384–397.

Middleton, F. A. and P. L. Strick (2002). "Basal-ganglia 'projections' to the prefrontal cortex of the primate." Cereb Cortex **12**(9): 926–935.

Miller, E. K. and J. D. Cohen (2001). "An integrative theory of prefrontal cortex function." Annu Rev Neurosci **24**: 167–202.

Miller, E. K., C. A. Erickson and R. Desimone (1996). "Neural mechanisms of visual working memory in prefrontal cortex of the macaque." Journal of Neuroscience **16**(16): 5154–5167.

Miller, E. K., L. Li and R. Desimone (1991). "A neural mechanism for working and recognition memory in inferior temporal cortex." Science **254**(5036): 1377–1379.

Miller, E. K., L. Li and R. Desimone (1993). "Activity of neurons in anterior inferior temporal cortex during a short-term memory task." Journal of Neuroscience **13**(4): 1460–1478.

Miller, E. K., M. Lundqvist and A. M. Bastos (2018). "Working Memory 2.0." Neuron **100**(2): 463–475.

Miller, G. A. (1960). Plans and the structure of behavior. New York,, Holt.

Miller, J. A. and C. Constantinidis (2024). "Timescales of learning in prefrontal cortex." Nat Rev Neurosci **25**(9): 597–610.

Miller, J. A., A. Tambini, A. Kiyonaga and M. D'Esposito (2022). "Long-term learning transforms prefrontal cortex representations during working memory." Neuron **110**(22): 3805–3819 e3806.

Mishkin, M., L. G. Ungerleider and K. A. Macko (1983). "Object vision and spatial vision: Two cortical pathways." Trends in Neuroscience **6**: 414–417.

Miyake, A., N. P. Friedman, M. J. Emerson, A. H. Witzki, A. Howerter and T. D. Wager (2000). "The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis." Cogn Psychol **41**(1): 49–100.

Miyashita, Y. and H. S. Chang (1988). "Neuronal correlate of pictorial short-term memory in the primate temporal cortex." Nature **331**(6151): 68–70.

Moneta, N., M. M. Garvert, H. R. Heekeren and N. W. Schuck (2023). "Task state representations in vmPFC mediate relevant and irrelevant value signals and their behavioral influence." Nat Commun **14**(1): 3156.

Mongillo, G., O. Barak and M. Tsodyks (2008). "Synaptic theory of working memory." Science **319**(5869): 1543–1546.

Montez, D., F. Calabro and B. Luna (2017). "The expression of established cognitive brain states stabilizes with working memory development." eLife **Epub ahead of print**.

Montez, D. F., F. J. Calabro and B. Luna (2019). "Working memory improves developmentally as neural processes stabilize." PLoS One **14**(3): e0213010.

Moreno-Bote, R., J. Beck, I. Kanitscheider, X. Pitkow, P. Latham and A. Pouget (2014). "Information-limiting correlations." Nat Neurosci **17**(10): 1410–1417.

Morris, R. G. and A. D. Baddeley (1988). "Primary and working memory functioning in Alzheimer-type dementia." J Clin Exp Neuropsychol **10**(2): 279–296.

Mozumder, R. and C. Constantinidis (2023). "Single-neuron and population measures of neuronal activity in working memory tasks." J Neurophysiol **130**(3): 694–705.

Muir, J. L., B. J. Everitt and T. W. Robbins (1996). "The cerebral cortex of the rat and visual attentional function: dissociable effects of mediofrontal, cingulate, anterior dorsolateral, and parietal cortex lesions on a five-choice serial reaction time task." Cereb Cortex **6**(3): 470–481.

Murray, J. D., A. Bernacchia, N. A. Roy, C. Constantinidis, R. Romo and X. J. Wang (2017). "Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex." Proc Natl Acad Sci U S A **114**(2): 394–399.

Nair, A., T. Karigo, B. Yang, S. Ganguli, M. J. Schnitzer, S. W. Linderman, D. J. Anderson and A. Kennedy (2023). "An approximate line attractor in the hypothalamus encodes an aggressive state." Cell **186**(1): 178–193 e115.

Nakamura, K. and K. Kubota (1995). "Mnemonic firing of neurons in the monkey temporal pole during a visual recognition memory task." J Neurophysiol **74**(1): 162–178.

Naya, Y., M. Yoshida and Y. Miyashita (2001). "Backward spreading of memory-retrieval signal in the primate temporal cortex." Science **291**(5504): 661–664.

Nieder, A., D. J. Freedman and E. K. Miller (2002). "Representation of the quantity of visual items in the primate prefrontal cortex." Science **297**(5587): 1708–1711.

Niki, H. (1974). "Prefrontal unit activity during delayed alternation in the monkey. I. Relation to direction of response." Brain Research **68**(2): 185–196.

Ninokura, Y., H. Mushiake and J. Tanji (2004). "Integration of Temporal Order and Object Information in the Monkey Lateral Prefrontal Cortex." J Neurophysiol **91**(1): 555–560.

Norris, D. (2019). "Even an activated long-term memory system still needs a separate short-term store: A reply to Cowan (2019)." Psychol Bull **145**(8): 848–853.

Noudoost, B., K. L. Clark and T. Moore (2014). "A distinct contribution of the frontal eye field to the visual representation of saccadic targets." J Neurosci **34**(10): 3687–3698.

Nystrom, L. E., T. S. Braver, F. W. Sabb, M. R. Delgado, D. C. Noll and J. D. Cohen (2000). "Working memory for letters, shapes, and locations: fMRI evidence against stimulus-based regional organization in human prefrontal cortex." Neuroimage **11**(5 Pt 1): 424–446.

O Scalaidhe, S., F. A. Wilson and P. S. Goldman-Rakic (1997). "Areal segregation of face-processing neurons in prefrontal cortex." Science **278**(5340): 1135–1138.

O Scalaidhe, S. P., F. A. Wilson and P. S. Goldman-Rakic (1999). "Face-selective neurons during passive viewing and working memory performance of rhesus monkeys: evidence for intrinsic specialization of neuronal coding." Cerebral Cortex **9**(5): 459–475.

O'Doherty, J. P., U. Rutishauser and K. Iigaya (2021). "The hierarchical construction of value." Curr Opin Behav Sci **41**: 71–77.

Oberauer, K. (2019). "Is Rehearsal an Effective Maintenance Strategy for Working Memory?" Trends Cogn Sci **23**(9): 798–809.

Oberauer, K. (2019). "Working Memory and Attention - A Conceptual Analysis and Review." J Cogn **2**(1): 36.

Oberauer, K. and E. Awh (2022). "Is There an Activity-silent Working Memory?" J Cogn Neurosci **34**(12): 2360–2374.

Offen, S., D. Schluppeck and D. J. Heeger (2009). "The role of early visual cortex in visual short-term memory and visual attention." Vision research **49**(10): 1352–1362.

Olesen, P. J., J. Macoveanu, J. Tegner and T. Klingberg (2007). "Brain activity related to working memory and distraction in children and adults." Cereb Cortex **17**(5): 1047–1054.

Olesen, P. J., Z. Nagy, H. Westerberg and T. Klingberg (2003). "Combined analysis of DTI and fMRI data reveals a joint maturation of white and grey matter in a fronto-parietal network." Brain Res Cogn Brain Res **18**(1): 48–57.

Olesen, P. J., H. Westerberg and T. Klingberg (2004). "Increased prefrontal and parietal activity after training of working memory." Nat Neurosci **7**(1): 75–79.

Ott, T., S. N. Jacob and A. Nieder (2014). "Dopamine receptors differentially enhance rule coding in primate prefrontal cortex neurons." Neuron **84**(6): 1317–1328.

Owen, A. M., A. Hampshire, J. A. Grahn, R. Stenton, S. Dajani, A. S. Burns, R. J. Howard and C. G. Ballard (2010). "Putting brain training to the test." Nature **465**(7299): 775–778.

Owen, A. M., C. E. Stern, R. B. Look, I. Tracey, B. R. Rosen and M. Petrides (1998). "Functional organization of spatial and nonspatial working memory processing within the human lateral frontal cortex." Proceedings of the National Academy of Sciences of the United States of America **95**(13): 7721–7726.

Panichello, M. F., B. DePasquale, J. W. Pillow and T. J. Buschman (2019). "Error-correcting dynamics in visual working memory." Nat Commun **10**(1): 3366.

Pantelis, C., T. R. Barnes, H. E. Nelson, S. Tanner, L. Weatherley, A. M. Owen and T. W. Robbins (1997). "Frontal-striatal cognitive deficits in patients with chronic schizophrenia." Brain **120**(Pt 10): 1823–1843.

Papadimitriou, C., A. Ferdoash and L. H. Snyder (2015). "Ghosts in the machine: memory interference from the previous trial." J Neurophysiol **113**(2): 567–577.

Papadimitriou, C., C. D. Holmes and L. H. Snyder (2021). "Primate Spatial Memory Cells Become Tuned Early and Lose Tuning at Cell-Specific Times." Cereb Cortex.

Park, S., P. S. Holzman and P. S. Goldman-Rakic (1995). "Spatial working memory deficits in the relatives of schizophrenic patients." Archives of General Psychiatry **52**(10): 821–828.

Parkin, A. J. (1998). "The central executive does not exist." J Int Neuropsychol Soc **4**(5): 518–522.

Parthasarathy, A., R. Herikstad, J. H. Bong, F. S. Medina, C. Libedinsky and S. C. Yen (2017). "Mixed selectivity morphs population codes in prefrontal cortex." Nat Neurosci **20**(12): 1770–1779.

Parthasarathy, A., C. Tang, R. Herikstad, L. F. Cheong, S. C. Yen and C. Libedinsky (2019). "Time-invariant working memory representations in the presence of code-morphing in the lateral prefrontal cortex." Nat Commun **10**(1): 4995.

Pashler, H. (1988). "Familiarity and visual change detection." Percept Psychophys **44**(4): 369–378.

Pasternak, T. and M. W. Greenlee (2005). "Working memory in primate sensory systems." Nat Rev Neurosci **6**(2): 97–107.

Peijnenborgh, J. C., P. M. Hurks, A. P. Aldenkamp, J. S. Vles and J. G. Hendriksen (2015). "Efficacy of working memory training in children and adolescents with learning disabilities: A review study and meta-analysis." Neuropsychol Rehabil: 1–28.

Peijnenborgh, J. C., P. M. Hurks, A. P. Aldenkamp, J. S. Vles and J. G. Hendriksen (2016). "Efficacy of working memory training in children and adolescents with learning disabilities: A review study and meta-analysis." Neuropsychol Rehabil **26**(5-6): 645–672.

Peng, X., M. E. Sereno, A. K. Silva, S. R. Lehky and A. B. Sereno (2008). "Shape selectivity in primate frontal eye field." J Neurophysiol **100**(2): 796–814.

Pertzov, Y., S. Manohar and M. Husain (2017). "Rapid forgetting results from competition over time between items in visual working memory." J Exp Psychol Learn Mem Cogn **43**(4): 528–536.

Pesaran, B., J. S. Pezaris, M. Sahani, P. P. Mitra and R. A. Andersen (2002). "Temporal structure in neuronal activity during working memory in macaque parietal cortex." Nat Neurosci **5**(8): 805–811.

Peterson, L. and M. J. Peterson (1959). "Short-term retention of individual verbal items." Journal of Experimental Psychology **58**(3): 193–198.

Petrides, M. (1996). "Specialized systems for the processing of mnemonic information within the primate frontal cortex." Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences **351**(1346): 1455–1461; discussion 1461–1452.

Petrides, M. (2000). "The role of the mid-dorsolateral prefrontal cortex in working memory." Exp Brain Res **133**(1): 44–54.

Petrides, M. (2005). "Lateral prefrontal cortex: architectonic and functional organization." Philos Trans R Soc Lond B Biol Sci **360**(1456): 781–795.

Petrides, M. and D. N. Pandya (1984). "Projections to the frontal cortex from the posterior parietal region in the rhesus monkey." Journal of Comparative Neurology **228**(1): 105–116.

Pfefferbaum, A., D. H. Mathalon, E. V. Sullivan, J. M. Rawles, R. B. Zipursky and K. O. Lim (1994). "A quantitative magnetic resonance imaging study of changes in brain morphology from infancy to late adulthood." Arch Neurol **51**(9): 874–887.

Pinal, D., M. Zurron and F. Diaz (2014). "Effects of load and maintenance duration on the time course of information encoding and retrieval in working memory: from perceptual analysis to post-categorization processes." Front Hum Neurosci **8**: 165.

Plakke, B., J. Hwang and L. M. Romanski (2015). "Inactivation of Primate Prefrontal Cortex Impairs Auditory and Audiovisual Working Memory." J Neurosci **35**(26): 9666–9675.

Plakke, B., C. W. Ng and A. Poremba (2013). "Neural correlates of auditory recognition memory in primate lateral prefrontal cortex." Neuroscience **244**: 62–76.

Plancher, G. and P. Barrouillet (2020). "On some of the main criticisms of the modal model: Reappraisal from a TBRS perspective." Mem Cognit **48**(3): 455–468.

Plant, T. M., S. Ramaswamy, D. Simorangkir and G. R. Marshall (2005). "Postnatal and pubertal development of the rhesus monkey (Macaca mulatta) testis." Ann N Y Acad Sci **1061**: 149–162.

Poeppel, D. and F. Adolfi (2020). "Against the Epistemological Primacy of the Hardware: The Brain from Inside Out, Turned Upside Down." eNeuro **7**(4).

Popivanov, I. D., J. Jastorff, W. Vanduffel and R. Vogels (2014). "Heterogeneous single-unit selectivity in an fMRI-defined body-selective patch." J Neurosci **34**(1): 95–111.

Portrat, S., P. Barrouillet and V. Camos (2008). "Time-related decay or interference-based forgetting in working memory?" J Exp Psychol Learn Mem Cogn **34**(6): 1561–1564.

Postle, B. R. (2006). "Working memory as an emergent property of the mind and brain." Neuroscience **139**(1): 23–38.

Preuss, T. M. (1995). "Do rats have prefrontal cortex? The rose-woolsey-akert program reconsidered." J Cogn Neurosci **7**(1): 1–24.

Preuss, T. M. and P. S. Goldman-Rakic (1991). "Architectonics of the parietal and temporal association cortex in the strepsirhine primate Galago compared to the anthropoid primate Macaca." Journal of Comparative Neurology **310**(4): 475–506.

Proctor, R., Fagnani, C. (1978). "Effects of distractor-stimulus modality in the Brown–Peterson distractor task.

." Journal of Experimental Psychology: Human Learning and Memory **4**(6): 676–684.

Pucak, M. L., J. B. Levitt, J. S. Lund and D. A. Lewis (1996). "Patterns of intrinsic and associational circuitry in monkey prefrontal cortex." Journal of Comparative Neurology **376**(4): 614–630.

Puig, M. V. and E. K. Miller (2012). "The role of prefrontal dopamine D1 receptors in the neural mechanisms of associative learning." Neuron **74**(5): 874–886.

Qi, X. L. and C. Constantinidis (2012). "Correlated discharges in the primate prefrontal cortex before and after working memory training " European Journal of Neuroscience **36**(11): 3538–3548.

Qi, X. L. and C. Constantinidis (2012). "Correlated discharges in the primate prefrontal cortex before and after working memory training." Eur J Neurosci **36**(11): 3538–3548.

Qi, X. L. and C. Constantinidis (2012). "Variability of prefrontal neuronal discharges before and after training in a working memory task." PLoS ONE **7**(7): e41053.

Qi, X. L. and C. Constantinidis (2013). "Neural changes after training to perform cognitive tasks." Behav Brain Res **241**: 235–243.

Qi, X. L., A. C. Elworthy, B. C. Lambert and C. Constantinidis (2015). "Representation of remembered stimuli and task information in the monkey dorsolateral prefrontal and posterior parietal cortex." J Neurophysiol **113**(1): 44–57.

Qi, X. L., F. Katsuki, T. Meyer, J. B. Rawley, X. Zhou, K. L. Douglas and C. Constantinidis (2010). "Comparison of neural activity related to working memory in primate dorsolateral prefrontal and posterior parietal cortex." Front Syst Neurosci **4**: 12.

Qi, X. L., R. Liu, A. I. Vazdarjanova, D. T. Blake and C. Constantinidis (2019). "Nucleus Basalis Stimulation Enhances Working Memory and Stabilizes Attractor Networks in Prefrontal Cortex." bioRxiv: 674465.

Qi, X. L., T. Meyer, T. R. Stanford and C. Constantinidis (2011). "Changes in Prefrontal Neuronal Activity after Learning to Perform a Spatial Working Memory Task." Cerebral Cortex **21**: 2722–2732.

Qi, X. L., T. Meyer, T. R. Stanford and C. Constantinidis (2012). "Neural correlates of a decision variable before learning to perform a match/non-match task." J Neurosci **32**(18): 6161–6169.

Qi, X. L., T. Meyer, T. R. Stanford and C. Constantinidis (2012). "Neural correlates of a decision variable before learning to perform a Match/Nonmatch task." Journal of Neuroscience **32**(18): 6161–6169.

Qi, X. L., X. Zhou and C. Constantinidis (2015). Neurophysiological Mechanisms of Working Memory: Cortical Specialization & Plasticity. Attention and Performance XXV. P. Jolicoeur, C. Lefebre and J. C. Martinez-Trujillo. London, Academic Press**:** 171–186.

Quintana, J., J. Yajeya and J. M. Fuster (1988). "Prefrontal representation of stimulus attributes during delay tasks. I. Unit activity in cross-temporal integration of sensory and sensory-motor information." Brain Research **474**(2): 211–221.

Raffone, A. and G. Wolters (2001). "A cortical mechanism for binding in visual working memory." J Cogn Neurosci **13**(6): 766–785.

Rainer, G., W. F. Asaad and E. K. Miller (1998). "Memory fields of neurons in the primate prefrontal cortex." Proceedings of the National Academy of Sciences of the United States of America **95**(25): 15008–15013.

Rainer, G., W. F. Asaad and E. K. Miller (1998). "Selective representation of relevant information by neurons in the primate prefrontal cortex." Nature **393**(6685): 577–579.

Rainer, G. and E. K. Miller (2000). "Effects of visual experience on the representation of objects in the prefrontal cortex." Neuron **27**(1): 179–189.

Rama, P., A. Poremba, J. B. Sala, L. Yee, M. Malloy, M. Mishkin and S. M. Courtney (2004). "Dissociable functional cortical topographies for working memory maintenance of voice identity and location." Cereb Cortex **14**(7): 768–780.

Ranganath, C., M. X. Cohen and C. J. Brozinsky (2005). "Working memory maintenance contributes to long-term memory formation: neural and behavioral evidence." J Cogn Neurosci **17**(7): 994–1010.

Ranganath, C., M. K. Johnson and M. D'Esposito (2003). "Prefrontal activity associated with working memory and episodic long-term memory." Neuropsychologia **41**(3): 378–389.

Rao, S. C., G. Rainer and E. K. Miller (1997). "Integration of what and where in the primate prefrontal cortex." Science **276**(5313): 821–824.

Rao, S. G., G. V. Williams and P. S. Goldman-Rakic (1999). "Isodirectional tuning of adjacent interneurons and pyramidal cells during working memory: evidence for microcolumnar organization in PFC." Journal of Neurophysiology **81**(4): 1903–1916.

Rao, S. G., G. V. Williams and P. S. Goldman-Rakic (2000). "Destruction and creation of spatial tuning by disinhibition: GABA(A) blockade of prefrontal cortical neurons engaged by working memory." Journal of Neuroscience **20**(1): 485–494.

Rao, S. M., D. L. Harrington, K. Y. Haaland, J. A. Bobholz, R. W. Cox and J. R. Binder (1997). "Distributed neural systems underlying the timing of movements." J Neurosci **17**(14): 5528–5535.

Rauschecker, J. P., B. Tian and M. Hauser (1995). "Processing of complex sounds in the macaque nonprimary auditory cortex." Science **268**(5207): 111–114.

Rawley, J. B. and C. Constantinidis (2009). "Neural correlates of learning and working memory in the primate posterior parietal cortex." Neurobiol Learn Mem **91**: 129–138.

Reiner, A., D. J. Perkel, L. L. Bruce, A. B. Butler, A. Csillag, W. Kuenzel, L. Medina, G. Paxinos, T. Shimizu, G. Striedter, M. Wild, G. F. Ball, S. Durand, O. Gunturkun, D. W. Lee, C. V. Mello, A. Powers, S. A. White, G. Hough, L. Kubikova, T. V. Smulders, K. Wada, J. Dugas-Ford, S. Husband, K. Yamamoto, J. Yu, C. Siang, E. D. Jarvis and F. Avian Brain Nomenclature (2004). "Revised nomenclature for avian telencephalon and some related brainstem nuclei." J Comp Neurol **473**(3): 377–414.

Rigotti, M., O. Barak, M. R. Warden, X. J. Wang, N. D. Daw, E. K. Miller and S. Fusi (2013). "The importance of mixed selectivity in complex cognitive tasks." Nature **497**(7451): 585–590.

Rigotti, M., D. Ben Dayan Rubin, X. J. Wang and S. Fusi (2010). "Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses." Front Comput Neurosci **4**: 24.

Riley, M. R. and C. Constantinidis (2015). "Role of Prefrontal Persistent Activity in Working Memory." Front Syst Neurosci **9**: 181.

Riley, M. R. and C. Constantinidis (2016). "Role of prefrontal persistent activity in working memory." Front Syst Neurosci **9**: 181.

Riley, M. R., X. L. Qi and C. Constantinidis (2017). "Functional specialization of areas along the anterior-posterior axis of the primate prefrontal cortex." Cereb Cortex **27**(7): 3683–3697.

Riley, M. R., X. L. Qi and C. Constantinidis (2017). "Functional specialization of areas along the anterior-posterior axis of the primate prefrontal cortex." Cereb Cortex **27**: 3683–3697.

Riley, M. R., X. L. Qi, X. Zhou and C. Constantinidis (2018). "Anterior-posterior gradient of plasticity in primate prefrontal cortex." Nat Commun **9**(1): 3790.

Rinnert, P., M. E. Kirschhock and A. Nieder (2019). "Neuronal Correlates of Spatial Working Memory in the Endbrain of Crows." Curr Biol **29**(16): 2616–2624 e2614.

Rodermund, P., S. Westendorff and A. Nieder (2020). "Blockage of NMDA- and GABA(A) Receptors Improves Working Memory Selectivity of Primate Prefrontal Neurons." J Neurosci **40**(7): 1527–1537.

Roelfsema, P. R. (2015). "The role of the different layers of primary visual cortex in working memory." J Vis **15**(12): 1406.

Romanski, L. M., J. F. Bates and P. S. Goldman-Rakic (1999). "Auditory belt and parabelt projections to the prefrontal cortex in the rhesus monkey." Journal of Comparative Neurology **403**(2): 141–157.

Romanski, L. M., M. Giguere, J. F. Bates and P. S. Goldman-Rakic (1997). "Topographic organization of medial pulvinar connections with the prefrontal cortex in the rhesus monkey." J Comp Neurol **379**(3): 313–332.

Romanski, L. M. and P. S. Goldman-Rakic (2002). "An auditory domain in primate prefrontal cortex." Nat Neurosci **5**(1): 15–16.

Romanski, L. M., B. Tian, J. Fritz, M. Mishkin, P. S. Goldman-Rakic and J. P. Rauschecker (1999). "Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex." Nature Neuroscience **2**(12): 1131–1136.

Romo, R., C. D. Brody, A. Hernandez and L. Lemus (1999). "Neuronal correlates of parametric working memory in the prefrontal cortex." Nature **399**(6735): 470–473.

Romo, R., A. Hernandez, A. Zainos, L. Lemus and C. D. Brody (2002). "Neuronal correlates of decision-making in secondary somatosensory cortex." Nat Neurosci **5**(11): 1217–1225.

Romo, R. and E. Salinas (2003). "Flutter discrimination: neural codes, perception, memory and decision making." Nat Rev Neurosci **4**(3): 203–218.

Rose, N. S., J. J. LaRocque, A. C. Riggall, O. Gosseries, M. J. Starrett, E. E. Meyering and B. R. Postle (2016). "Reactivation of latent working memories with transcranial magnetic stimulation." Science **354**(6316): 1136–1139.

Rossi, A. F., N. P. Bichot, R. Desimone and L. G. Ungerleider (2007). "Top down attentional deficits in macaques with lesions of lateral prefrontal cortex." J Neurosci **27**(42): 11306–11314.

Rouder, J. N., R. D. Morey, C. C. Morey and N. Cowan (2011). "How to measure working memory capacity in the change detection paradigm." Psychon Bull Rev **18**(2): 324–330.

Rouiller, E. M., J. Tanne, V. Moret and D. Boussaoud (1999). "Origin of thalamic inputs to the primary, premotor, and supplementary motor cortical areas and to

area 46 in macaque monkeys: a multiple retrograde tracing study." J Comp Neurol **409**(1): 131–152.

Roux, F. and P. J. Uhlhaas (2014). "Working memory and neural oscillations: alpha–gamma versus theta–gamma codes for distinct WM information?" Trends in Cognitive Sciences **18**(1): 16–25.

Roy, J. E., T. J. Buschman and E. K. Miller (2014). "PFC neurons reflect categorical decisions about ambiguous stimuli." J Cogn Neurosci **26**(6): 1283–1291.

Rubinstein, J. S., D. E. Meyer and J. E. Evans (2001). "Executive control of cognitive processes in task switching." J Exp Psychol Hum Percept Perform **27**(4): 763–797.

Ruchkin, D. S., J. Grafman, K. Cameron and R. S. Berndt (2003). "Working memory retention systems: a state of activated long-term memory." Behav Brain Sci **26**(6): 709–728; discussion 728–777.

Rushworth, M. F., P. D. Nixon, M. J. Eacott and R. E. Passingham (1997). "Ventral prefrontal cortex is not essential for working memory." J Neurosci **17**(12): 4829–4838.

Rygula, R., S. C. Walker, H. F. Clarke, T. W. Robbins and A. C. Roberts (2010). "Differential contributions of the primate ventrolateral prefrontal and orbitofrontal cortex to serial reversal learning." J Neurosci **30**(43): 14552–14559.

Sakagami, M., K. Tsutsui, J. Lauwereyns, M. Koizumi, S. Kobayashi and O. Hikosaka (2001). "A code for behavioral inhibition on the basis of color, but not motion, in ventrolateral prefrontal cortex of macaque monkey." J Neurosci **21**(13): 4801–4808.

Salazar, R. F., N. M. Dotson, S. L. Bressler and C. M. Gray (2012). "Content-specific fronto-parietal synchronization during visual working memory." Science **338**(6110): 1097–1100.

Salinas, E. and T. J. Sejnowski (2000). "Impact of correlated synaptic input on output firing rate and variability in simple neuronal models." J Neurosci **20**(16): 6193–6209.

Sandhu, R. and B. J. Dyson (2013). "Modality and task switching interactions using bi-modal and bivalent stimuli." Brain Cogn **82**(1): 90–99.

Sarma, A., N. Y. Masse, X. J. Wang and D. J. Freedman (2016). "Task-specific versus generalized mnemonic representations in parietal and prefrontal cortices." Nat Neurosci **19**(1): 143–149.

Saunders, N., R. Downham, B. Turman, J. Kropotov, R. Clark, R. Yumash and A. Szatmary (2015). "Working memory training with tDCS improves behavioral and neurophysiological symptoms in pilot group with post-traumatic stress disorder (PTSD) and with poor working memory." Neurocase **21**(3): 271–278.

Sawaguchi, T. and P. S. Goldman-Rakic (1991). "D1 dopamine receptors in prefrontal cortex: involvement in working memory." Science **251**(4996): 947–950.

Sawaguchi, T. and P. S. Goldman-Rakic (1994). "The role of D1-dopamine receptor in working memory: local injections of dopamine antagonists into the prefrontal cortex of rhesus monkeys performing an oculomotor delayed-response task." Journal of Neurophysiology **71**(2): 515–528.

Sawaguchi, T. and M. Iba (2001). "Prefrontal cortical representation of visuospatial working memory in monkeys examined by local inactivation with muscimol." J Neurophysiol **86**(4): 2041–2053.

Schneegans, S. and P. M. Bays (2018). "Drift in Neural Population Activity Causes Working Memory to Deteriorate Over Time." J Neurosci **38**(21): 4859–4869.

Schneiders, J. A., B. Opitz, C. M. Krick and A. Mecklinger (2011). "Separating intra-modal and across-modal training effects in visual working memory: an fMRI investigation." Cereb Cortex **21**(11): 2555–2564.

Schoenbaum, G., A. A. Chiba and M. Gallagher (2000). "Changes in functional connectivity in orbitofrontal cortex and basolateral amygdala during learning and reversal training." J Neurosci **20**(13): 5179–5189.

Schurgin, M. W., J. T. Wixted and T. F. Brady (2020). "Psychophysical scaling reveals a unified theory of visual memory strength." Nat Hum Behav **4**(11): 1156–1172.

Schwaighofer, M., F. Fischer and M. Buhner (2015). "Does Working Memory Training Transfer? A Meta-Analysis Including Training Conditions as Moderators." Educational Psychologist **50**(2): 138–166.

Schweizer, S., J. Grahn, A. Hampshire, D. Mobbs and T. Dalgleish (2013). "Training the emotional brain: improving affective control through emotional working memory training." J Neurosci **33**(12): 5301–5311.

Scott, B. B., C. M. Constantinople, A. Akrami, T. D. Hanks, C. D. Brody and D. W. Tank (2017). "Fronto-parietal Cortical Circuits Encode Accumulated Evidence with a Diversity of Timescales." Neuron **95**(2): 385–398 e385.

Scott, B. H. and M. Mishkin (2016). "Auditory short-term memory in the primate auditory cortex." Brain Res **1640**(Pt B): 264–277.

Scott, B. H., M. Mishkin and P. Yin (2014). "Neural correlates of auditory short-term memory in rostral superior temporal cortex." Curr Biol **24**(23): 2767–2775.

Seamans, J. K., D. Durstewitz, B. R. Christie, C. F. Stevens and T. J. Sejnowski (2001). "Dopamine D1/D5 receptor modulation of excitatory synaptic inputs to layer V prefrontal cortex neurons." Proc Natl Acad Sci U S A **98**(1): 301–306.

Selemon, L. D. and P. S. Goldman-Rakic (1985). "Longitudinal topography and interdigitation of corticostriatal projections in the rhesus monkey." J Neurosci **5**(3): 776–794.

Selemon, L. D. and P. S. Goldman-Rakic (1988). "Common cortical and subcortical targets of the dorsolateral prefrontal and posterior parietal cortices in the rhesus monkey: evidence for a distributed neural network subserving spatially guided behavior." Journal of Neuroscience **8**(11): 4049–4068.

Selemon, L. D. and P. S. Goldman-Rakic (1988). "Common cortical and subcortical targets of the dorsolateral prefrontal and posterior parietal cortices in the rhesus monkey: evidence for a distributed neural network subserving spatially guided behavior." J Neurosci **8**(11): 4049–4068.

Sereno, A. B. and J. H. Maunsell (1998). "Shape selectivity in primate lateral intraparietal cortex." Nature **395**(6701): 500–503.

Seung, H. S., D. D. Lee, B. Y. Reis and D. W. Tank (2000). "Stability of the memory of eye position in a recurrent network of conductance-based model neurons." Neuron **26**(1): 259–271.

Shanahan, M., V. P. Bingman, T. Shimizu, M. Wild and O. Gunturkun (2013). "Large-scale network organization in the avian forebrain: a connectivity matrix and theoretical analysis." Front Comput Neurosci **7**: 89.

Shima, K., M. Isoda, H. Mushiake and J. Tanji (2007). "Categorization of behavioural sequences in the prefrontal cortex." Nature **445**(7125): 315–318.

Siegel, M., M. R. Warden and E. K. Miller (2009). "Phase-dependent neuronal coding of objects in short-term memory." Proc Natl Acad Sci U S A **106**(50): 21341–21346.

Sigala, N., M. Kusunoki, I. Nimmo-Smith, D. Gaffan and J. Duncan (2008). "Hierarchical coding for sequential task events in the monkey prefrontal cortex." Proc Natl Acad Sci U S A **105**(33): 11969–11974.

Sigala, N. and N. K. Logothetis (2002). "Visual categorization shapes feature selectivity in the primate temporal cortex." Nature **415**(6869): 318–320.

Simmonds, D. J., M. N. Hallquist and B. Luna (2017). "Protracted development of executive and mnemonic brain systems underlying working memory in adolescence: a longitudinal fMRI study." Neuroimage.

Simons, D. J. and C. F. Chabris (1999). "Gorillas in our midst: sustained inattentional blindness for dynamic events." Perception **28**(9): 1059–1074.

Sinz, F. H., X. Pitkow, J. Reimer, M. Bethge and A. S. Tolias (2019). "Engineering a Less Artificial Intelligence." Neuron **103**(6): 967–979.

Skirzewski, M., S. Molotchnikoff, L. F. Hernandez and J. F. Maya-Vetencourt (2021). "Multisensory Integration: Is Medial Prefrontal Cortex Signaling Relevant for the Treatment of Higher-Order Visual Dysfunctions?" Front Mol Neurosci **14**: 806376.

Sommer, M. A. and E. J. Tehovnik (1997). "Reversible inactivation of macaque frontal eye field." Experimental Brain Research **116**(2): 229–249.

Soto, D., T. Mantyla and J. Silvanto (2011). "Working memory without consciousness." Curr Biol **21**(22): R912–913.

Soto, D. and J. Silvanto (2014). "Reappraising the relationship between working memory and conscious awareness." Trends Cogn Sci **18**(10): 520–525.

Sowell, E. R., D. Delis, J. Stiles and T. L. Jernigan (2001). "Improved memory functioning and frontal lobe maturation between childhood and adolescence: a structural MRI study." J Int Neuropsychol Soc **7**(3): 312–322.

Spaak, E., K. Watanabe, S. Funahashi and M. G. Stokes (2017). "Stable and Dynamic Coding for Working Memory in Primate Prefrontal Cortex." J Neurosci **37**(27): 6503–6516.

Sperling, G. (1960). The information available in brief visual presentations. Washington, American Psychological Association.

Spingath, E. Y., H. S. Kang, T. Plummer and D. T. Blake (2011). "Different neuroplasticity for task targets and distractors." PLoS One **6**(1): e15342.

Sprague, T. C., E. F. Ester and J. T. Serences (2014). "Reconstructions of information in visual spatial working memory degrade with memory load." Curr Biol **24**(18): 2174–2180.

Sprague, T. C., E. F. Ester and J. T. Serences (2016). "Restoring Latent Visual Working Memory Representations in Human Cortex." Neuron **91**(3): 694–707.

Spronk, M., E. K. Vogel and L. M. Jonkman (2012). "Electrophysiological evidence for immature processing capacity and filtering in visuospatial working memory in adolescents." PLoS One **7**(8): e42262.

Squire, L. R. and J. T. Wixted (2011). "The cognitive neuroscience of human memory since H.M." Annu Rev Neurosci **34**: 259–288.

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

Stanley, D. A., J. E. Roy, M. C. Aoi, N. J. Kopell and E. K. Miller (2018). "Low-Beta Oscillations Turn Up the Gain During Category Judgments." Cereb Cortex **28**(1): 116–130.

Stein, H., J. Barbosa, M. Rosa-Justicia, L. Prades, A. Morato, A. Galan-Gadea, H. Arino, E. Martinez-Hernandez, J. Castro-Fornieles, J. Dalmau and A. Compte (2020). "Reduced serial dependence suggests deficits in synaptic potentiation in anti-NMDAR encephalitis and schizophrenia." Nat Commun **11**(1): 4250.

Steinmetz, M. A., C. E. Connor, C. Constantinidis and J. R. McLaughlin (1994). "Covert attention suppresses neuronal responses in area 7a of the posterior parietal cortex." Journal of Neurophysiology **72**(2): 1020–1023.

Stern, C. E., S. J. Sherman, B. A. Kirchhoff and M. E. Hasselmo (2001). "Medial temporal and prefrontal contributions to working memory tasks with novel and familiar stimuli." Hippocampus **11**(4): 337–346.

Stokes, M. G. (2015). "'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework." Trends Cogn Sci **19**(7): 394–405.

Stokes, M. G., M. Kusunoki, N. Sigala, H. Nili, D. Gaffan and J. Duncan (2013). "Dynamic coding for cognitive control in prefrontal cortex." Neuron **78**(2): 364–375.

Subramaniam, K., T. L. Luks, M. Fisher, G. V. Simpson, S. Nagarajan and S. Vinogradov (2012). "Computerized cognitive training restores neural activity within the reality monitoring network in schizophrenia." Neuron **73**(4): 842–853.

Suchow, J. W., D. Fougnie, T. F. Brady and G. A. Alvarez (2014). "Terms of the debate on the format and structure of visual memory." Atten Percept Psychophys **76**(7): 2071–2079.

Sugase-Miyamoto, Y., Z. Liu, M. C. Wiener, L. M. Optican and B. J. Richmond (2008). "Short-term memory trace in rapidly adapting synapses of inferior temporal cortex." PLoS Comput Biol **4**(5): e1000073.

Sugihara, T., M. D. Diltz, B. B. Averbeck and L. M. Romanski (2006). "Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex." J Neurosci **26**(43): 11138–11147.

Super, H., H. Spekreijse and V. A. Lamme (2001). "A neural correlate of working memory in the monkey primary visual cortex." Science **293**(5527): 120–124.

Suzuki, M. and J. Gottlieb (2013). "Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe." Nat Neurosci **16**(1): 98–104.

Swaminathan, S. K. and D. J. Freedman (2012). "Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex." Nat Neurosci **15**(2): 315–320.

Sweller, J. (2016). "Working Memory, Long-term Memory, and Instructional Design." Journal of Applied Research in Memory and Cognition **5**(4): 360–367.

Sylwestrak, E. L., Y. Jo, S. Vesuna, X. Wang, B. Holcomb, R. H. Tien, D. K. Kim, L. Fenno, C. Ramakrishnan, W. E. Allen, R. Chen, K. V. Shenoy, D. Sussillo and K. Deisseroth (2022). "Cell-type-specific population dynamics of diverse reward computations." Cell **185**(19): 3568–3587 e3527.

Takeda, K. and S. Funahashi (2002). "Prefrontal task-related activity representing visual cue location or saccade direction in spatial working memory tasks." J Neurophysiol **87**(1): 567–588.

Takeuchi, H., A. Sekiguchi, Y. Taki, S. Yokoyama, Y. Yomogida, N. Komuro, T. Yamanouchi, S. Suzuki and R. Kawashima (2010). "Training of working memory impacts structural connectivity." J Neurosci **30**(9): 3297–3303.

Takeuchi, H., Y. Taki, R. Nouchi, H. Hashizume, A. Sekiguchi, Y. Kotozaki, S. Nakagawa, C. M. Miyauchi, Y. Sassa and R. Kawashima (2013). "Effects of working memory training on functional connectivity and cerebral blood flow during rest." Cortex **49**(8): 2106–2125.

Tanaka, K., H. Saito, Y. Fukada and M. Moriya (1991). "Coding visual images of objects in the inferotemporal cortex of the macaque monkey." J Neurophysiol **66**(1): 170–189.

Tang, H., X. L. Qi, M. R. Riley and C. Constantinidis (2019). "Working memory capacity is enhanced by distributed prefrontal activation and invariant temporal dynamics." Proc Natl Acad Sci U S A **116**(14): 7095–7100.

Tang, H., M. R. Riley and C. Constantinidis (2017). "Lateralization of Executive Function: Working Memory Advantage for Same Hemifield Stimuli in the Monkey." Front Neurosci **11**: 532.

Tang, H., M. R. Riley, B. Singh, X. L. Qi, D. T. Blake and C. Constantinidis (2022). "Prefrontal cortical plasticity during learning of cognitive tasks." Nat Commun **13**(1): 90.

Tanibuchi, I. and P. S. Goldman-Rakic (2003). "Dissociation of spatial-, object-, and sound-coding neurons in the mediodorsal nucleus of the primate thalamus." J Neurophysiol **89**(2): 1067–1077.

Taylor, J. and Y. Xu (2024). "Using fMRI to examine nonlinear mixed selectivity tuning to task and category in the human brain." Imaging Neurosci (Camb) **2**.

Thompson, T. W., M. L. Waskom and J. D. Gabrieli (2016). "Intensive Working Memory Training Produces Functional Changes in Large-scale Frontoparietal Networks." J Cogn Neurosci **28**(4): 575–588.

Todd, J. J. and R. Marois (2004). "Capacity limit of visual short-term memory in human posterior parietal cortex." Nature **428**(6984): 751–754.

Tomov, M. S., S. Yagati, A. Kumar, W. Yang and S. J. Gershman (2020). "Discovery of hierarchical representations for efficient planning." PLoS Comput Biol **16**(4): e1007594.

Tong, F. and M. S. Pratte (2012). "Decoding patterns of human brain activity." Annual review of psychology **63**: 483–509.

Torres-Gomez, S., J. D. Blonde, D. Mendoza-Halliday, E. Kuebler, M. Everest, X. J. Wang, W. Inoue, M. O. Poulter and J. Martinez-Trujillo (2020). "Changes in the Proportion of Inhibitory Interneuron Types from Sensory to Executive Areas of the Primate Neocortex: Implications for the Origins of Working Memory Representations." Cereb Cortex **30**(8): 4544–4562.

Treue, S. and J. H. Maunsell (1996). "Attentional modulation of visual motion processing in cortical areas MT and MST." Nature **382**(6591): 539–541.

Tsao, D. Y., W. A. Freiwald, R. B. Tootell and M. S. Livingstone (2006). "A cortical region consisting entirely of face-selective cells." Science **311**(5761): 670–674.

Tye, K. M., E. K. Miller, F. H. Taschbach, M. K. Benna, M. Rigotti and S. Fusi (2024). "Mixed selectivity: Cellular computations for complexity." Neuron **112**(14): 2289–2303.

Ungerleider, L. G., S. M. Courtney and J. V. Haxby (1998). "A neural system for human visual working memory." Proceedings of the National Academy of Sciences of the United States of America **95**(3): 883–890.

Ungerleider, L. G. and M. Mishkin (1982). Two cortical visual systems. Analysis of Visual Behavior. D. J. Ingle, M. A. Goodale and R. J. W. Mansfield. Cambridge, MA, MIT Press**:** 549–586.

Unsworth, N. and M. K. Robison (2020). "Working memory capacity and sustained attention: A cognitive-energetic perspective." J Exp Psychol Learn Mem Cogn **46**(1): 77–103.

Veit, L., K. Hartmann and A. Nieder (2014). "Neuronal correlates of visual working memory in the corvid endbrain." J Neurosci **34**(23): 7778–7786.

Veit, L. and A. Nieder (2013). "Abstract rule neurons in the endbrain support intelligent behaviour in corvid songbirds." Nat Commun **4**: 2878.

Vergara, J., N. Rivera, R. Rossi-Pool and R. Romo (2016). "A neural parametric code for storing information of more than one sensory modality in working memory." Neuron.

Verhaeghen, P. and A. Marcoen (1996). "On the mechanisms of plasticity in young and older adults after instruction in the method of loci: evidence for an amplification model." Psychol Aging **11**(1): 164–178.

Vermeij, A., R. P. Kessels, L. Heskamp, E. M. Simons, P. L. Dautzenberg and J. A. Claassen (2016). "Prefrontal activation may predict working-memory training gain in normal aging and mild cognitive impairment." Brain Imaging Behav.

Vijayraghavan, S., M. Wang, S. G. Birnbaum, G. V. Williams and A. F. Arnsten (2007). "Inverted-U dopamine D1 receptor actions on prefrontal neurons engaged in working memory." Nat Neurosci **10**(3): 376–384.

Vogel, E. K. and M. G. Machizawa (2004). "Neural activity predicts individual differences in visual working memory capacity." Nature **428**(6984): 748–751.

Volle, E., S. Kinkingnehun, J. B. Pochon, K. Mondon, M. Thiebaut de Schotten, M. Seassau, H. Duffau, Y. Samson, B. Dubois and R. Levy (2008). "The functional architecture of the left posterior and lateral prefrontal cortex in humans." Cereb Cortex **18**(10): 2460–2469.

Walker, A. E. (1940). "A cytoarchitectural study of the prefrontal area of the macaque monkey." Journal of Comparative Neurology **73**: 59–86.

Wallis, J. D., K. C. Anderson and E. K. Miller (2001). "Single neurons in prefrontal cortex encode abstract rules." Nature **411**(6840): 953–956.

Wang, L., X. Li, S. S. Hsiao, F. A. Lenz, M. Bodner, Y. D. Zhou and J. M. Fuster (2015). "Differential roles of delay-period neural activity in the monkey dorsolateral prefrontal cortex in visual-haptic crossmodal working memory." Proc Natl Acad Sci U S A **112**(2): E214–219.

Wang, M. and A. F. Arnsten (2015). "Contribution of NMDA receptors to dorsolateral prefrontal cortical networks in primates." Neurosci Bull **31**(2): 191–197.

Wang, M., N. J. Gamo, Y. Yang, L. E. Jin, X. J. Wang, M. Laubach, J. A. Mazer, D. Lee and A. F. Arnsten (2011). "Neuronal basis of age-related working memory decline." Nature **476**(7359): 210–213.

Wang, M., B. P. Ramos, C. D. Paspalas, Y. Shu, A. Simen, A. Duque, S. Vijayraghavan, A. Brennan, A. Dudley, E. Nou, J. A. Mazer, D. A. McCormick and A. F.

Arnsten (2007). "Alpha2A-adrenoceptors strengthen working memory networks by inhibiting cAMP-HCN channel signaling in prefrontal cortex." Cell **129**(2): 397–410.

Wang, M., S. Vijayraghavan and P. S. Goldman-Rakic (2004). "Selective D2 receptor actions on the functional circuitry of working memory." Science **303**(5659): 853–856.

Wang, M., A. H. Wong and F. Liu (2012). "Interactions between NMDA and dopamine receptors: a potential therapeutic target." Brain Res **1476**: 154–163.

Wang, M., Y. Yang, C. J. Wang, N. J. Gamo, L. E. Jin, J. A. Mazer, J. H. Morrison, X. J. Wang and A. F. Arnsten (2013). "NMDA receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex." Neuron **77**(4): 736–749.

Wang, X. J. (1999). "Synaptic Basis of Cortical Persistent Activity: the Importance of NMDA Receptors to Working Memory." J. Neurosci. **19**(21): 9587–9603.

Wang, X. J. (2001). "Synaptic reverberation underlying mnemonic persistent activity." Trends Neurosci **24**(8): 455–463.

Wang, X. J., J. Tegner, C. Constantinidis and P. S. Goldman-Rakic (2004). "Division of labor among distinct subtypes of inhibitory neurons in a cortical microcircuit of working memory." Proc Natl Acad Sci U S A **101**(5): 1368–1373.

Wardak, C., E. Olivier and J. R. Duhamel (2002). "Saccadic target selection deficits after lateral intraparietal area inactivation in monkeys." J Neurosci **22**(22): 9877–9884.

Wardak, C., E. Olivier and J. R. Duhamel (2004). "A deficit in covert attention after parietal cortex inactivation in the monkey." Neuron **42**(3): 501–508.

Warden, M. R. and E. K. Miller (2007). "The representation of multiple objects in prefrontal neuronal delay activity." Cereb Cortex **17 Suppl 1**: i41–50.

Warden, M. R. and E. K. Miller (2010). "Task-dependent changes in short-term memory in the prefrontal cortex." J Neurosci **30**(47): 15801–15810.

Watanabe, K. and S. Funahashi (2014). "Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex." Nat Neurosci **17**(4): 601–611.

Watanabe, Y. and S. Funahashi (2004). "Neuronal activity throughout the primate mediodorsal nucleus of the thalamus during oculomotor delayed-responses. I. Cue-, delay-, and response-period activity." J Neurophysiol **92**(3): 1738–1755.

Watanabe, Y. and S. Funahashi (2004). "Neuronal Activity Throughout the Primate Mediodorsal Nucleus of the Thalamus During Oculomotor Delayed-Responses. II. Activity Encoding Visual Versus Motor Signal." J Neurophysiol **92**(3): 1756–1769.

Westerberg, H., T. Hirvikoski, H. Forssberg and T. Klingberg (2004). "Visuo-spatial working memory span: a sensitive measure of cognitive deficits in children with ADHD." Child Neuropsychol **10**(3): 155–161.

Westerberg, H., H. Jacobaeus, T. Hirvikoski, P. Clevberger, M. L. Ostensson, A. Bartfai and T. Klingberg (2007). "Computerized working memory training after stroke--a pilot study." Brain Inj **21**(1): 21–29.

White, I. M. and S. P. Wise (1999). "Rule-dependent neuronal activity in the prefrontal cortex." Exp Brain Res **126**(3): 315–335.

White, J. M., D. L. Sparks and T. R. Stanford (1994). "Saccades to remembered target locations: an analysis of systematic and variable errors." Vision Res **34**(1): 79–92.

Whitney, P., P. A. Arnett, A. Driver and D. Budd (2001). "Measuring central executive functioning: what's in a reading span?" Brain Cogn **45**(1): 1–14.

Wickelgren, W. A. (1965). "Acoustic Similarity and Intrusion Errors in Short-Term Memory." J Exp Psychol **70**: 102–108.

Wilke, M., I. Kagan and R. A. Andersen (2012). "Functional imaging reveals rapid reorganization of cortical activity after parietal inactivation in monkeys." Proc Natl Acad Sci U S A **109**(21): 8274–8279.

Williams, G. V. and P. S. Goldman-Rakic (1995). "Modulation of memory fields by dopamine D1 receptors in prefrontal cortex." Nature **376**(6541): 572–575.

Wilson, F. A., S. P. O Scalaidhe and P. S. Goldman-Rakic (1993). "Dissociation of object and spatial processing domains in primate prefrontal cortex." Science **260**(5116): 1955–1958.

Wilson, F. A., S. P. Scalaidhe and P. S. Goldman-Rakic (1993). "Dissociation of object and spatial processing domains in primate prefrontal cortex." Science **260**(5116): 1955–1958.

Wimmer, K., D. Q. Nykamp, C. Constantinidis and A. Compte (2014). "Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory." Nat Neurosci **17**(3): 431–439.

Wise, S. P., E. A. Murray and C. R. Gerfen (1996). "The frontal cortex-basal ganglia system in primates." Critical Reviews in Neurobiology **10**(3&4): 317–356.

Wixted, J. T. (1991). "Conditions and consequences of maintenance rehearsal." J Exp Psychol Learn Mem Cogn **17**(5): 963–973.

Woloszyn, L. and D. L. Sheinberg (2009). "Neural dynamics in inferior temporal cortex during a visual working memory task." J Neurosci **29**(17): 5494–5507.

Womelsdorf, T., P. Fries, P. P. Mitra and R. Desimone (2006). "Gamma-band synchronization in visual cortex predicts speed of change detection." Nature **439**(7077): 733–736.

Wutz, A., R. Loonis, J. E. Roy, J. A. Donoghue and E. K. Miller (2018). "Different Levels of Category Abstraction by Different Dynamics in Different Prefrontal Areas." Neuron **97**(3): 716–726 e718.

Xie, Y., Y. H. Liu, C. Constantinidis and X. Zhou (2022). "Neural Mechanisms of Working Memory Accuracy Revealed by Recurrent Neural Networks." Front Syst Neurosci **16**: 760864.

Xing, Y., T. Ledgeway, P. V. McGraw and D. Schluppeck (2013). "Decoding working memory of stimulus contrast in early visual cortex." J Neurosci **33**(25): 10301–10311.

Xu, M., T. Hosokawa, K. I. Tsutsui and K. Aihara (2024). "Dynamic tuning of neural stability for cognitive control." Proc Natl Acad Sci U S A **121**(49): e2409487121.

Xu, Y. (2002). "Encoding color and shape from different parts of an object in visual short-term memory." Percept Psychophys **64**(8): 1260–1280.

Yakovlev, P. I. and A. R. Lecours, Eds. (1967). The myelogenetic cycles of regional maturation of the brain. Regional Development of the Brain in Early Life. Oxford, Blackwell.

Yang, C. R. and J. K. Seamans (1996). "Dopamine D1 receptor actions in layers V-VI rat prefrontal cortex neurons in vitro: modulation of dendritic-somatic signal integration." J Neurosci **16**(5): 1922–1935.

Yang, G. R., M. R. Joglekar, H. F. Song, W. T. Newsome and X. J. Wang (2019). "Task representations in neural networks trained to perform many cognitive tasks." Nat Neurosci **22**(2): 297–306.

Yang, Y., C. D. Paspalas, L. E. Jin, M. R. Picciotto, A. F. Arnsten and M. Wang (2013). "Nicotinic alpha7 receptors enhance NMDA cognitive circuits in dorsolateral prefrontal cortex." Proc Natl Acad Sci U S A **110**(29): 12078–12083.

Yeung, M. S., S. Zdunek, O. Bergmann, S. Bernard, M. Salehpour, K. Alkass, S. Perl, J. Tisdale, G. Possnert, L. Brundin, H. Druid and J. Frisen (2014). "Dynamics of oligodendrocyte generation and myelination in the human brain." Cell **159**(4): 766–774.

Yokokura, M., K. Takebasashi, A. Takao, K. Nakaizumi, E. Yoshikawa, M. Futatsubashi, K. Suzuki, K. Nakamura, H. Yamasue and Y. Ouchi (2020). "In vivo

imaging of dopamine D1 receptor and activated microglia in attention-deficit/hyperactivity disorder: a positron emission tomography study." Mol Psychiatry.

Zaksas, D. and T. Pasternak (2006). "Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task." J Neurosci **26**(45): 11726–11742.

Zhang, W. H. and Z. M. Williams (2015). "Frontal neurons modulate memory retrieval across widely varying temporal scales." Learn Mem **22**(6): 299–306.

Zhao, D., Y. D. Zhou, M. Bodner and Y. Ku (2018). "The Causal Role of the Prefrontal Cortex and Somatosensory Cortex in Tactile Working Memory." Cereb Cortex **28**(10): 3468–3477.

Zhou, X., F. Katsuki, X. L. Qi and C. Constantinidis (2012). "Neurons with inverted tuning during the delay periods of working memory tasks in the dorsal prefrontal and posterior parietal cortex." J Neurophysiol **108**(1): 31–38.

Zhou, X., X. L. Qi, K. Douglas, K. Palaninathan, H. S. Kang, J. J. Buccafusco, D. T. Blake and C. Constantinidis (2011). "Cholinergic modulation of working memory activity in primate prefrontal cortex." J Neurophysiol **106**(5): 2180–2188.

Zhou, X., D. Zhu, F. Katsuki, X. L. Qi, C. J. Lees, A. J. Bennett, E. Salinas, T. R. Stanford and C. Constantinidis (2014). "Age-dependent changes in prefrontal intrinsic connectivity." Proc Natl Acad Sci U S A **111**(10): 3853–3858.

Zhou, X., D. Zhu, X. L. Qi, C. J. Lees, A. J. Bennett, E. Salinas, T. R. Stanford and C. Constantinidis (2013). "Working Memory Performance and Neural Activity in the Prefrontal Cortex of Peri-pubertal Monkeys." J Neurophysiol **110**: 2648–2660.

Zhou, X., D. Zhu, X. L. Qi, S. Li, S. G. King, E. Salinas, T. R. Stanford and C. Constantinidis (2016). "Neural correlates of working memory development in adolescent primates." Nat Commun **7**: 13423.

Zhou, Y. D. and J. M. Fuster (1996). "Mnemonic neuronal activity in somatosensory cortex." Proceedings of the National Academy of Sciences of the United States of America **93**(19): 10533–10537.

Zhu, J., C. M. Garin, X. L. Qi, A. Machado, Z. Wang, S. B. Hamed, T. R. Stanford, E. Salinas, C. T. Whitlow, A. W. Anderson, X. M. Zhou, F. J. Calabro, B. Luna and C. Constantinidis (2024). "Brain structure and activity predicting cognitive maturation in adolescence." bioRxiv.

Zick, J. L., R. K. Blackman, D. A. Crowe, B. Amirikian, A. L. DeNicola, T. I. Netoff and M. V. Chafee (2018). "Blocking NMDAR Disrupts Spike Timing and Decouples Monkey Prefrontal Circuits: Implications for Activity-Dependent Disconnection in Schizophrenia." Neuron **98**(6): 1243–1255 e1245.

CURRICULUM VITAE

RYE JAFFE 121 Acklen Park Drive
Neuroscience PhD Candidate, Wake Forest University Apt B, Nashville, TN

SELECTED AWARDS

- Ruth L. Kirschstein Pre-Doctoral National Research Service Award, 2023-2026
- Departmental Honors in Neuroscience from Oberlin College, 2018

PEER-REVIEWED PUBLICATIONS

- Jaffe, R., W. Dang, T. Gao, J. Zhu and C. Constantinidis (2026). "Prefrontal Mechanisms of Rule Learning." *bioRxiv.* https://doi.org/10.64898/2026.04.01.715865
- Sun, Y., W. Dang, R. G. Jaffe and C. Constantinidis (2024). "Local organization of spatial and shape information in the primate prefrontal cortex." Cereb Cortex **34**(9).
- Thrower, L., W. Dang, R. G. Jaffe, J. D. Sun and C. Constantinidis (2023). "Decoding working memory information from neurons with and without persistent activity in the primate prefrontal cortex." J Neurophysiol **130**(6): 1392–1402.
- Garin, C. M., M. Garin, L. Silenzi, R. Jaffe and C. Constantinidis (2022). "Multilevel atlas comparisons reveal divergent evolution of the primate brain." Proc Natl Acad Sci U S A **119**(25): e2202491119.
- Jaffe, R. J. and C. Constantinidis (2021). "Working Memory: From Neural Activity to the Sentient Mind." Compr Physiol **11**(4): 2547–2587.
- Dang, W., R. J. Jaffe, X. L. Qi and C. Constantinidis (2021). "Emergence of Nonlinear Mixed Selectivity in Prefrontal Cortex after Training." J Neurosci **41**(35): 7420–7434.
- Trivedi, J., D. Mahajan, R. J. Jaffe, A. Acharya, D. Mitra and S. N. Byrareddy (2020). "Recent Advances in the Development of Integrase Inhibitors for HIV Treatment." Curr HIV/AIDS Rep **17**(1): 63–75.
- Jaffe, R. J., R. S. Dave and S. N. Byrareddy (2019). "Meningeal lymphatics in aging and Alzheimer's disease." Ann Transl Med **7**(Suppl 1): S2.
- Foster, J. J., E. M. Bsales, R. J. Jaffe and E. Awh (2017). "Alpha-Band Activity Reveals Spontaneous Representations of Spatial Position in Visual Working Memory." Curr Biol **27**(20): 3216–3223 e3216.

CONFERENCE PRESENTATIONS

Jaffe R, Dang W, Gao T, Constantinidis C. Common Mechanisms of Rule Learning Across Tasks in the Prefrontal Cortex. Society for Neuroscience 2025 meeting. San Diego, CA.